Probability and statistics for geophysical processes

Demetris Koutsoyiannis
National Technical University of Athens

Chapter 1: The utility of probability – Chapter 2: Basic concepts of probability – Chapter 3: Elementary statistical concepts – Chapter 4: Special concepts of probability theory in geophysical applications – Chapter 5: Typical univariate statistical analysis in geophysical processes – Chapter 6: Typical distribution functions in geophysics, hydrology and water resources – Appendix (Statistical tables)

Chapter 1

The utility of probability

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

Summary

Commonly, probability is regarded to be a branch of applied mathematics that provides tools for data analysis. Nonetheless, probability is a more general concept that helps shape a consistent, realistic and powerful view of the world. Historically, the modern science was initiated from deterministic views of the world, in which probability had a marginal role for peculiar unpredictable phenomena. However, in the turn of the nineteenth century, radical developments in physics, and particularly thermodynamics, dynamical systems and quantum physics, as well as in mathematics has given the probability theory a central role in the scientific scene, in the understanding and the modelling of natural phenomena. Furthermore, probability has provided grounds for philosophical concepts such as indeterminism and causality, as well as for extending the typical mathematical logic, offering the mathematical foundation of induction. In typical scientific and technological applications, probability provides the tools to quantify uncertainty, rationalize decisions under uncertainty, and make predictions of future events under uncertainty, in lieu of unsuccessful deterministic predictions. Uncertainty seems to be an intrinsic property of nature, as it can emerge even from pure and simple deterministic dynamics, and cannot been eliminated. This is demonstrated here using a working example with simple deterministic equations and showing that deterministic methods may be good for short-term predictions but for long horizons their predictive capacity is cancelled, whereas the probabilistic methods can perform much better.

1.1 Determinism and indeterminism

The philosophical proposition of *determinism* is widely accepted in science. It is manifested in the idea of a clockwork universe, which comes from the French philosopher and scientist René Descartes (1596-1650) and was perfected by the French mathematician and astronomer Pierre-Simon Laplace (1749-1827). It is vividly expressed in the metaphor of *Laplace's demon*, a hypothetical all-knowing entity that knows the precise location and momentum of every atom in the universe at present, and therefore could use Newton's laws to reveal the entire course of cosmic events, past and future. (Isaac Newton – 1643-1727 – however, rejected cartesian thinking and especially the clockwork idea; he was aware of the fragility the world and believed that God had to keep making adjustments all the time to correct the emerging chaos.) The demon who knows the present perfectly is of course a metaphor; what is more important in this idea is that knowing the present perfectly one can deduce the future and the past using Newton's laws. The metaphor helps us understand that, according to deterministic thinking, the roots of uncertainty about future should be subjective, i.e. rely on the fact that we do not know exactly the present, or we do not have good enough methods and models. It is then a matter of time to eliminate uncertainty, acquiring better data (observations) and building better models.

However, according to *indeterminism*, a philosophical belief contradictory to determinism, uncertainty may be a structural element of nature and thus cannot be eliminated. Indeterminism has its origin in the Greek philosophers Heraclitus (*ca*. 535–475 BC) and Epicurus (341–270 BC). In science, indeterminism largely relies on the notion of *probability*, which according to the Austrian-British philosopher Karl Popper (1902-1994) is the extension (quantification) of the Aristotelian idea of *potentia* (Popper, 1982, p. 133). Practically, the idea is that several outcomes can be produced by a specified cause, while in deterministic thinking only one outcome is possible (but it may be difficult to predict which one). Probability is a quantification of the likelihood of each outcome or of any set of outcomes. In this chapter we use the term probability in a loose manner. In the next chapter we will provide a precise description of the term using the axiomatization introduced by the soviet mathematician Andrey Nikolaevich Kolmogorov (1903-1987).

In everyday problems deterministic thinking may lead to deadlocks, for instance in dealing with the outcome of a dice throw or a roulette spin. The movements of both obey Newton's laws; however, application of these laws did not help anyone to become rich predicting the dice outcomes. In an attempt to rectify such deadlocks, some have been tempted to divide the natural phenomena into two categories, deterministic (e.g. the movement of planets) and random (e.g. the movement of dice). We maintain that this is a fallacy (both planets and dice obey to the same Newton's laws). Another very common fallacy of the same type (in fact, an extension of the former) is the attempt to separate natural processes into deterministic and random components, one superimposed (usually added) to the other. Both fallacies can be avoided by abandoning the premise of determinism.

1.2 Deduction and induction

In mathematical logic, determinism can be paralleled to the premise that all truth can be revealed by *deductive reasoning* or *deduction* (the Aristotelian *apodeixis*). This type of reasoning consists of repeated application of strong syllogisms such as:

If A is true, then B is true
A is true
Therefore, B is true

and

If A is true, then B is true;
B is false
Therefore, A is false

Deduction uses a set of axioms to prove propositions known as theorems, which, given the axioms, are irrefutable, absolutely true statements. It is also irrefutable that deduction is the preferred route to truth; the question is, however, whether or not it has any limits. David Hilbert (1862-1943) expressed his belief that there are no limits in his slogan (from his talk in 1930; also inscribed in his tombstone at Göttingen): "*Wir müssen wissen, wir werden wissen* - We must know, we will know". His idea, more formally known as *completeness*, is that any mathematical statement could be proved or disproved by deduction from axioms.

In everyday life, however, we use weaker syllogisms of the type:

If A is true, then B is true;
B is true
Therefore, A becomes more plausible

and

If A is true, then B is true;
A is false
Therefore, B becomes less plausible

The latter type of syllogism is called *induction* (the Aristotelian *epagoge*). It does not offer a proof that a proposition is true or false and may lead to errors. However, it is very useful in decision making, when deduction is not possible.

An important achievement of probability is that it quantifies (expresses in the form of a number between 0 and 1) the degree of plausibility of a certain proposition or statement. The formal probability framework uses both deduction, for proving theorems, and induction, for inference with incomplete information or data.

1.3 The illusion of certainty and its collapse

Determinism in physics and completeness in mathematics reflect the idea that uncertainty could in principle be eliminated. However, in the turn of the nineteenth century and the first half of the twentieth century this idea proved to be an illusion as it received several blows in four major scientific areas, summarized below.

1.3.1 Statistical physics and maximum entropy

In its initial steps, thermodynamics was based on purely deterministic concepts and particularly on the notion of the caloric fluid, a hypothetical fluid (a weightless gas) that flows from hotter to colder bodies (passes in pores of solids and liquids). The caloric theory was proposed in 1783 by Antoine Lavoisier and persisted in scientific literature until the end of the 19th century. In 1902 the term *statistical thermodynamics* was coined by the American mathematical-engineer, physicist, and chemist J. Willard Gibbs. The statistical theory of thermodynamics is essentially based on the probabilistic description of kinetic properties of atoms and molecules and was very successful in explaining all concepts and phenomena related to heat transfer.

The concept of *entropy* (from the Greek *εντροπία*), which was essential for the formulation of the second law of thermodynamics (by Rudolf Clausius in 1850), was given a statistical interpretation by Ludwig Boltzmann (in 1872). The second law says that the entropy of an isolated system which will tend to increase over time, approaching a maximum value at equilibrium. Boltzmann showed that entropy can be defined in terms of the number of possible microscopic configurations that result in the observed macroscopic description of a thermodynamic system. In 1878, Gibbs extended this notion of entropy introducing the idea of the statistical (or thermodynamic) ensemble, an idealization consisting of a large number (sometimes infinitely many) of mental copies of a system, each of which represents a possible state that the real system might be in. In 1948, Claude E. Shannon generalized the concept of entropy and gave it an abstract probabilistic definition applicable for any random variable, thus essentially showing that entropy is a measure of uncertainty of a system. Kolmogorov and his student Sinai went far beyond and suggested a definition of the metric entropy for dynamical systems (their results were published in 1959). In 1957, the American mathematician and physicist Edwin Thompson Jaynes extended Gibbs' statistical mechanics ideas showing that they can be applied for statistical inference about any type of a system. Specifically, he showed that the *principle of maximum entropy* can be used as a general method to infer the unknown probability distribution of any random variable. For instance, the principle of maximum entropy can easily produce that the probability of the landing of a die in each of its six faces will be 1/6 (any departure from equality of all six probabilities would decrease the uncertainty of the event). It is thus impressive that the principle that predicts that heat spontaneously flows from a hot to a cold body, is the same principle that can give the probability distribution of dice.

Thus, statistical thermodynamics has formed a nice paradigm entirely based on probability as a tool for both explanation and mathematical description of natural behaviours. Furthermore, the second law of thermodynamics essentially shows that nature works in a way that maximizes uncertainty in complex systems. Following nature's behaviour and applying the principle of maximum entropy (maximum uncertainty) to any type of system we can infer useful knowledge about the system's behaviour. This knowledge, however, is no longer expressed in terms of certainty about the sharp states of the system, but rather in terms of probabilities of these states. In large systems however, it turns out that this knowledge can lead to nearly precise descriptions of macroscopical properties, despite the maximum uncertainty at the microscopical level. For instance, we can easily infer that the average of the outcomes of 45 000 dice throws will be between 3.49 and 3.51 with probability 99.99%. From a practical point of view such a statement is almost equivalent to certainty; however, it does not preclude the case that all 45 000 will be sixes (and the average will be also six).

1.3.2 Dynamical systems and chaos

Chaos (from the Greek χάος) is most often referred to as a deterministic notion (deterministic chaos). Yet it offers an excellent insight of uncertainty, even in the case of purely deterministic dynamics. The basic concepts of chaos are due to the French mathematician Jules Henri Poincaré (1854–1912). In 1890, Poincaré's memoir on the three body problem was published in the journal Acta Mathematica as the winning entry in the international prize competition sponsored by Oscar II, King of Sweden and Norway, to mark his 60th birthday. Today this paper is renowned for containing the first mathematical description of chaotic behavior in a dynamical system (Barrow-Green, 1994). It was the first time that the complexity of Newtonian dynamics was demonstrated, even in a system as apparently simple as three gravitational bodies. Poincaré gave the first example of the *sensitive dependence on initial conditions*, a characteristic of chaotic behaviour that is met in unstable dynamical systems.

Ironically, however, the prize winning work of Poincaré was not exactly the published one. In contrast, in his original work Poincaré, had found certain stability results for the three-body problem. After the prize award (1889) and after the prize winning essay had been printed (but not distributed), Poincaré discovered a fatal flaw in his proof that was supposed to show that the universe worked like clockwork. Poincaré then had to spend his monetary prize plus 1000 Crowns to withdraw the printed volumes with the erroneous version of the memoir, as well as several months of work to correct the error. In the final paper he had reinstated the chaos in the movement of the astral bodies and brought down for ever the idea of a clockwork universe.

We can understand the emergence of chaos and chance from purely deterministic dynamics reading his own words (from Henri Poincaré, Science et méthode, 1908; reproduced in Poincaré, 1956, p. 1382):

> *A very small cause, which escapes us, determines a considerable effect which we cannot help seeing, and then we say that the effect is due to chance. If we could know exactly the laws of nature and the situation of the universe at the initial instant, we should be able to predict the situation of this same universe at a subsequent instant. But even when the natural laws should have no further secret for us, we could know the initial situation only* ***approximately****. If that permits us to foresee the succeeding situation* ***with the same degree of approximation****, that is all we require, we say the phenomenon had been predicted, that it is ruled by laws. But it is not always the case; it may happen that slight differences in the initial conditions produce very great differences in the final phenomena; a slight error in the former would make an enormous error in the latter. Prediction becomes impossible, and we have the fortuitous phenomenon.*

> *... Why have the meteorologists such difficulty in predicting the weather with any certainty? Why do the rains, the tempests themselves seem to us to come by chance, so that many persons find it quite natural to pray for rain or shine, when they would think it ridiculous to pray for an eclipse? We see that great perturbations generally happen in regions where the atmosphere is in unstable equilibrium... Here again we find the same contrast between a very slight cause, inappreciable to the observer, and important effects, which are sometimes tremendous disasters.*

Non-linear chaotic dynamics remained in the backwoods of mathematics and physics until the 1960s, even though some of the leading mathematicians, mostly in Russia/USSR (Lyapunov, Kolmogorov, Andronov), worked on it. Then the use of computers made it possible to experiment with chaos in numerical applications. The American meteorologist Edward Norton Lorenz was an early pioneer of experimenting chaos with computers; also he coined the term *butterfly effect* to encapsulate the notion of sensitive dependence on initial conditions in chaotic systems: a butterfly's wings (a small change in the initial condition of the atmospheric system) might create tiny changes in the atmosphere that ultimately cause a tornado to appear.

Now the mathematical theory of nonlinear complex chaotic dynamic systems is centre stage and mainstream. A prominent characteristic of the notion of chaos is that it is easily understandable, as it may involve simple deterministic dynamics, and allows the experimentation with very simple examples that exhibit chaotic behaviour. Such a simple example we will study in the next section. It is fascinating that a simple nonlinear deterministic system (such as the gravitational movement of three bodies or the hydrological system studied below) can have a complex, erratic evolution. Sadly, however, most of hydrological studies understood this in the opposite direction: they attempted to show, making naïve and mistaken use of tools from dynamical systems, that complexity in hydrological phenomena implies that their dynamics are simple (Koutsoyiannis, 2006).

1.3.3 Quantum physics

While chaotic systems demonstrated that uncertainty can be produced even in a purely deterministic framework, quantum physics has shown that uncertainty is an intrinsic characteristic of nature. In this respect, probability is not only a necessary epistemic addition or luxury for modelling natural phenomena. Rather it is a structural element of nature, an ontological rather than epistemic concept.

Quantum physics has put limits to the knowledge we can obtain from observation of a microscopic system and has shown that exact measurements are impossible. The outcome of even an ideal measurement of a system is not sharp (exact), but instead is characterized by a probability distribution. The Heisenberg uncertainty principle gives a lower bound on the product of the uncertainty measures of position and momentum for a system, implying that it is impossible to have a particle that has an arbitrarily well-defined position and momentum simultaneously. Thus, our familiar deterministic description proves to be impossible for the microscopic world.

A famous example that shows how fundamental the notion of probability is in nature is the double-slit experiment. Light is shined at a thin, solid barrier that has two slits cut into it. A photographic plate is set up behind the barrier to record what passes through slits. When only one slit is open, there is only one possibility for a photon, to pass through the open slit. Indeed the plate shows a single dark line where all the photons are accumulated. However, when both slits are open and only one photon at a time is fired at the barrier, there are two possibilities for the photon, which however are not mutually exclusive because, according to the uncertainty principle, the position of the photon is not sharp. Thus, it seems that the photon passes from both slits simultaneously. This will be recorded in the photographic plate, which shows regions of brightness and darkness (interference fringes). It seems that a single photon materializes the theoretical probability distribution in each case. According to our macroscopic experience the phonon would follow one of the two available options, and at the time it passes though the barrier it would be in one of the two slits with equal probabilities. However, in a quantum physics description the photon is simultaneously in both slits and the two probabilities interfere.

Such phenomena are difficult to describe or explain based on our experience (and language) of the macroscopic world. However, the phenomena of the quantum physics are reflected in the macroscopic world too (e.g. in the double-slit experiment), and thus cannot be irrelevant to our description of macrocosmos. For instance, statistical physics is strongly influenced by quantum physics.

1.3.4 Incompleteness

While the three previous developments eventually deal with physics, this fourth one concerns pure mathematical logic. In 1931 the Austrian mathematician Kurt Gödel proved two important theorems, so-called *incompleteness theorems*, stating inherent limitations of mathematical logic. The theorems are also important in the philosophy of mathematics and in wider areas of philosophy. The first incompleteness theorem practically says that any system with some axioms, containing the natural numbers and basic arithmetic (addition, multiplication) is necessarily incomplete: it contains undecidable statements, i.e. statements that are neither provably true nor provably false. Furthermore, if an undecidable statement is added to the system as an axiom, there will always be other statements that still cannot be proved as true, even with the new axiom. The second theorem says that if the system can prove that it is consistent, then it is inconsistent. That is to say, we can never know that a system is consistent, meaning that it does not contain a contradiction. Note that if the system contains a contradiction, i.e. a case where a proposition and its negation are both provably true, then every proposition becomes true.

Ironically, Gödel had presented his incompleteness results the day before Hilbert pronounced his slogan discussed above (*Wir müssen wissen, wir werden wissen*). Obviously, the slogan received a strong blow by Gödel's results. The conjectured almightiness of deduction was vitiated. In other words, Gödel's results show that uncertainty is not eliminable. Simultaneously, they enhance the role of probability theory, as extended logic, and the necessity of induction (see also Jaynes, 2003, p. 47).

1.3.5 The positive side of uncertainty

Surprisingly, the new role of probability is not well assimilated in the scientific community. The quest of determinism and uncertainty elimination still dominates in science. Another symptom of this type is the exorcism of probability and its replacement with any type of substitutes. One good example for this is provided by the *fuzzy methods* which are regarded much more fashioned than probability. However, no solutions using fuzzy approaches could not have been achieved at least as effectively using probability and statistics (Laviolette, *et al.,* 1995). The Education still promotes deterministic thinking as if all above fundamental changes in science had not happened. Hopes are expressed that these results are flawed and determinism will be reinstated. These results are considered negative and pessimistic by many. We maintain that they are absolutely positive and optimistic. Life would not have any meaning without uncertainty. This is well known by people working in the media, who spend much money to show live (i.e. with uncertain outcome) reportages and sports games; had determinism been more fascinating, they would show recorded versions in the next day, with eliminated uncertainty (e.g. the score of the game would be known). Without uncertainty concepts such as *hope*, *will* (particularly *free will*), *freedom*, *expectation*, *optimism*, *pessimism* etc. would hardly make sense.

1.4 A working example

With this example we will see that, contrary to intuition, pure determinism does not help very much to predict the future, even in very simple systems. The example studies a hydrological system that is fully deterministic and is deliberately made extremely simple. This system is a natural plain with water stored in the soil, which sustains some vegetation. We assume that each year a constant amount of water I = 250 mm enters the soil and that the potential evapotranspiration is also constant, PET = 1000 mm. (Obviously in reality the inflow and potential evapotranspiration – especially the former – vary in an irregular manner but we deliberately assumed constant rates to simplify the example and make it fully deterministic). The actual evapotranspiration is $E \leq$ PET. We assume that a fraction f of the total plain area is covered by vegetation, and that the evapotranspiration rate in this area equals PET and in all other area is zero (assuming no route of soil water to the surface), so that in the entire plain, the average actual evapotranspiration will be

$$E = \mathrm{PET}\, f \tag{1.1}$$

It is easy to see that if $f = I / \mathrm{PET} = 0.25$ then $E = I = 250$ mm, i.e. the input equals the output and the system stays at an equilibrium; the water stored in the soil stays at a constant value. The situation becomes more interesting if at some time $f \neq 0.25$. In this case f may vary in time. It is natural to assume that f will increase if there is plenty of water stored in the soil (the vegetation will tend to expand) and to decrease otherwise. We denote s the water stored in the soil and we assume a certain reference level for which we set $s = 0$, so that $s > 0$ stands for soil water excess and $s < 0$ for soil water deficit.

Our system is described by the two state variables, the soil water s and the vegetation cover f, which can vary in time. If $i = 1, 2, \dots$ denotes time in years, then the water balance equation for our system is

$$s_i = s_{i-1} + I - \mathrm{PET}\, f_{i-1} \tag{1.2}$$

Since our system is described by two state variables, we need one more equation to fully describe its dynamics (i.e. its evolution in time). Naturally, the second equation should be sought in the dynamics of grow and decay of plants, which however may be too complicated. Here we will approach it in an extremely simplified, conceptual manner. We set a basic desideratum that f should increase when $s > 0$ and decrease otherwise. A second desideratum is the consistency with the fact that $0 \leq f \leq 1$.

Such desiderata are fulfilled by the curves shown in Fig. 1.1. The curves are described by the following equation, which takes an input x and produces an output y, depending on a parameter a that can take any real value, positive or negative:

$$y = g(x, a) := \frac{\max(1 + a, 1)x}{\max(1 - a, 1) + ax} \tag{1.3}$$

By inspection it can be verified that if $0 \leq x \leq 1$, then $0 \leq y \leq 1$, whatever the value of a is. Furthermore, it can be seen that if $a = 0$ then $y = x$, when $a > 0$ then $y > x$, and when $a < 0$ then $y < x$.

Thus, if in equation (1.3) we replace x with f_{i-1}, y with f_i, and a with some increasing function of s_{i-1} such that it takes the value 0 when $s_{i-1} = 0$, then we obtain an equation that is conceptually consistent with our desiderata. For the latter let us set $a \equiv (s_{i-1}/s^*)^3$, where s^* is a standardizing constant assumed to be $s^* = 100$ mm. Hence, the equation that completes the system dynamics becomes

$$f_i = g(f_{\mathrm{i}-1}, (s_{i-1}/s^*)^3) \tag{1.4}$$

or

$$f_i = \frac{\max(1 + (s_{i-1} / s^*)^3, 1) f_{i-1}}{\max(1 - (s_{i-1} / s^*)^3, 1) + (s_{i-1} / s^*)^3 f_{i-1}} \tag{1.5}$$

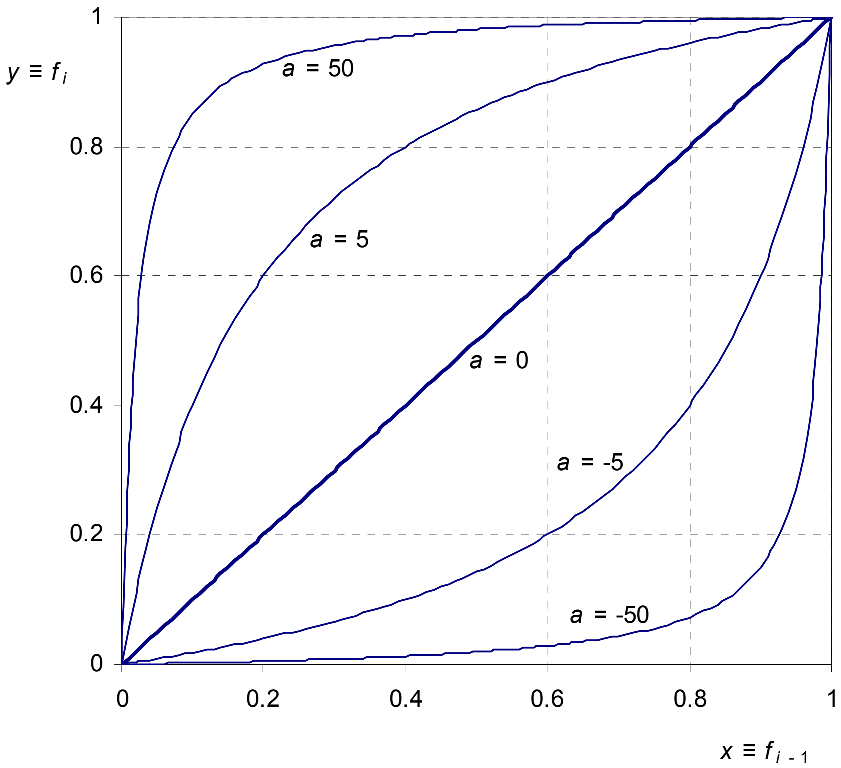

Fig. 1.1 Graphical depiction of equation (1.3) for several values of the parameter a, which is an increasing function of the soil water s.

From the system dynamics (1.2) and (1.5), it can be easily verified that if the initial conditions at time $i = 0$ are $f_0 = 0.25$, $s_0 = 0$, then the system will stay for ever at state $f_i = 0.25$, $s_i = 0$ for any time i. Now let us assume that the initial conditions depart from these conditions of stability. For instance we consider $f_0 = 0.30$, $s_0 = 100$ mm. From the system dynamics (1.2) and (1.5) we can easily find that, at time $i = 1$, $f_1 = 0.462$ (the vegetation cover was increased because of surplus water) and $s_1 = -111.5$ mm (the increased vegetation consumed more water, so that the surplus was exhausted and now there is deficit). Continuing in this manner we can calculate (f_2, s_2), (f_3, s_3) etc. It is a matter of a few minutes to set up a spreadsheet with two columns that evaluate equations (1.2) and (1.5), and calculate the system state (f_i, s_i) at time $i = 1$ to, say, 10 000, given the initial state (f_0, s_0) (homework). Fig. 1.2 depicts the first 100 values of the evolution of system state. It is observed that the system does not tend to the stable state discussed above. Rather, the vegetation cover fluctuates around 0.25 (roughly between 0 and 0.8) and the soil water fluctuates around 0 (roughly between -400 and 400 mm). These fluctuations seem to have a period of roughly 4-5 years but are not perfectly periodic.

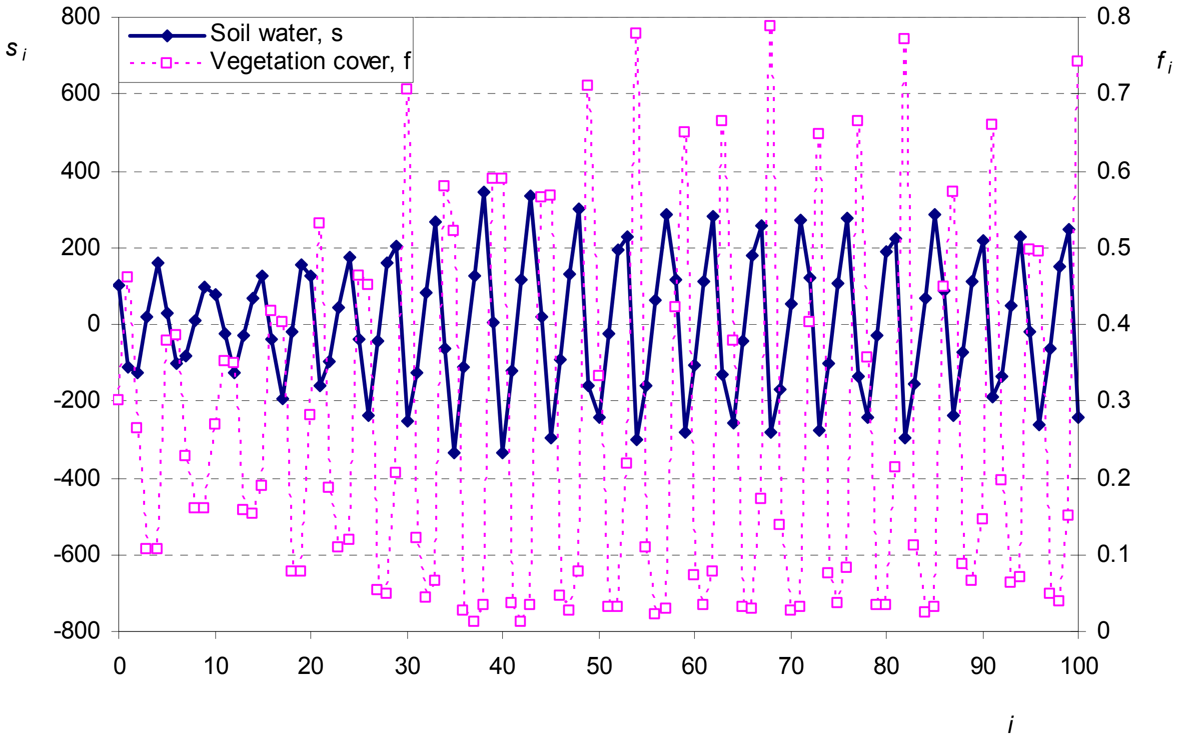

Fig. 1.2 Graphical depiction of the system evolution for time up to 100.

Despite fluctuating behaviour, it appears that we can predict exactly any future state given the initial conditions f_0, s_0. The question we wish to examine here is this: Will predictions represent reality? We can split this question to two: (1) Is our model a perfect representation of reality? (2) Is our knowledge of the initial conditions perfectly accurate? The reply to both questions should be negative. No model can represent nature perfectly; all models are just approximations. Furthermore, our knowledge of initial conditions at the best case comes from measurements and all measurements include some error or uncertainty.

Let us circumvent the first problem, and assume that our model is perfect. Put it in a different way, let us temporarily forget that the mathematical system with dynamics (1.2) and (1.5) aims to represent a natural system, so that we do not care about model errors. What is then the effect of imperfect knowledge of the initial conditions? To demonstrate this, we assume that the initial conditions set above are obtained by rounding off some true values, which introduces some small error. (We suppose that rounding off mimics the measurement error in a natural system). Our true conditions are assumed to be $f_0 = 0.2999$, $s_0 = 100.01$ mm and our approximations are $f'_0 = 0.30$, $s'_0 = 100$ mm, as above; the errors in f_0 and s_0 are -0.0001 and 0.01 mm, respectively. Repeating our calculations (with our spreadsheet) with the true conditions, we obtain a set of *true* values that are depicted in Fig. 1.3, along with the *approximate* values. By approximate we mean the values that were obtained by the rounded off initial conditions f'_0 and s'_0; these values are precisely those shown in Fig. 1.2.

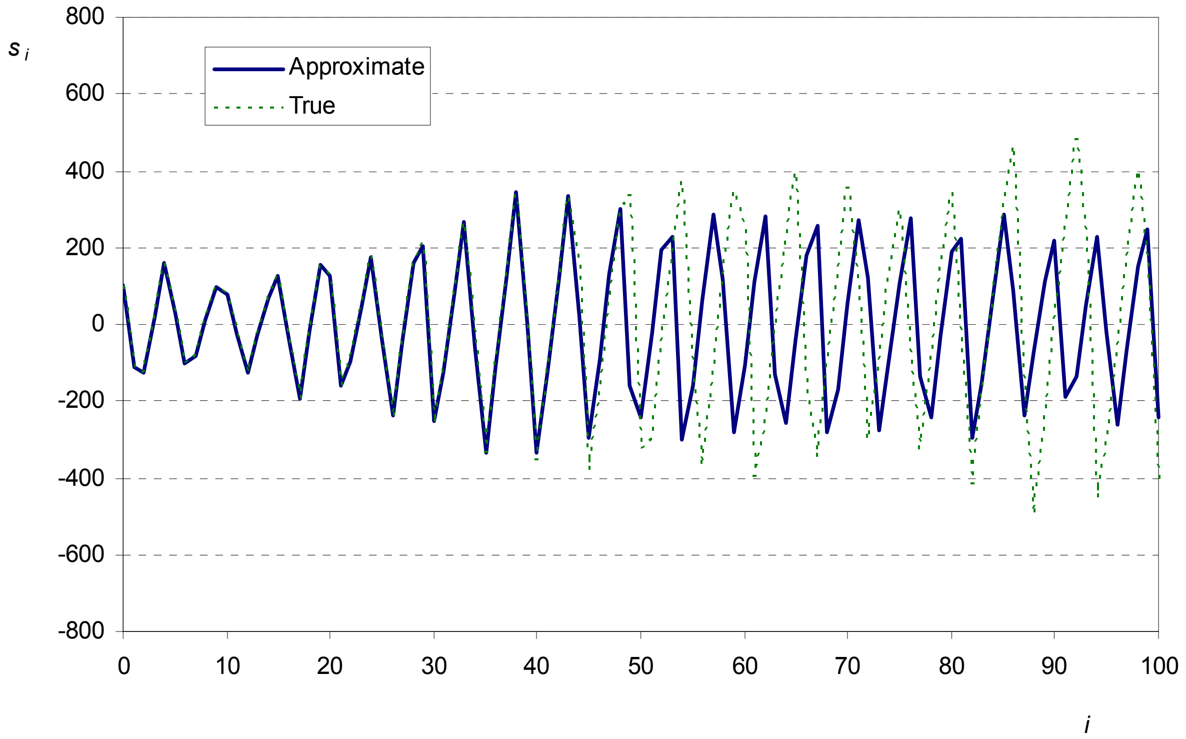

Fig. 1.3 Graphical depiction of the true system evolution and its approximation for time up to 100.

We can observe in Fig. 1.3 that the approximation is almost perfect for times up to 40 but becomes imperfect for larger times. For instance, the true value at time i = 60 is s_{60} = 245.9 mm whereas the approximate value is s'_{60} = –105.4 mm. Thus a small error of 0.01 mm in the initial conditions is magnified to 245.9 – (–105.4) = 491.8 mm in 60 time steps. Here we have used this definition of error:

$$e_i := s_i - s'_i \tag{1.6}$$

This large error clearly suggests that deterministic dynamics, even perfectly known and simple, may be unable to give deterministic future predictions for long lead times.

Nevertheless, in engineering applications it is often necessary to cast predictions for long time horizons. For instance, when we design a major project, we may have a planning horizon of say 100 years and we wish to know the behaviour of the natural system for the next 100 years. However, in most situations we are interested about the events that may occur and particularly about their magnitude while we are not interested about the exact time of occurrence. Such predictions can be obtained in a different manner, which may not need to know the deterministic dynamics of the system. Rather, it is based on the statistical properties of the system *trajectory* as reflected in a time series of the system evolution.

In the simplest case, a statistical prediction is obtained by taking the average of the time series. In our system this average of s is around 0, so that the prediction for any future time is simply s'_i = 0. As strange as it may seem, for large lead times this prediction is better (i.e.

gives a smaller error) than obtained by running the deterministic model. For instance, at time $i = 60$, $e_i = 245.9 - 0 = 245.9 < 491.8$ mm. A graphical depiction of prediction errors of both the deterministic and statistical method (where in the second method $e_i = s_i - 0 = s_i$) is shown in Fig. 1.4

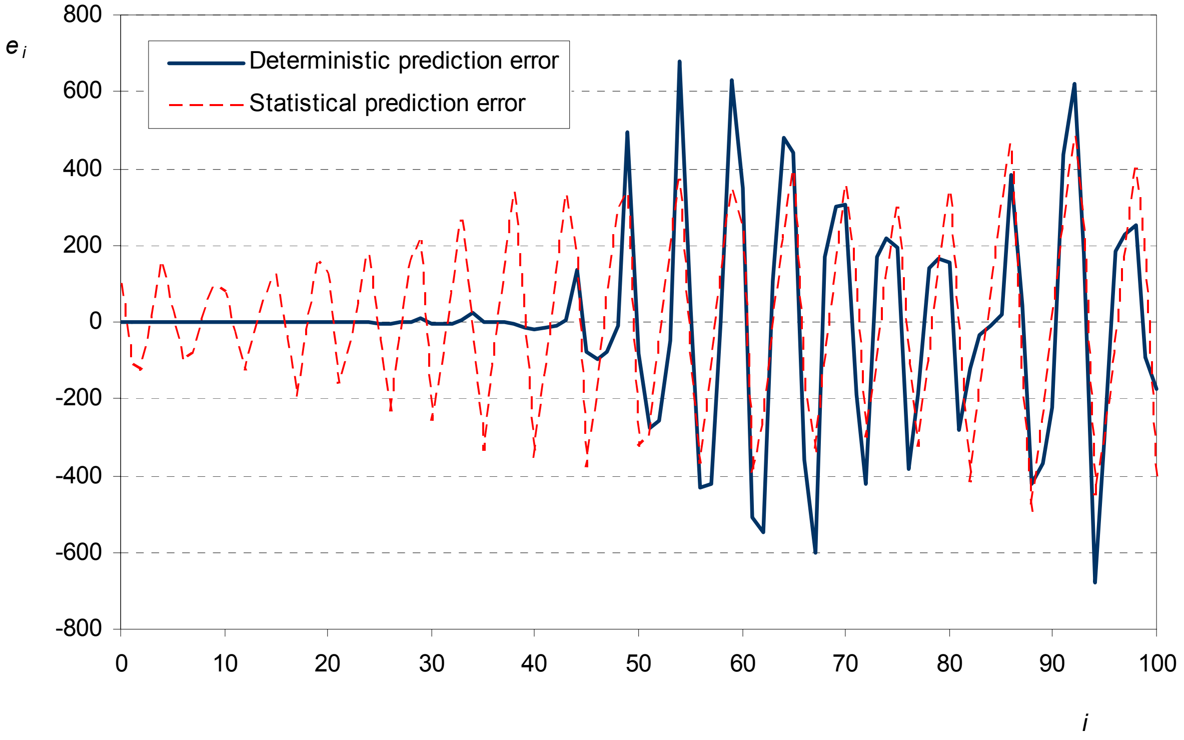

Fig. 1.4 Comparison of prediction errors of the deterministic and statistical methods for time up to 100.

We can observe that the deterministic method yields zero error for time up to 25 and negligible error for time up to 40. Then the error becomes high, fluctuating between about –800 and 800. The error of the statistical prediction fluctuates in a narrower range, between –500 and 500 mm. Statistics gives us a way to give a quantitative global measure of the error and compare the errors quantitatively rather than graphically. Thus for the n-year period $[l, l + n - 1]$ we define the root mean square (RMS) error as

$$e_{\mathrm{RMS}} := \frac{1}{n}\sqrt{\sum_{i=l}^{l+n-1} e_i^2} \tag{1.7}$$

The logic in taking the squares of errors and then summing up is to avoid an artificial cancelling up of negative and positive errors. Thus, for the last 50 years ($l = 51$, $n = 50$) the RMS error (calculated in the spreadsheet) is 342.0 and 277.9 mm, respectively. This verifies that the statistical prediction is better than the deterministic one for times > 50. On the other hand, Fig. 1.4 clearly shows that the deterministic prediction is better than the statistical for times < 50.

This happens in most real world systems, but the time horizon, up to which a deterministic prediction is reliable, varies and depends on the system dynamics. For instance, we know that a weather prediction, obtained by solving the differential equations describing the global atmospheric system dynamics, is very good for the first couple of days but is totally unreliable for more than a week or ten days lead time. After that time, statistical predictions of weather conditions, based on records of previous years for the same time of the year, are more reliable.

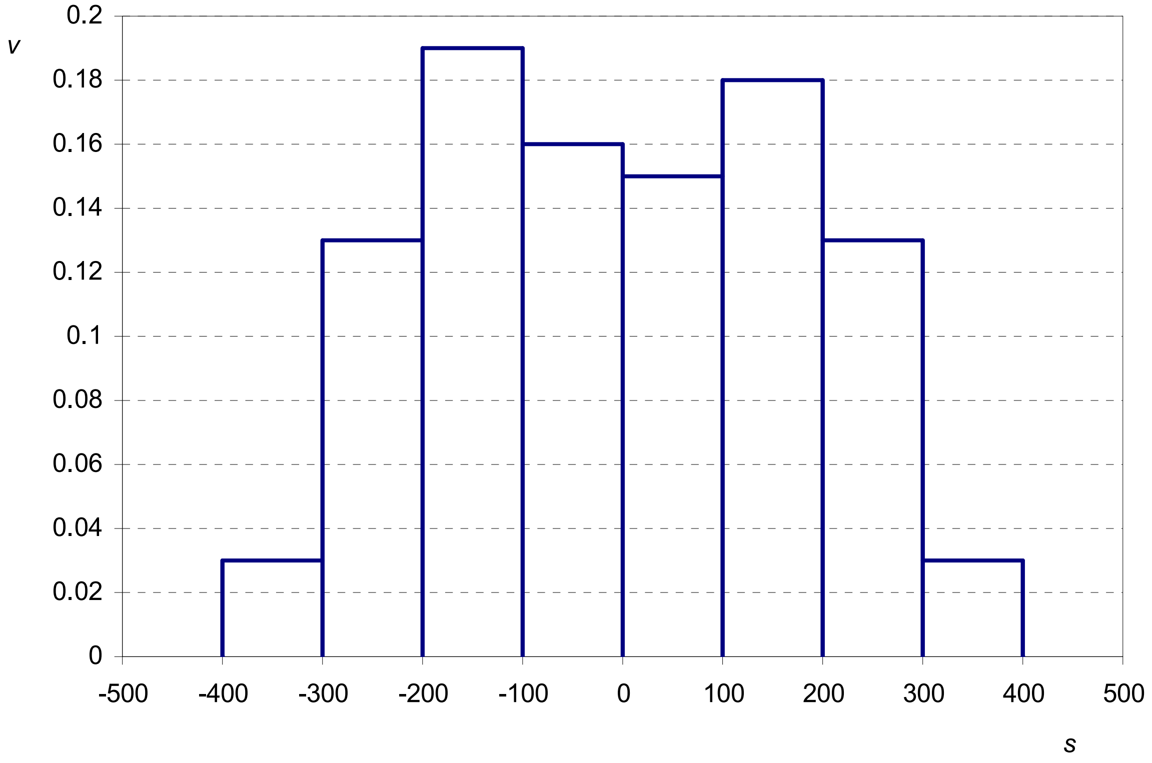

Fig. 1.5 Relative frequency *v* of the intervals of *s,* each with length 100 mm, as determined from the time series shown in Fig. 1.2.

A statistical prediction is generally more powerful than indicated in the example above. Instead of providing a single value (the value 0 in the example) that is a likely future state of the system, it can give ranges of likely values and a likelihood measure for each range. This measure is an empirical estimate of *probability* obtained by analyzing the available time series and using the theory of probability and statistics. That is to say, it is obtained by induction and not by deduction. In our example, analyzing the time series of Fig. 1.2, we can construct the *histogram* shown in Fig. 1.5, which represents empirically estimated probabilities for ranges of values of the soil water *s*. The histogram shows for instance that with probability 16%, *s* will be between –100 mm and 0, or that with probability 3%, *s* will be between –400 and –300 mm. We must be careful, however, about the validity of empirical inferences of this type. For instance, extending this logic we may conclude from Fig. 1.5 that with probability 100% the soil water will be between –400 and 400 mm. This is a mistaken conclusion: we cannot exclude values of soil water smaller than –400 mm or higher than 400 mm. The probabilities of such extreme (very low or very high) events are nonzero. To find them the empirical observations do not suffice and we need some theoretical tools, i.e. deductive reasoning. The tools are provided by the *probability theory* and the related areas of *statistics* and *stochastics*. Particularly, the latter area deals with processes that possess some dependence in time and perhaps cyclical behaviour (as happens in our example), and endeavour to incorporate any known deterministic laws within a unified, probability based, mathematical description.

1.5 Concluding remarks

If we try to make the above example more realistic, we should do several changes. Particularly: (a) the input (soil infiltration) should vary in time (and in space) in a rather irregular (random) manner; and (b) the relationship between soil water and vegetation cover should be revisited in light of some observational data in the specific area. For step (a) we need to build an additional model to simulate the input. This model should utilize infiltration data in the area, if available, or other hydrological data (rainfall, runoff) of the area; in the latter case an additional model that transforms rainfall to infiltration and runoff will be required. In all cases, the building of the model will require tools from probability, statistics, and stochastics. For step (b), which aims at establishing a deterministic relationship, it is wise to admit from the beginning the great difficulty or impossibility to establish the relationship by pure theoretical (deductive) reasoning. Usually a mixed approach is followed: (b1) a plausible (conceptual) mathematical expression is assumed that contains some parameters strongly affecting its shape; and (b2) an available time series of measurements is used to estimate its parameters. Step b2 is clearly based on a statistical/inductive approach and will always give some error; in fact the parameter estimation is done with the target to minimize (but not to eliminate) the error. This error should be modelled itself, again using tools from probability, statistics, and stochastics.

It may seem contradictory, at first glance, that in the establishment of a deterministic relationship we use statistical tools. As strange as it may seem, this happens all the time. The detection of deterministic controls, based on observed field or laboratory data, and the establishment of deterministic relationships, again based on data, is always done using tools from probability, statistics, and stochastics. A variety of such tools, all probability-based, are available: least squares estimation, Bayesian estimation, spectral analysis, time delay embedding (based on the entropy concept) and others. Here it should be added that even purely deterministic problems such as the numerical optimization of a purely deterministic non-convex function and the numerical integration of a multivariate purely deterministic function can be handled more efficiently and effectively by probability-based methods (evolutionary algorithms and Monte Carlo integration, respectively) rather than by deterministic methods.

Obviously, in a realistic setting of our example problem, the system trajectory should look more irregular than demonstrated above and the horizon for a reliable deterministic prediction should decrease significantly, perhaps to zero. In this case, a probabilistic-statistical treatment of the problem should be attempted *from the outset*, not for long horizons only. In this case we need not disregard the deterministic dynamics, if identified. On the contrary, stochastic methods are able to make explicit use of any identified deterministic control, so as to improve predictions as much as possible. That is to say, a stochastic approach *from the outset* does not deny *causality* and deterministic controls; rather it poses them in a more consistent framework admitting that uncertainty is inherent in natural systems. Here we should clarify that causality is conceptually different in a deterministic and a probabilistic approach. In the former case causality (or causation) is a directional relationship between one event (called cause) and another event (called effect), which is the consequence (result) of the first. In a stochastic view of the world, the definition of causality can be generalized in the following way (Suppes, 1970): An event A is the *prima facie* cause of an event B if and only if (i) A occurs earlier than B, (ii) A has a nonzero probability of occuring, and (iii) the conditional probability[*] of B occurring when A occurs is greater than the unconditional probability of B occurring.

It is, however, possible that in a real world problem our attempt to establish a causal relationship between our state variables fails. In a probabilistic framework this is not a tragedy, provided that we have a sufficient series of observations. We can build a model (for instance for the soil water s) without having identified the system dynamics. This is actually done in many cases of hydrological simulations.

All graphs in the above example indicate that the trajectories of the state variables of our system are irregular; simultaneously, they do not look like a purely random phenomenon, such as a series of roulette outcomes. This is very important and should be taken into serious consideration in any modelling attempt using probabilistic tools. In fact, the trajectories of natural systems never look like our more familiar purely random systems.[†] One major difference is the dependence in time, which may be very complex, contrary to the independence of roulette outcomes or to simple type of dependence (e.g. Markovian) encountered in simplistic stochastic models. In one of the next chapters we will examine these properties (revisiting the above example). We note, however, that such properties of natural processes, which seem peculiar in comparison to simple random systems, have been overlooked for years. Even worse, the standard statistical framework that was developed for independent events has been typically used in hydrological and geophysical applications, and this gave rise to erroneous results and conceptions.

Acknowledgment I thank Richard Mackey for his discussions that have influenced and fed this text.

[*] For a formal definition of conditional probability see chapter 2.

[†] We will revisit these differences in chapter 4.

References

Barrow-Green, J., Oscar II's prize competition and the error in Poincaré's memoir on the three body problem, *Archive for History of Exact Sciences*, 48(2), 107-131, 1994.

Jaynes, E.T., *Probability Theory: The Logic of Science*, Cambridge University Press, 2003.

Koutsoyiannis, D., On the quest for chaotic attractors in hydrological processes, *Hydrological Sciences Journal*, 51(6), 1065-1091, 2006.

Laviolette, M., J. W. Seaman, Jr., J.D. Barrett and W.H. Woodall, A probabilistic and statistical view of fuzzy methods, *Technometrics*, 37(3), 249-261, 1995.

Poincaré, H. Chance, *The World of Mathematics*, Simon & Schuster, New York, 1956.

Popper, K., *Quantum Theory and the Schism in Physics*, Routledge, 1982.

Suppes, P., *A Probabilistic Theory of Causality*, North-Holland, Amsterdam, 1970.

Chapter 2

Basic concepts of probability

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

Summary

This chapter aims to serve as a reminder of basic concepts of probability theory, rather than a systematic and complete presentation of the theory. The text follows Kolmogorov's axiomatic foundation of probability and defines and discusses concepts such as random variables, distribution functions, independent and dependent events, conditional probability, expected values, moments and L moments, joint, marginal and conditional distributions, stochastic processes, stationarity, ergodicity, the central limit theorem, and the normal, χ^2 and Student distributions. Although the presentation is general and abstract, several examples with analytical and numerical calculations, as well as practical discussions are given, which focus on geophysical, and particularly hydrological, processes.

2.1 Axiomatic foundation of probability theory

For the understanding and the correct use of probability, it is very important to insist on the definitions and clarification of its fundamental concepts. Such concepts may differ from other, more familiar, arithmetic and mathematical concepts, and this may create confusion or even collapse of our cognitive construction, if we do not base it in concrete fundaments. For instance, in our everyday use of mathematics, we expect that all quantities are expressed by numbers and that the relationship between two quantities is expressed by the notion of a function, which to a numerical input quantity associates (maps) another numerical quantity, a unique output. Probability too does such a mapping, but the input quantity is not a number but an event, which mathematically can be represented as a set. Probability is then a quantified likelihood that the specific event will happen. This type of representation was proposed by Kolmogorov (1956)[*]. There are other probability systems different from Kolmogorov's axiomatic system, according to which the input is not a set. Thus, in Jaynes (2003)[†] the input of the mapping is a logical proposition and probability is a quantification of the plausibility of the proposition. The two systems are conceptually different but the differences mainly rely on

[*] Here we cite the English translation, second edition, whilst the original publication was in German in 1933.

[†] Jaynes's book that we cite here was published after his death in 1998.

interpretation rather than on the mathematical results. Here we will follow Kolmogorov's system.

Kolmogorov's approach to probability theory is based on the notion of *measure*, which maps *sets* onto *numbers*. The objects of probability theory, the *events*, to which probability is assigned, are thought of as sets. For instance the outcome of a roulette spin, i.e. the pocket in which the ball eventually falls on to the wheel is one of 37 (in a European roulette – 38 in an American one) pockets numbered 0 to 36 and coloured black or red (except 0 which is coloured green). Thus all sets {0}, {1}, … {36} are events (also called elementary events). But they are not the only ones. All possible subsets of Ω, including the empty set Ø, are events. The set Ω := {0, 1, …, 36} is an event too. Because any possible outcome is contained in Ω, the event Ω occurs in any case and it is called the *certain event*. The sets ODD := {1, 3, 5, …, 35}, EVEN := {2, 4, 6, …, 36}, RED := {1, 3, 5, 7, 9, 12, 14, 16, 18, 19, 21, 23, 25, 27, 30, 32, 34, 36}, and BLACK := Ω – RED – {0} are also events (in fact, betable). While events are represented as sets, in probability theory there are some differences from set theory in terminology and interpretation, which are shown in Table 2.1.

Table 2.1 Terminology correspondence in set theory and probability theory (adapted from Kolmogorov, 1956)

Set theory	Events
$A = \varnothing$	Event A is impossible
$A = \Omega$	Event A is certain
$AB = \varnothing$ (or $A \cap B = \varnothing$; disjoint sets)	Events A and B are incompatible (mutually exclusive)
$AB...N = \varnothing$	Events A, B, …, N are incompatible
$X := AB...N$	Event X is defined as the simultaneous occurrence of A, B, …, N
$X := A + B + \ ... \ + N$ (or $X := A \cup B \cup \ ... \ \cup N$)	Event X is defined as the occurrence of at least one of the events A, B, …, N
$X := A - B$	Event X is defined as the occurrence of A and, at the same time, the non-occurrence of B
$\overline{A}$ (the complementary of A)	The opposite event $\overline{A}$ consisting of the non-occurrence of A
$B \subset A$ (B is a subset of A)	From the occurrence of event B follows the inevitable occurrence of event A

Based on Kolmogorov's (1956) axiomatization, probability theory is based on three fundamental concepts and four axioms. The concepts are:

1. A non-empty set Ω, sometimes called the *basic set*, *sample space* or the *certain event* whose elements ω are known as *outcomes* or *states*.

2. A set Σ known as *σ-algebra* or *σ-field* whose elements E are subsets of Ω, known as *events*. Ω and Ø are both members of Σ, and, in addition, (a) if E is in Σ then the complement $\Omega - E$ is in Σ; (b) the union of countably many sets in Σ is also in Σ.
3. A function P called *probability* that maps events to real numbers, assigning each event E (member of Σ) a number between 0 and 1.

The triplet (Ω, Σ, P) is called *probability space*.

The four axioms, which define properties of P, are

Non-negativity: For any event A, $P(A) \geq 0$ (2.1.I)

Normalization: $P(\Omega) = 1$ (2.1.II)

Additivity: For any events A, B with $AB = \varnothing$, $P(A + B) = P(A) + P(B)$ (2.1.III)

IV. *Continuity at zero*: If $A_1 \supset A_2 \supset \ldots \supset A_n \supset \ldots$ is a decreasing sequence of events, with $A_1A_2\ldots A_n\ldots = \varnothing$, then $\lim_{n\to\infty}P(A_n) = 0$ (2.1.IV)

In the case that Σ is finite, axiom IV follows from axioms I-III; in the general case, however, it should be put as an independent axiom.

2.2 Random variables

A random variable X is a function that maps outcomes to numbers, i.e. quantifies the sample space Ω. More formally, a real single-valued function $X(\omega)$, defined on the basic set Ω, is called a *random variable* if for each choice of a real number a the set $\{X < a\}$ for all ω for which the inequality $X(\omega) < \alpha$ holds true, belongs to Σ.

With the notion of the random variable we can conveniently express events using basic mathematics. In most cases this is done almost automatically. For instance in the roulette case a random variable X that takes values 0 to 36 is intuitively assumed when we deal with a roulette experiment.

We must be attentive that a random variable is not a number but a function. Intuitively, we could think of a random variable as an object that represents simultaneously all possible states and only them. A particular value that a random variable may take in a random experiment, else known as a *realization* of the variable is a number. Usually we denote a random variable by an upper case letter, e.g. X, and its realization by a lower case letter, e.g. x. The two should not be confused. For example, if X represents the rainfall depth expressed in millimetres for a given rainfall episode (in this case Ω is the set of all possible rainfall depths) then $\{X \leq 1\}$ represents an *event* in the probability notion (a subset of Ω and a member of Σ – not to be confused with a physical event or episode) and has a probability $P\{X \leq 1\}$.[*] If x is a realization of X then $x \leq 1$ is not an event but a relationship between the two numbers x and 1,

[*] The consistent notation here would be $P(\{X \leq 1\})$. However, we simplified it dropping the parentheses; we will follow this simplification throughout this text. Some texts follow another convention, i.e., they drop the curly brackets.

which can be either true or false. In this respect it has no meaning to write $P\{x \le 1\}$. Furthermore, if we consider the two variables X and Y it is meaningful to write $P\{X \geq Y\}$ (i.e. $\{X \geq Y\}$ represents an event) but there is no meaning in the expression $P\{x \geq y\}$.

2.3 Distribution function

Distribution function is a function of the real variable x defined by

$$F_X(x) := P\{X \leq x\} \tag{2.2}$$

where X is a random variable[*]. Clearly, $F_X(x)$ maps numbers (x values) to numbers. The random variable to which this function refers (is associated) is not an argument of the function; it is usually denoted as a subscript of F (or even omitted if there is no risk of confusion). Typically $F_X(x)$ has some mathematical expression depending on some parameters β_i. The domain of $F_X(x)$ is not identical to the range of the random variable X; rather it is always the set of real numbers. The distribution function is a non-decreasing function obeying the relationship

$$0 = F_X(-\infty) \leq F_X(x) \leq F_X(+\infty) = 1 \tag{2.3}$$

For its non-decreasing attitude, in the English literature the distribution function is also known as *cumulative distribution function* (cdf) – though cumulative is not necessary here. In hydrological applications the distribution function is also known as non-exceedence probability. Correspondingly, the quantity

$$F_X^*(x) := P\{X > x\} = 1 - F_X(x) \tag{2.4}$$

is known as exceedence probability, is a non-increasing function and obeys

$$1 = F_X^*(-\infty) \geq F_X^*(x) \geq F_X^*(+\infty) = 0 \tag{2.5}$$

The distribution function is always continuous on the right; however, if the basic set Ω is finite or countable, $F_X(x)$ is discontinuous on the left at all points x_i that correspond to outcomes ω_i, and it is constant in between consecutive points. In other words, the distribution function in these cases is staircase-like and the random variable is called *discrete*. If $F_X(x)$ is continuous, then the random variable is called *continuous*. A *mixed* case with a continuous part and a discrete part is also possible. In this case the distribution function has some discontinuities on the left, without being staircase-like.

The derivative of the distribution function

$$f_X(x) := \frac{dF(x)}{dx} \tag{2.6}$$

[*] In original Kolmogorov's writing $F_X(x)$ is defined as $P\{X < x\}$; however replacing '<' with '≤' makes the handling of distribution function more convenient and has prevailed in later literature.

is called the *probability density function* (sometimes abbreviated as pdf). In continuous variables, this function is defined everywhere but this is not the case in discrete variables, unless we use Dirac's δ functions. The basic properties of $f_X(x)$ are

$$f_X(x) \ge 0, \quad \int_{-\infty}^{\infty} f_X(x)dx = 1 \tag{2.7}$$

Obviously, the probability density function does not represent a probability; therefore it can take values higher than 1. Its relationship with probability is described by the following equation:

$$f_X(x) = \lim_{\Delta x \to 0} \frac{P\{x \le X \le x + \Delta x\}}{\Delta x} \tag{2.8}$$

The distribution function can be calculated from the density function by the following relationship, inverse of (2.6)

$$F_X(x) = \int_{-\infty}^{x} f_X(\xi)d\xi \tag{2.9}$$

For continuous random variables, the inverse function F_X^{-1} of $F_X(x)$ exists. Consequently, the equation $u = F_X(x)$ has a unique solution for x, that is $x_u = F_X^{-1}(u)$. The value x_u, which corresponds to a specific value u of the distribution function, is called *u-quantile* of the variable X.

2.3.1 An example of the basic concepts of probability

For clarification of the basic concepts of probability theory, we give the following example from hydrology. We are interested on the mathematical description of the possibilities that a certain day in a specific place and time of the year is wet or dry. These are the outcomes or states of our problem, so the basic set or sample space is

$$\Omega = \{\text{wet, dry}\}$$

The field Σ contains all possible events, i.e.,

$$\Sigma = \{\varnothing, \{\text{wet}\}, \{\text{dry}\}, \Omega\}$$

To fully define probability on Σ it suffices to define the probability of one of either states, say P(wet). In fact this is not easy – usually it is done by induction, and it needs a set of observations to be available and concepts of the *statistics* theory (see chapter 3) to be applied. For the time being let us arbitrarily assume that $P\{\text{wet}\} = 0.2$. The remaining probabilities are obtained by applying the axioms. Clearly, $P(\Omega) = 1$ and $P(\varnothing) = 0$. Since "wet" and "dry" are incompatible, $P\{\text{wet}\} + P\{\text{dry}\} = P(\{\text{wet}\} + \{\text{dry}\}) = P(\Omega) = 1$, so $P\{\text{dry}\} = 0.8$.

We define a random variable X based on the rule

$$X(\text{dry}) = 0, \quad X(\text{wet}) = 1$$

We can now easily determine the distribution function of X. For any $x < 0$,

$$F_X(x) = P\{X \le x\} = 0$$

(because X, cannot take negative values). For $0 \le x < 1$,

$$F_X(x) = P\{X \le x\} = P\{X = 0\} = 0.8$$

Finally, for $x \ge 1$,

$$F_X(x) = P\{X \le x\} = P\{X = 0\} + P\{X = 1\} = 1$$

The graphical depiction of the distribution function is shown on Fig. 2.1. The staircase-like shape reflects the fact that random variable is discrete.

If this mathematical model is to represent a physical phenomenon, we must have in mind that all probabilities depend on a specific location and a specific time of the year. So the model cannot be a global representation of the wet and dry state of a day. The model as formulated here is extremely simplified, because it does not make any reference to the succession of dry or wet states in different days. This is not an error; it simply diminishes the predictive capacity of the model. A better model would describe separately the probability of a wet day following a wet day, a wet day following a dry day (we anticipate that the latter should be smaller than the former), etc. We will discuss this case in section 2.4.2.

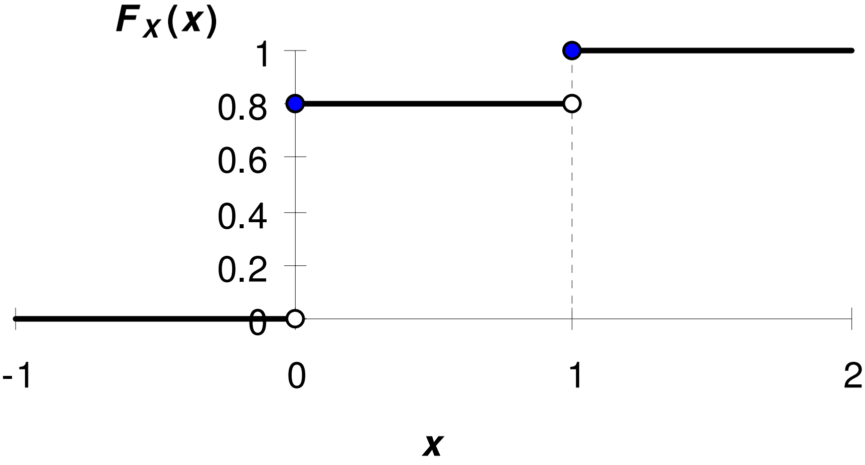

Fig. 2.1 Distribution function of a random variable representing the dry or wet state of a given day at a certain area and time of the year.

2.4 Independent and dependent events, conditional probability

Two events A and B are called *independent* (or *stochastically independent*), if

$$P(AB) = P(A)P(B) \tag{2.10}$$

Otherwise A and B are called (*stochastically*) *dependent*. The definition can be extended to many events. Thus, the events A_1, A_2, …, are *independent* if

$$P(A_{i_1} A_{i_2} \cdots A_{i_n}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_n}) \tag{2.11}$$

for any finite set of distinct indices i_1, i_2, …, i_n.

The handling of probabilities of independent events is thus easy. However, this is a special case because usually natural events are dependent. In the handling of dependent events the notion of *conditional probability* is vital. By definition (Kolmogorov, 1956), conditional probability of the event A given B (i.e. under the condition that the event B has occurred) is the quotient

$$P(A \mid B) := \frac{P(AB)}{P(B)} \tag{2.12}$$

Obviously, if $P(B) = 0$, this conditional probability cannot be defined, while for independent A and B, $P(A|B) = P(A)$. From (2.12) it follows that

$$P(AB) = P(A \mid B)P(B) = P(B \mid A)P(A) \tag{2.13}$$

and

$$P(B \mid A) := P(B)\frac{P(A \mid B)}{P(A)} \tag{2.14}$$

The latter equation is known as the *Bayes theorem*. It is easy to prove that the generalization of (2.11) for dependent events takes the forms

$$P(A_n \cdots A_1) = P(A_n \mid A_{n-1} \cdots A_1) \cdots P(A_2 \mid A_1)P(A_1) \tag{2.15}$$

$$P(A_n \cdots A_1 \mid B) = P(A_n \mid A_{n-1} \cdots A_1 B) \cdots P(A_2 \mid A_1 B)P(A_1 \mid B) \tag{2.16}$$

which are known as the *chain rules*. It is also easy to prove (homework) that if A and B are mutually exclusive, then

$$P(A + B \mid C) = P(A \mid C) + P(B \mid C) \tag{2.17}$$

$$P(C \mid A + B) = \frac{P(C \mid A)P(A) + P(C \mid B)P(B)}{P(A) + P(B)} \tag{2.18}$$

2.4.1 Some examples on independent events

a. Based on the example of section 2.3.1, calculate the probability that two consecutive days are wet assuming that the events in the two days are independent.

Let A := {wet} the event that a day is wet and $\overline{A}$ = {dry} the complementary event that a day is dry. As in section 2.3.1 we assume that $p := P(A) = 0.2$ and $q := P(\overline{A}) = 0.8$. Since we are interested on two consecutive days, our basic set will be

$$\Omega = \left\{A_1 A_2, \overline{A}_1 A_2, A_1 \overline{A}_2, \overline{A}_1 \overline{A}_2\right\}$$

where indices 1 and 2 correspond to the first and second day, respectively. By the independence assumption, the required probability will be

$$P_1 := (A_1 A_2) = P(A_1)P(A_2) = p^2 = 0.04$$

For completeness we also calculate the probabilities of all other events, which are:

$$P(\overline{A}_1 A_2) = P(A_1 \overline{A}_2) = pq = 0.16, \quad P(\overline{A}_1 \overline{A}_2) = q^2 = 0.64$$

As anticipated, the sum of probabilities of all events is 1.

b. Calculate the probability that two consecutive days are wet if it is known that one day is wet.

Knowing that one day is wet means that the event $\overline{A}_1\overline{A}_2$ should be excluded (has not occurred) or that the composite event $A_1A_2 + \overline{A}_1A_2 + A_1\overline{A}_2$ has occurred. Thus, we seek the probability

$$P_2 := P(A_1A_2 \mid A_1A_2 + \overline{A}_1A_2 + A_1\overline{A}_2)$$

which according to the definition of conditional probability is

$$P_2 = \frac{P\left(A_1A_2\left(A_1A_2 + \overline{A}_1A_2 + A_1\overline{A}_2\right)\right)}{P\left(A_1A_2 + \overline{A}_1A_2 + A_1\overline{A}_2\right)}$$

Considering that all combinations of events are mutually exclusive, we obtain

$$P_2 = \frac{P(A_1A_2)}{P(A_1A_2) + P(\overline{A}_1A_2) + P(A_1\overline{A}_2)} = \frac{p^2}{p^2 + 2pq} = \frac{p}{p + 2q} = 0.111\ldots$$

c. Calculate the probability that two consecutive days are wet if it is known that the first day is wet

Even though it may seem that this question is identical to the previous one, in fact it is not. In the previous question we knew that one day is wet, without knowing which one exactly. Here we have additional information, that the wet day is the first one. This information alters the probabilities as we will verify immediately.

Now we know that the composite event $A_1A_2 + A_1\overline{A}_2$ has occurred (events $\overline{A}_1A_2$ and $\overline{A}_1\overline{A}_2$ should be excluded). Consequently, the probability sought is

$$P_3 := P(A_1A_2 \mid A_1A_2 + A_1\overline{A}_2)$$

which according to the definition of conditional probability is

$$P_3 = \frac{P\left(A_1A_2\left(A_1A_2 + A_1\overline{A}_2\right)\right)}{P\left(A_1A_2 + A_1\overline{A}_2\right)}$$

or

$$P_3 = \frac{P(A_1A_2)}{P(A_1A_2) + P(A_1\overline{A}_2)} = \frac{p^2}{p^2 + pq} = \frac{p}{p + q} = p = 0.2$$

It is not a surprise that this is precisely the probability that one day is wet, as in section 2.3.1.

With these examples we demonstrated two important thinks: (a) that the prior information we have in a problem may introduce dependences in events that are initially assumed independent, and, more generally, (b) that the probability is not an objective and invariant quantity, characteristic of physical reality, but a quantity that depends on our knowledge or information on the examined phenomenon. This should not seem strange as it is always the case in science. For instance the location and velocity of a moving particle are not absolute objective quantities; they depend on the observer's coordinate system. The dependence of probability on given information, or its "subjectivity" should not be taken as ambiguity; there was nothing ambiguous in calculating the above probabilities, based on the information given each time.

2.4.2 An example on dependent events

The independence assumption in problem 2.4.1a is obviously a poor representation of the physical reality. To make a more realistic model, let us assume that the probability of today being wet (A_2) or dry $\overline{A}_2$ depend on the state yesterday (A_1 or $\overline{A}_1$). It is reasonable to assume that the following inequalities hold:

$$P(A_2 \mid A_1) > P(A_2) = p\,,\ P(\overline{A}_2 \mid \overline{A}_1) > P(\overline{A}_2) = q$$

$$P(A_2 \mid \overline{A}_1) < P(A_2) = p\,,\ P(\overline{A}_2 \mid A_1) < P(\overline{A}_2) = q$$

The problem now is more complicated than before. Let us arbitrarily assume that

$$P(A_2 \mid A_1) = 0.40, \quad P(A_2 \mid \overline{A}_1) = 0.15$$

Since

$$P(A_2 \mid A_1) + P(\overline{A}_2 \mid A_1) = 1$$

we can calculate

$$P(\overline{A}_2 \mid A_1) = 1 - P(A_2 \mid A_1) = 0.60$$

Similarly,

$$P(\overline{A}_2 \mid \overline{A}_1) = 1 - P(A_2 \mid \overline{A}_1) = 0.85$$

As the event $A_1 + \overline{A}_1$ is certain (i.e. $P(A_1 + \overline{A}_1) = 1$) we can write

$$P(A_2) = P(A_2 \mid A_1 + \overline{A}_1)$$

and using (2.18) we obtain

$$P(A_2) = P(A_2 \mid A_1)P(A_1) + P(A_2 \mid \overline{A}_1)P(\overline{A}_1) \tag{2.19}$$

It is reasonable to assume that the unconditional probabilities do not change after one day, i.e. that $P(A_2) = P(A_1) = p$ and $P(\overline{A}_2) = P(\overline{A}_1) = q = 1 - p$. Thus, (2.19) becomes

$$p = 0.40\,p + 0.15\,(1 - p)$$

from which we find p = 0.20 and q = 0.80. (Here we have deliberately chosen the values of $P(A_2 \mid A_1)$ and $P(A_2 \mid \overline{A}_1)$ such as to find the same p and q as in 2.4.1a).

Now we can proceed to the calculation of the probability that both days are wet:

$$P(A_2 A_1) = P(A_2 \mid A_1)P(A_1) = 0.4 \times 0.2 = 0.08 > p^2 = 0.04$$

For completeness we also calculate the probabilities of all other events, which are:

$$P(A_2 \overline{A}_1) = P(A_2 \mid \overline{A}_1)P(\overline{A}_1) = 0.15 \times 0.80 = 0.12\text{,}\ P(\overline{A}_2 A_1) = P(\overline{A}_2 \mid A_1)P(A_1) = 0.60 \times 0.20 = 0.12$$

$$P(\overline{A}_2 \overline{A}_1) = P(\overline{A}_2 \mid \overline{A}_1)P(\overline{A}_1) = 0.85 \times 0.80 = 0.68 > q^2 = 0.64$$

Thus, the dependence resulted in higher probabilities of consecutive events that are alike. This corresponds to a general natural behaviour that is known as *persistence* (see also chapter 4).

2.5 Expected values and moments

If X is a continuous random variable and $g(X)$ is an arbitrary function of X, then we define as the *expected value* or *mean* of $g(X)$ the quantity

$$E[g(X)] := \int_{-\infty}^{\infty} g(x) f_X(x) dx \tag{2.20}$$

Correspondingly, for a discrete random variable X, taking on the values x_1, x_2, …,

$$E[g(X)] := \sum_{i=1}^{\infty} g(x_i) P(X = x_i) \tag{2.21}$$

For certain types of functions $g(X)$ we take very commonly used statistical parameters, as specified below:

1. For $g(X) = X^r$, where $r = 0, 1, 2, \ldots$, the quantity

$$m_X^{(r)} := E[X^r] \tag{2.22}$$

   is called the r*th moment* (or the r*th moment about the origin*) of X. For $r = 0$, obviously the moment is 1.

2. For $g(X) = X$, the quantity

$$m_X := E[X] \tag{2.23}$$

   (that is the first moment) is called the *mean* of X. An alternative, commonly used, symbol for $E[X]$ is μ_X.

3. For $g(X) = (X - m_X)^r$, where $r = 0, 1, 2, \ldots$, the quantity

$$\mu_X^{(r)} := E\left[(X - m_X)^r\right] \tag{2.24}$$

   is called the r*th central moment* of X. For $r = 0$ and 1 the central moments are respectively 1 and 0. The central moments are related to the moments about the origin by

$$\mu_X^{(r)} = m_X^{(r)} - \binom{r}{1} m_X^{(r-1)} m_X + \cdots + (-1)^j \binom{r}{j} m_X^{(r-j)} m_X^j + \cdots (-1)^r m_X^{(0)} m_X^r \tag{2.25}$$

These take the following forms for small *r*

$$\mu_X^{(2)} = m_X^{(2)} - m_X^2$$

$$\mu_X^{(3)} = m_X^{(3)} - 3m_X^{(2)}m_X + 2m_X^3 \tag{2.26}$$

$$\mu_X^{(4)} = m_X^{(4)} - 4m_X^{(3)}m_X + 6m_X^{(2)}m_X^2 - 3m_X^4$$

and can be inverted to read:

$$m_X^{(2)} = \sigma_X^2 + m_X^2$$

$$m_X^{(3)} = \mu_X^{(3)} + 3\sigma_X^2 m_X + m_X^3 \tag{2.27}$$

$$m_X^{(4)} = \mu_X^{(4)} + 4\mu_X^{(3)}m_X + 6\sigma_X^2 m_X^2 + m_X^4$$

4. For $g(X) = (X - m_X)^2$, the quantity

$$\sigma_X^2 := \mu_X^{(2)} = E\left[(X - m_X)^2\right] = E[X^2] - m_X^2 \tag{2.28}$$

(that is the second central moment) is called the *variance* of *X*. The variance is also denoted as $\operatorname{Var}[X]$. Its square root, denoted as σ_X or StD[*X*] is called the standard deviation of *X*.

The above families of moments are the classical ones having been used for more than a century. More recently, other types of moments have been introduced and some of them have been already in wide use in hydrology. We will discuss two families.

5. For $g(X) = X\,[F(X)]^r$, where $r = 0, 1, 2, \ldots$, the quantity

$$\beta_X^{(r)} := E\{X\,[F(X)]^r\} = \int_{-\infty}^{\infty} x\,[F(x)]^r f(x)\,dx = \int_0^1 x(u)\,u^r\,du \tag{2.29}$$

is called the r*th probability weighted moment* of *X* (Greenwood et al., 1979). All probability weighted moments have dimensions identical to those of *X* (this is not the case in the other moments described earlier).

6. For $g(X) = X\,P_{r-1}^*(F(X))$, where $r = 1, 2, \ldots$, $P_r^*(u)$ is the *r*th shifted Legendre polynomial, i.e.,

$$P_r^*(u) := \sum_{k=0}^{r} p_{r,k}^* u^k \text{ with } p_{r,k}^* := (-1)^{r-k}\binom{r}{k}\binom{r+k}{k} = \frac{(-1)^{r-k}(r+k)!}{(k!)^2(r-k)!}$$

the quantity

$$\lambda_X^{(r)} := E[X P_{r-1}^*(F(X))] = \int_0^1 x(u)\;P_{r-1}^*(u)\,du \tag{2.30}$$

is called the r*th L moment* of X (Hosking, 1990). Similar to the probability weighted moments, the L moments have dimensions identical to those of X. The L moments are related to the probability weighted moments by

$$\lambda_X^{(r)} := \sum_{k=0}^{r-1} p^*_{r,k} \ \beta_X^{(r)} \tag{2.31}$$

which for the most commonly used r takes the specific forms

$$\begin{aligned}
\lambda_X^{(1)} &= \beta_X^{(0)} \ (= m_X) \\
\lambda_X^{(2)} &= 2\ \beta_X^{(1)} - \beta_X^{(0)} \\
\lambda_X^{(3)} &= 6\beta_X^{(2)} - 6\beta_X^{(1)} + \beta_X^{(0)} \\
\lambda_X^{(4)} &= 20\beta_X^{(3)} - 30\beta_X^{(2)} + 12\ \beta_X^{(1)} - \beta_X^{(0)}
\end{aligned} \tag{2.32}$$

In all above quantities the index X may be omitted if there is no risk of confusion. The first four moments, central moments and L moments are widely used in hydrological statistics as they have a conceptual or geometrical meaning easily comprehensible. Specifically, they describe the location, dispersion, skewness and kurtosis of the distribution as it is explained below. Alternatively, other statistical parameters with similar meaning are also used, which are also explained below.

2.5.1 Location parameters

Essentially, the mean describes the location of the centre of gravity of the shape defined by the probability density function and the horizontal axis (Fig. 2.2a). It is also equivalent with the static moment of this shape about the vertical axis (given that the area of the shape equals 1). Often, the following types of location parameters are also used:

1. The *mode*, or most probable value, x_p, is the value of x for which the density $f_X(x)$ becomes maximum, if the random variable is continuous, or, for discrete variables, the probability becomes maximum. If $f_X(x)$ has one, two or many maxima, we say that the distribution is unimodal, bi-modal or multi-modal, respectively.

2. The *median*, $x_{0.5}$, is the value for which $P\{X \le x_{0.5}\} = P\{X \ge x_{0.5}\} = 1/2$, if the random variable is continuous (analogously we can define it for a discrete variable). Thus, a vertical line at the median separates the shape of the density function in two equivalent parts each having an area of 1/2.

Generally, the mean, the mode and the median are not identical unless the density is has a symmetrical and unimodal shape.

2.5.2 Dispersion parameters

The variance of a random variable and its square root, the standard deviation, which has same dimensions as the random variable, describe a measure of the scatter or dispersion of the probability density around the mean. Thus, a small variance shows a concentrated distribution (Fig. 2.2b). The variance cannot be negative. The lowest possible value is zero and this corresponds to a variable that takes one value only (the mean) with absolute certainty. Geometrically it is equivalent with the moment of inertia about the vertical axis passing from the centre of gravity of the shape defined by the probability density function and the horizontal axis.

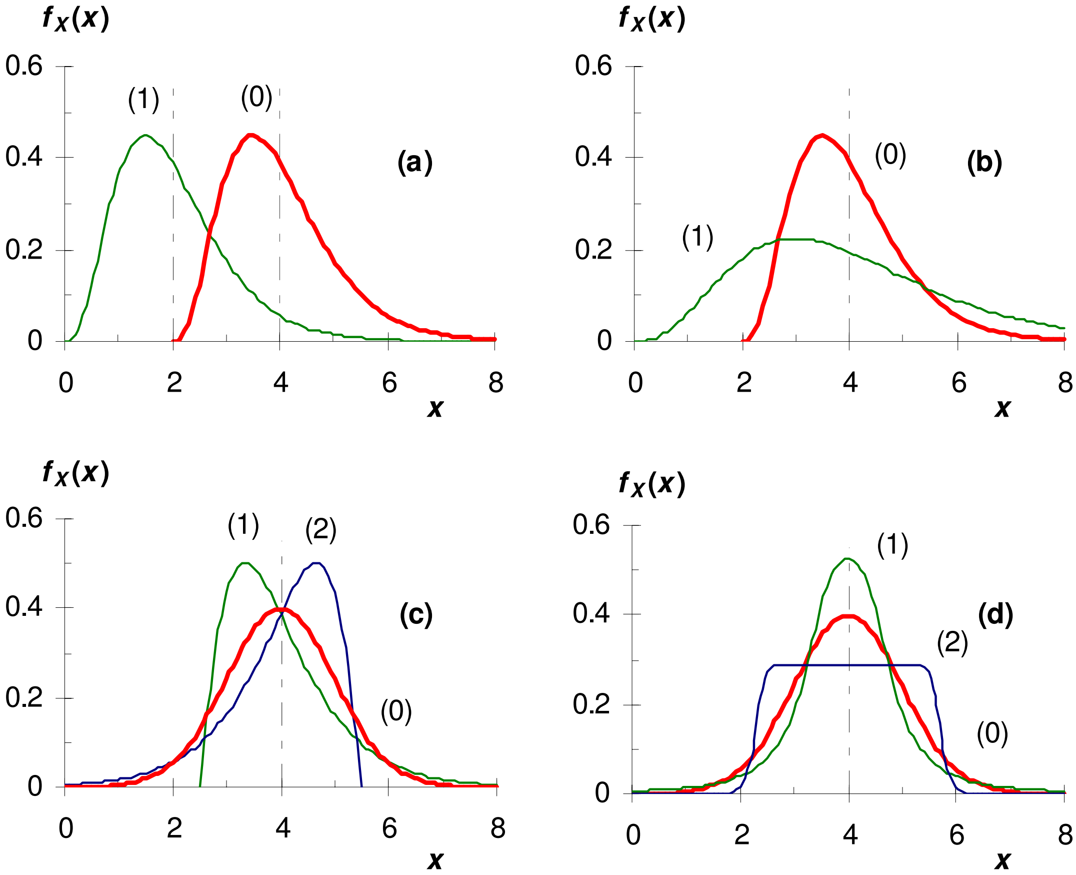

Fig. 2.2 Demonstration of the shape characteristics of the probability density function in relation to various parameters of the distribution function: *(a) Effect of the mean.* Curves (0) and (1) have means 4 and 2, respectively, whereas they both have standard deviation 1, coefficient of skewness 1 and coefficient of kurtosis 4.5. *(b) Effect of the standard deviation.* Curves (0) and (1) have standard deviation 1 and 2 respectively, whereas they both have mean 4, coefficient of skewness 1 and coefficient of kurtosis 4.5. *(c) Effect of the coefficient of skewness.* Curves (0), (1) and (2) have coefficients of skewness 0, +1.33 and -1.33, respectively, but they all have mean 4 and standard deviation 1; their coefficients of kurtosis are 3, 5.67 and 5.67, respectively. *(d) Effect of the coefficient of kurtosis.* Curves (0), (1) and (2) have coefficients of kurtosis 3, 5 and 2, respectively, whereas they all have mean 4, standard deviation 1 and coefficient of skewness 0.

Alternative measures of dispersion are provided by the so-called interquartile range, defined as the difference $x_{0.75} - x_{0.25}$, i.e. the difference of the 0.75 and 0.25 quantiles (or upper and lower quartiles) of the random variable (they define an area in the density function equal to 0.5), as well as the second L moment. This is well justified as it can be shown that the second L moment is the expected value of the difference between any two random realizations of the random variable.

If the random variable is positive, as happens with most hydrological variables, two dimensionless parameters are also used as measures of dispersion. These are called the *coefficient of variation* and the *L coefficient of variation*, and are defined, respectively, by:

$$C_{\mathrm{v}_X} := \frac{\sigma_X}{m_X}, \quad \tau_X^{(2)} := \frac{\lambda_X^{(2)}}{m_X} \tag{2.33}$$

2.5.3 Skewness parameters

The third central moment and the third L moment are used as measures of skewness. A zero value indicates that the density is symmetric. This can be easily verified from the definition of the third central moment. Furthermore, the third L moment indicates the expected value of the difference between the middle of three random realizations of a random variable from the average of the other two values (the smallest and the largest); more precisely the third central moment is the 2/3 of this expected value. Clearly then, in a symmetric distribution the distances of the middle value to the smallest and largest ones will be equal to each other and thus the third L moment will be zero. If the third central or L moment is positive or negative, we say that the distribution is positively or negatively skewed respectively (Fig. 2.2c). In a positively skewed unimodal distribution the following inequality holds: $x_\mathrm{p} \le x_{0.5} \le m_X$; the reverse holds for a negatively skewed distribution. More convenient measures of skewness are the following dimensionless parameters, named the *coefficient of skewness* and the *L coefficient of skewness*, respectively:

$$C_{\mathrm{s}_X} := \frac{\mu_X^{(3)}}{\sigma_X^3}, \quad \tau_X^{(3)} := \frac{\lambda_X^{(3)}}{\lambda_X^{(2)}} \tag{2.34}$$

2.5.4 Kurtosis parameters

The term kurtosis describes the "peakedness" of the probability density function around its mode. Quantification of this property provide the following dimensionless coefficients, based on the fourth central moment and the fourth L moment, respectively:

$$C_{\mathrm{k}_X} := \frac{\mu_X^{(4)}}{\sigma_X^4}, \quad \tau_X^{(4)} := \frac{\lambda_X^{(4)}}{\lambda_X^{(2)}} \tag{2.35}$$

These are called the c*oefficient of kurtosis* and the *L coefficient of kurtosis*. Reference values for kurtosis are provided by the normal distribution (see section 2.10.2), which has C_{k_X} = 3 and $\tau_X^{(4)}$ = 0.1226. Distributions with kurtosis greater than the reference values are called *leptokurtic* (acute, sharp) and have typically fat tails, so that more of the variance is due to infrequent extreme deviations, as opposed to frequent modestly-sized deviations. Distributions with kurtosis less than the reference values are called *platykurtic* (flat; Fig. 2.2d).

2.5.5 A simple example of a distribution function and its moments

We assume that the daily rainfall depth during the rain days, *X*, expressed in mm, for a certain location and time period, can be modelled by the *exponential distribution*, i.e.,

$$F_X(x) = 1 - e^{-x/\lambda}, \quad x \ge 0$$

where λ = 20 mm. We will calculate the location, dispersion, skewness and kurtosis parameters of the distribution.

Taking the derivative of the distribution function we calculate the probability density function:

$$f_X(x) = (1/\lambda)e^{-x/\lambda}, \quad x \ge 0$$

Both the distribution and the density functions are plotted in Fig. 2.3. To calculate the mean, we apply (2.20) for $g(X) = X$:

$$m_X = E[X] = \int_{-\infty}^{\infty} x f_X(x)dx = (1/\lambda)\int_{0}^{\infty} x e^{-x/\lambda} dx$$

After algebraic manipulations:

$$m_X = \lambda = 20\,\text{mm}$$

In a similar manner we find that for any $r \ge 0$

$$m_X^{(r)} = E[X^r] = r!\lambda^r$$

and finally, applying (2.26)

$$\sigma_X^2 = \lambda^2 = 400\,\text{mm}^2, \mu_X^{(3)} = 2\lambda^3 = 16000\,\text{mm}^3$$

$$\mu_X^{(4)} = 9\lambda^4 = 1440000\,\text{mm}^4$$

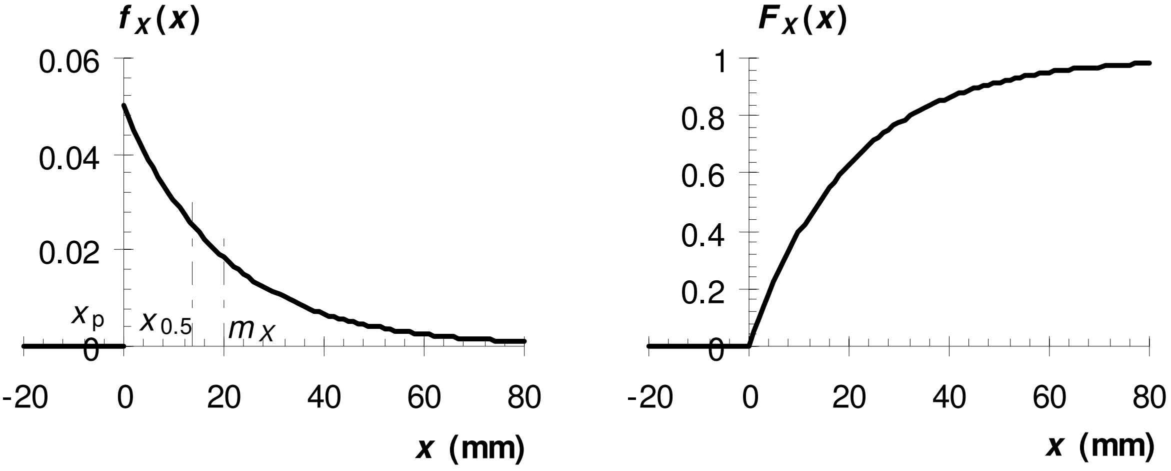

Fig. 2.3 Probability density function and probability distribution function of the exponential distribution, modelling the daily rainfall depth at a hypothetical site and time period.

The mode is apparently zero (see Fig. 2.3). The inverse of the distribution function is calculated as follows:

$$F_X(x_u) = u \rightarrow 1 - e^{-x_u/\lambda} = u \rightarrow x_u = -\lambda \ln(1-u)$$

Thus, the median is $x_{0.5} = -20 \times \ln 0.5 = 13.9\,\text{mm}$. We verify that the inequality $x_p \le x_{0.5} \le m_X$, which characterizes positively skewed distributions, holds.

The standard deviation is σ_X = 20 mm and the coefficient of variation C_{v_X} = 1. This is a very high value indicating high dispersion.

The coefficient of skewness is calculated for (2.34):

$$C_{s_X} = 2\lambda^3 / \lambda^3 = 2$$

This verifies the positive skewness of the distribution, as also shown in Fig. 2.3. More specifically, we observe that the density function has an inverse-J shape, in contrast to other, more familiar densities (e.g. in Fig. 2.2) that have a bell-shape.

The coefficient of kurtosis is calculated from (2.35):

$$C_{k_X} = 9\lambda^4 / \lambda^4 = 9$$

Its high value shows that the distribution is leptokurtic, as also depicted in Fig. 2.3.

We proceed now in the calculations of probability weighted and L moments as well as other parameters based on these. From (2.29) we find

$$\beta_X^{(r)} = \int_0^1 x(u)\, u^r\, du = -\lambda \int_0^1 \ln(1-u)\, u^r\, du = \frac{\lambda}{r+1} \sum_{i=1}^{r+1} \frac{1}{i} \tag{2.36}$$

(This was somewhat tricky to calculate). This results in

$$\beta_X^{(0)} = \lambda,\ \ \beta_X^{(1)} = \frac{3\lambda}{4},\ \ \beta_X^{(2)} = \frac{11\lambda}{18},\ \ \beta_X^{(3)} = \frac{25\lambda}{48} \tag{2.37}$$

Then, from (2.32) we find the first four L moments and the three L moment dimensionless coefficients as follows:

$$\lambda_X^{(1)} = \lambda = 20\text{ mm } (= m_X)$$

$$\lambda_X^{(2)} = 2\frac{3\lambda}{4} - \lambda = \frac{\lambda}{2} = 10\text{ mm}$$

$$\lambda_X^{(3)} = 6\frac{11\lambda}{18} - 6\frac{3\lambda}{4} + \lambda = \frac{\lambda}{6} = 3.33\text{ mm}$$

$$\lambda_X^{(4)} = 20\frac{25\lambda}{48} - 30\frac{11\lambda}{18} + 12\frac{3\lambda}{4} - \lambda = \frac{\lambda}{12} = 1.67\text{ mm}$$

$$\tau_X^{(2)} = \frac{\lambda_X^{(2)}}{\lambda_X^{(1)}} = \frac{1}{2} = 0.5,\ \ \tau_X^{(3)} = \frac{\lambda_X^{(3)}}{\lambda_X^{(2)}} = \frac{1}{3} = 0.333,\ \ \tau_X^{(4)} = \frac{\lambda_X^{(4)}}{\lambda_X^{(2)}} = \frac{1}{6} = 0.\ 167$$

Despite the very dissimilar values in comparison to those of classical moments, the results indicate the same behaviour, i.e., that the distribution is positively skewed and leptokurtic. In the following chapters we will utilize both classical and L moments in several hydrological problems.

2.5.6 Time scale and distribution shape

In the above example we saw that the distribution of a natural quantity such as rainfall, which is very random and simultaneously takes only nonnegative values, at a fine timescale, such as daily, exhibits high variation, strongly positive skewness and inverted-J shape of probability density function, which means that the most probable value (mode) is zero. Clearly, rainfall cannot be negative, so its distribution cannot be symmetric. It happens that the main body of rainfall values are close to zero, but a few values are extremely high (with low probability), which creates the distribution tail to the right. As we will see in other chapters, the distribution tails are even longer (or fatter, stronger, heavier) than described by this simple exponential distribution. In the exponential distribution, as demonstrated above, all moments (for any arbitrarily high but finite value of r) exist, i.e. take finite values. This is not, however, the case in long-tail distributions, whose moments above a certain rank r^* diverge, i.e. are infinite.

As we proceed from fine to coarser scales, e.g. from the daily toward the annual scale, aggregating more and more daily values, all moments increase but the standard deviation increases at a smaller rate in comparison to the mean, so the coefficient of variation decreases. In a similar manner, the coefficients of skewness and kurtosis decrease. Thus, the distributions tend to become more symmetric and the density functions take a more bell-shaped pattern. As we will se below, there are theoretical reasons for this behaviour for coarse timescales, which are related to the *central limit theorem* (see section 2.10.1). A more general theoretical explanation of the observed natural behaviours both in fine and coarse timescales is offered by the principle of *maximum entropy* (Koutsoyiannis, 2005a, b).

2.6 Change of variable

In hydrology we often prefer to use in our analyses, instead of the variable X that naturally describes a physical phenomenon (such as the rainfall depth in the example above), another variable Y which is a one-to-one mathematical transformation of X, e.g. $Y = g(X)$. If X is modelled as a random variable, then Y should be a random variable, too. The event $\{Y \le y\}$ is identical with the event $\{X \le g^{-1}(y)\}$ where g^{-1} is the inverse function of g. Consequently, the distribution functions of X and Y are related by

$$F_Y(y) = P\{Y \le y\} = P\{X \le g^{-1}(y)\} = F_X\left(g^{-1}(y)\right) \tag{2.38}$$

In the case that the variables are continuous and the function g differentiable, it can be shown that the density function of Y is given from that of X by

$$f_Y(y) = \frac{f_X\left(g^{-1}(y)\right)}{\left|g'\left(g^{-1}(y)\right)\right|} \tag{2.39}$$

where g' is the derivative of g. The application of (2.39) is elucidated in the following examples.

2.6.1 Example 1: the standardized variable

Very often the following transformation of a natural variable X is used:

$$Z = (X - m_X) / \sigma_X$$

This is called the *standardized variable*, is dimensionless and, as we will prove below, it has (a) zero mean, (b) unit standard deviation, and (c) third and fourth central moments equal to the coefficients of skewness and kurtosis of X, respectively.

From (2.38), setting $X = g^{-1}(Z) = \sigma_X Z + m_X$, we directly obtain

$$F_Z(z) = F_X\left(g^{-1}(z)\right) = F_X(\sigma_X z + m_X)$$

Given that $g'(x) = 1 / \sigma_X$, from (2.39) we obtain

$$f_Z(z) = \frac{f_X\left(g^{-1}(z)\right)}{\left|g'\left(g^{-1}(z)\right)\right|} = \sigma_X f_X(\sigma_X z + m_X)$$

Besides, from (2.20) we get

$$E[Z] = E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx = \int_{-\infty}^{\infty} \frac{x - m_X}{\sigma_X} f_X(x) dx =$$

$$= \frac{1}{\sigma_X} \int_{-\infty}^{\infty} x f_X(x) dx - \frac{m_X}{\sigma_X} \int_{-\infty}^{\infty} f_X(x) dx = \frac{1}{\sigma_X} m_X - \frac{m_X}{\sigma_X} 1$$

and finally

$$m_Z = E[Z] = 0$$

This entails that the moments about the origin and the central moments of Z are identical. Thus, the rth moment is

$$E\left[Z^r\right] = E\left[(g(X))^r\right] = \int_{-\infty}^{\infty} \left(\frac{x - m_X}{\sigma_X}\right)^r f_X(x) dx =$$

$$= \frac{1}{\sigma_X^r} \int_{-\infty}^{\infty} (x - m_X)^r f_X(x) dx = \frac{1}{\sigma_X^r} \mu_X^{(r)}$$

and finally

$$\mu_Z^{(r)} = m_Z^{(r)} = \frac{\mu_X^{(r)}}{\sigma_X^r}$$

2.6.2 Example 2: The exponential transformation and the Pareto distribution

Assuming that the variable X has exponential distribution as in the example of section 2.5.5, we will study the distribution of the transformed variable $Y = e^X$. The density and distribution of X are

$$f_X(x) = (1/\lambda)e^{-x/\lambda}, F_X(x) = 1 - e^{-x/\lambda}$$

and our transformation has the properties

$$Y = g(X) = e^X, g^{-1}(Y) = \ln Y, g'(X) = e^X$$

where $X \geq 0$ and $Y \geq 1$. From (2.38) we obtain

$$F_Y(y) = F_Y\left(g^{-1}(y)\right) = F_X(\ln y) = 1 - e^{-\ln y/\lambda} = 1 - y^{-1/\lambda}$$

and from (2.39)

$$f_Y(y) = \frac{f_X\left(g^{-1}(y)\right)}{\left|g'\left(g^{-1}(y)\right)\right|} = \frac{(1/\lambda)e^{-\ln y/\lambda}}{e^{\ln y}} = \frac{\lambda y^{-\lambda}}{y} = (1/\lambda)y^{-(1/\lambda+1)}$$

The latter can be more easily derived by taking the derivative of $F_Y(y)$.

This specific distribution is known as the *Pareto distribution*. The rth moment of this distribution is

$$m_Y^{(r)} = E\left[Y^r\right] = \int_{-\infty}^{\infty} y^r f_Y(y)dy = \int_1^{\infty} \lambda y^{r-1/\lambda-1}dx = \left.\frac{y^{r-1/\lambda}}{\lambda(r-1/\lambda)}\right]_{y=1}^{\infty} = \begin{cases} \dfrac{1}{1-r\lambda}, & r < 1/\lambda \\ \infty, & r \geq 1/\lambda \end{cases}$$

This clearly shows that only a finite number of moments ($r < 1/\lambda$) exist for this distribution, which means that the Pareto distribution has a long-tail.

2.7 Joint, marginal and conditional distributions

In the above sections, concepts of probability pertaining to the analysis of a single variable X have been described. Often, however, the simultaneous modelling of two (or more) variables is necessary. Let the couple of random variables (X, Y) represent two sample spaces (Ω_X, Ω_Y), respectively. The intersection of the two events $\{X \leq x\}$ and $\{Y \leq y\}$, denoted as $\{X \leq x\} \cap \{Y \leq y\} \equiv \{X \leq x, Y \leq y\}$ is an event of the sample space $\Omega_{XY} = \Omega_X \times \Omega_Y$. Based on the latter event, we can define the *joint probability distribution function* of (X, Y) as a function of the real variables (x, y):

$$F_{XY}(x, y) := P\{X \leq x, Y \leq y\} \tag{2.40}$$

The subscripts X, Y can be omitted if there is no risk of ambiguity. If F_{XY} is differentiable, then the function

$$f_{XY}(x, y) := \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y} \tag{2.41}$$

is the *joint probability density function* of the two variables. Obviously, the following equation holds:

$$F_{XY}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(\xi,\omega)d\omega d\xi \tag{2.42}$$

The functions

$$\begin{aligned} F_X(x) &:= P(X \le x) = \lim_{y\to\infty} F_{XY}(x,y) \\ F_Y(y) &:= P(Y \le y) = \lim_{x\to\infty} F_{XY}(x,y) \end{aligned} \tag{2.43}$$

are called the *marginal probability distribution functions* of *X* and *Y*, respectively. Also, the *marginal probability density functions* can be defined, from

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx \tag{2.44}$$

Of particular interest are the so-called *conditional probability distribution function* and *conditional probability density function* of *X* for a specified value of *Y* = *y*; these are given by

$$F_{X|Y}(x \mid y) = \frac{\int_{-\infty}^{x} f_{XY}(\xi,y)d\xi}{f_Y(y)}, \quad f_{X|Y}(x \mid y) = \frac{f_{XY}(x,y)}{f_Y(y)} \tag{2.45}$$

respectively. Switching *X* and *Y* we obtain the conditional functions of *Y*.

2.7.1 Expected values - moments

The expected value of any given function *g*(*X*, *Y*) of the random variables (*X*, *Y*) is defined by

$$E[g(X,Y)] := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)dydx \tag{2.46}$$

The quantity $E[X^p Y^q]$ is called *p* + *q* moment of *X* and *Y*. Likewise, the quantity $E[(X - m_X)^p (Y - m_Y)^q]$ is called the *p* + *q* central moment of *X* and *Y*. The most common of the latter case is the 1+1 moment, i.e.,

$$\sigma_{XY} := E[(X - m_X)(Y - m_Y)] = E[XY] - m_X m_Y \tag{2.47}$$

known as *covariance* of *X* and *Y* and also denoted as $\mathrm{Cov}[X,Y]$. Dividing this by the standard deviations σ_X and σ_Y we define the *correlation coefficient*

$$\rho_{XY} := \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\ \mathrm{Var}[Y]}} \equiv \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{2.48}$$

which is dimensionless with values $-1 \le \rho_{XY} \le 1$. As we will see later, this is an important parameter for the study of the correlation of two variables.

The *conditional expected value* of a function *g*(*X*) for a specified value *y* of *Y* is defined by

$$E[g(X) \mid y] \equiv E[g(X) \mid Y = y] := \int_{-\infty}^{\infty} g(x) f_{X|Y}(x \mid y) dx \tag{2.49}$$

An important quantity of this type is the conditional expected value of X:

$$E[X \mid y] \equiv E[X \mid Y = y] := \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) dx \tag{2.50}$$

Likewise, the conditional expected value of Y is defined. The conditional variance of X for a given $Y = y$ is defined as

$$\mathrm{Var}[X \mid Y = y] := E\left[(X - E[X \mid Y = y])^2 \mid Y = y\right] = \int_{-\infty}^{\infty} (x - E[X \mid Y = y])^2 f_{X|Y}(x \mid y) dx \tag{2.51}$$

or

$$\mathrm{Var}[X \mid y] \equiv \mathrm{Var}[X \mid Y = y] := E\left[X^2 \mid Y = y\right] - (E[X \mid Y = y])^2 \tag{2.52}$$

Both $E[X \mid Y = y] \equiv E[X \mid y] =: \eta(y)$ and $\mathrm{Var}[X \mid Y = y] \equiv \mathrm{Var}[X \mid y] =: \upsilon(y)$ are functions of the real variable y, rather than constants. If we do not specify in the condition the value y of the random variable Y, then the quantities $E[X \mid Y] = \eta(Y)$ and $\mathrm{Var}[X \mid Y] = \upsilon(Y)$ become functions of the random variable Y. Hence, they are random variables themselves and they have their own expected values, i.e.,

$$E[E[X \mid Y]] = \int_{-\infty}^{\infty} E[X \mid y] f_Y(y) dy, \quad E[\mathrm{Var}[X \mid Y]] = \int_{-\infty}^{\infty} \mathrm{Var}[X \mid y] f_Y(y) dy \tag{2.53}$$

It is easily shown that $E[E[X \mid Y]] = E[X]$.

2.7.2 Independent variables

The random variables (X, Y) are called independent if for any couple of values (x, y) the following equation holds:

$$F_{XY}(x, y) = F_X(x) F_Y(y) \tag{2.54}$$

The following equation also holds:

$$f_{XY}(x, y) = f_X(x) f_Y(y) \tag{2.55}$$

and is equivalent with (2.54). The additional equations

$$\sigma_{XY} = 0 \leftrightarrow \rho_{XY} = 0 \leftrightarrow E[XY] = E[X]E[Y] \tag{2.56}$$

$$E[X \mid Y = x] = E[X], \quad E[Y \mid X = x] = E[Y] \tag{2.57}$$

are simple consequences of (2.54) but not sufficient conditions for the variable (X, Y) to be independent. Two variables (X, Y) for which (2.56) holds are called *uncorrelated.*

2.7.3 Sums of variables

A consequence of the definition of the expected value (equation (2.46)) is the relationship

$$E[c_1 g_1(X,Y) + c_2 g_2(X,Y)] = c_1 E[g_1(X,Y)] + c_2 E[g_2(X,Y)] \tag{2.58}$$

where c_1 and c_2 are any constant values whereas g_1 and g_2 are any functions. Apparently, this property can be extended to any number of functions g_i. Applying (2.58) for the sum of two variables we obtain

$$E[X+Y] = E[X] + E[Y] \tag{2.59}$$

Likewise,

$$E\left[(X - m_X + Y - m_Y)^2\right] = E\left[(X - m_X)^2\right] + E\left[(Y - m_Y)^2\right] + 2E[(X - m_X)(Y - m_Y)] \tag{2.60}$$

which results in

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y] \tag{2.61}$$

The probability distribution function of the sum $Z = X + Y$ is generally difficult to calculate. However, if X and Y are independent then it can be shown that

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-w) f_Y(w) dw \tag{2.62}$$

The latter integral is known as the *convolution integral* of $f_X(x)$ and $f_Y(y)$.

2.7.4 An example of correlation of two variables

We study a lake with an area of 10 km^2 lying on an impermeable subsurface. The inflow to the lake during the month of April, composed of rainfall and catchment runoff, is modelled as a random variable with mean 4.0×10^6 m^3 and standard deviation 1.5×10^6 m^3. The evaporation from the surface of the lake, which is the only outflow, is also modelled as a random variable with mean 90.0 mm and standard deviation 20.0 mm. Assuming that inflow and outflow are stochastically independent, we seek to find the statistical properties of the water level change in April as well as the correlation of this quantity with inflow and outflow.

Initially, we express the inflow in the same units as the outflow. To this aim we divide the inflow volume by the lake area, thus calculating the corresponding change in water level. The mean is 4.0×10^6 / 10.0×10^6 = 0.4 m = 400.0 mm and the standard deviation 1.5×10^6 / 10.0×10^6 = 0.15 m = 150.0 mm.

We denote by X and Y the inflow and outflow in April, respectively and by Z the water level change in the same month. Apparently,

$$Z = X - Y \tag{2.63}$$

We are given the quantities

$$m_X = E[X] = 400.0\ \text{mm},\ \sigma_X = \sqrt{Var[X]} = 150.0\ \text{mm}$$

$$m_Y = E[Y] = 90.0\ \text{mm},\ \sigma_Y = \sqrt{Var[Y]} = 20.0\ \text{mm}$$

and we have assumed that the two quantities are independent, so that their covariance Cov[X, Y] = 0 (see 2.56) and their correlation $\rho_{XY} = 0$.

Combining (2.63) and (2.58) we obtain

$$E[Z] = E[X - Y] = E[X] - E[Y] \to m_Z = m_X - m_Y \tag{2.64}$$

or $m_Z = 310.0$ mm. Subtracting (2.63) and (2.64) side by side we obtain

$$Z - m_z = (X - m_X) - (Y - m_Y) \tag{2.65}$$

and squaring both sides we find

$$(Z - m_z)^2 = (X - m_X)^2 + (Y - m_Y)^2 - 2(X - m_X)(Y - m_Y)$$

which, by taking expected values in both sides, results in the following equation (similar to (2.61) except in the sign of the last term)

$$\mathrm{Var}[Z] = \mathrm{Var}[X - Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] - 2\mathrm{Cov}[X, Y] \tag{2.66}$$

Since Cov[X, Y] = 0, (2.66) gives

$$\sigma_Z^2 = 150.0^2 + 20.0^2 = 22900.0\ \mathrm{mm}^2$$

and $\sigma_Z = 151.3$ mm.

Multiplying both sides of (2.65) by ($X - m_X$) and then taking expected values we find

$$E[(Z - m_z)(X - m_X)] = E[(X - m_X)^2] - E[(X - m_X)(Y - m_Y)]$$

or

$$\mathrm{Cov}[Z, X] = \mathrm{Var}[X] - \mathrm{Cov}[X, Y] \tag{2.67}$$

in which the last term is zero. Thus,

$$\sigma_{ZY} = \sigma_X^2 = 150.0^2 = 22500.0\ \mathrm{mm}^2$$

Consequently, the correlation coefficient of X and Z is

$$\rho_{ZX} = \sigma_{ZX} / (\sigma_Z \sigma_X) = 22500.0 / (151.3 \times 150.0) = 0.991$$

Likewise,

$$\mathrm{Cov}[Z, Y] = \mathrm{Cov}[X, Y] - \mathrm{Var}[Y] \tag{2.68}$$

The first term of the right hand side is zero and thus

$$\sigma_{ZY} = -\sigma_Y^2 = -20.0^2 = -400.0\ \mathrm{mm}^2$$

Consequently, the correlation coefficient of Y and Z is

$$\rho_{ZY} = \sigma_{ZY} / (\sigma_Z \sigma_Y) = -400.0 / (151.3 \times 20.0) = -0.132$$

The positive value of ρ_{ZX} manifests the fact that the water level increases with the increase of inflow (positive correlation of X and Z). Conversely, the negative correlation of Y and Z ($\rho_{ZY} < 0$) corresponds to the fact that the water level decreases with the increase of outflow.

The large, close to one, value of ρ_{ZX} in comparison to the much lower (in absolute value) value of ρ_{ZY} reflects the fact that in April the change of water level depends primarily on the inflow and secondarily on the outflow, given that the former is greater than the latter and also has greater variability (standard deviation).

2.7.5 An example of dependent discrete variables

Further to the example of section 2.4.2, we introduce the random variables X and Y to quantify the events (wet or dry day) of today and yesterday, respectively. Values of X or Y equal to 0 and 1 correspond to a day being dry and wet, respectively. We use the values of conditional probabilities (also called transition probabilities) of section 2.4.2, which with the current notation are:

$$\pi_{1|1} := P\{X = 1|Y = 1\} = 0.40, \pi_{0|1} := P\{X = 0|Y = 1\} = 0.60$$

$$\pi_{1|0} := P\{X = 1|Y = 0\} = 0.15, \pi_{0|1} := P\{X = 0|Y = 1\} = 0.85$$

The unconditional or marginal probabilities, as found in section 2.4.2, are

$$p_1 := P\{X = 1\} = 0.20, p_0 := P\{X = 0\} = 0.80$$

and the joint probabilities, again as found in section 2.4.2, are

$$p_{11} := P\{X = 1, Y = 1\} = 0.08, p_{01} := P\{X = 0, Y = 1\} = 0.12$$

$$p_{10} := P\{X = 1, Y = 0\} = 0.12, p_{00} := P\{X = 0, Y = 0\} = 0.68$$

It is reminded that the marginal probabilities of Y were assumed equal to those of X, which resulted in time symmetry ($p_{01} = p_{10}$). It can be easily shown (homework) that the conditional quantities $\pi_i|_j$ can be determined from the joint p_{ij} and vice versa, and the marginal quantities p_i can be determined for either of the two series. Thus, from the set of the ten above quantities only two are independent (e.g. $\pi_{1|1}$ and $\pi_{1|0}$) and all others can be calculated from these two.

The marginal moments of X and Y are

$$E[X] = E[Y] = 0\,p_0 + 1\,p_1 = p_1 = 0.20,\ \ E[X^2] = E[Y^2] = 0^2\,p_0 + 1^2\,p_1 = p_1 = 0.20$$

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 = 0.2 - 0.2^2 = 0.16 = \mathrm{Var}[Y]$$

and the 1+1 joint moment is

$$E[XY] = 0 \times 0\,p_{00} + 0 \times 1\,p_{01} + 1 \times 0\,p_{10} + 1 \times 1\,p_{11} = p_{11} = 0.08$$

so that the covariance is

$$\sigma_{XY} \equiv \mathrm{Cov}[X, Y] = E[XY] - E[X]\,E[Y] = 0.08 - 0.2^2 = 0.04$$

and the correlation coefficient

$$\rho_{XY} := \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\ \mathrm{Var}[Y]}} \equiv \frac{0.04}{0.16} = 0.25$$

If we know that yesterday was a dry day, the moments for today are calculated from (2.49)-(2.52), replacing the integrals with sums and the conditional density $f_{X|Y}$ with the conditional probabilities $\pi_{i|j}$:

$$E[X|Y = 0] = 0\ \pi_{0|0} + 1\ \pi_{1|0} = \pi_{1|0} = 0.15,\ \ E[X^2|Y = 0] = 0^2\ \pi_{0|0} + 1^2\ \pi_{1|0} = \pi_{1|0} = 0.15$$

$$\mathrm{Var}[X|Y = 0] = 0.15 - 0.15^2 = 0.128$$

Likewise,

$$E[X|Y = 1] = 0\ \pi_{0|1} + 1\ \pi_{1|1} = \pi_{1|1} = 0.40,\ \ E[X^2|Y = 1] = 0^2\ \pi_{0|1} + 1^2\ \pi_{1|1} = \pi_{1|1} = 0.40$$

$$\mathrm{Var}[X|Y = 1] = 0.40 - 0.40^2 = 0.24$$

We observe that in the first case, $\mathrm{Var}[X|Y = 0] < \mathrm{Var}[X]$. This can be interpreted as a decrease of uncertainty for the event of today, caused by the information that we have for yesterday. However, in the second case $\mathrm{Var}[X|Y = 1] > \mathrm{Var}[X]$. Thus, the information that yesterday was wet, increases uncertainty for today. However, on the average the information about yesterday results in reduction of uncertainty. This can be expressed mathematically by $E[\mathrm{Var}[X|Y]]$ defined in (2.53), which is a weighted average of the two $\mathrm{Var}[X|Y = j]$:

$$E\{\mathrm{Var}[X|Y]\} := \mathrm{Var}[X|Y = 0]\, p_0 + \mathrm{Var}[X|Y = 1]\, p_1$$

This yields

$$E\{\mathrm{Var}[X|Y\} := 0.128 \times 0.8 + 0.24 \times 0.2 = 0.15 < 0.16 = \mathrm{Var}[X]$$

2.8 Many variables

All above theoretical analyses can be easily extended to more than two random variables. For instance, the distribution function of the n random variables $X_1, X_2, \ldots, X_n$ is

$$F_{X_1,\cdots,X_n}(x_1,\ldots,x_n) := P\{X_1 \le x_1,\ldots,X_n \le x_n\} \tag{2.69}$$

and is related to the n-dimensional probability density function by

$$F_{X_1,\cdots,X_n}(x_1,\ldots,x_n) = \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n} f_{X_1,\cdots,X_n}(\xi_1,\ldots,\xi_n) d\xi_n \cdots d\xi_1 \tag{2.70}$$

The variables $X_1, X_2, \ldots, X_n$ are independent if for any $x_1, x_2, \ldots, x_n$ the following holds true:

$$F_{X_1,\cdots,X_n}(x_1,\cdots,x_n) = F_{X_1}(x_1)\ldots F_{X_n}(x_n) \tag{2.71}$$

The expected values and moments are defined in a similar manner as in the case of two variables, and the property (2.58) is generalized for functions g_i of many variables.

2.9 The concept of a stochastic process

An arbitrarily (usually infinitely) large family of random variables $X(t)$ is called a *stochastic process* (Papoulis, 1991). To each one of them there corresponds an index t, which takes values from an *index set T*. Most often, the index set refers to time. The time t can be either *discrete* (when T is the set of integers) or continuous (when T is the set of real numbers); thus we have respectively a *discrete-time* or a *continuous-time* stochastic process. Each of the random variables $X(t)$ can be either discrete (e.g. the wet or dry state of a day) or continuous (e.g. the rainfall depth); thus we have respectively a *discrete-state* or a *continuous-state* stochastic process. Alternatively, a stochastic process may be denoted as X_t instead of $X(t)$; the notation X_t is more frequent for discrete-time processes. The index set can also be a vector space, rather than the real line or the set of integers; this is the case for instance when we assign a random variable (e.g. rainfall depth) to each geographical location (a two dimensional vector space) or to each location and time instance (a three-dimensional vector space). Stochastic processes with multidimensional index set are also known as random fields.

A realization $x(t)$ of a stochastic process $X(t)$, which is a regular (numerical) function of the time t, is known as a *sample function*. Typically, a realization is observed at countable time instances (not in continuous time, even in a continuous-time process). This sequence of observations is also called a *time series*. Clearly then, a time series is a sequence of numbers, whereas a stochastic process is a family of random variables. Unfortunately, a large literature body does not make this distinction and confuses stochastic processes with time series.

2.9.1 Distribution function

The distribution function of the random variable X_t, i.e.,

$$F(x;t) := P\{X(t) \le x\} \tag{2.72}$$

is called *first order distribution function* of the process. Likewise, the *second order distribution function* is

$$F(x_1, x_2; t_1, t_2) := P\{X(t_1) \le x_1, X(t_2) \le x_2\} \tag{2.73}$$

and the n*th order distribution function*

$$F(x_1, \dots, x_n; t_1, \dots, t_n) := P\{X(t_1) \le x_1, \dots, X(t_n) \le x_n\} \tag{2.74}$$

A stochastic process is completely determined if we know the nth order distribution function for any n. The nth order probability density function of the process is derived by taking the derivatives of the distribution function with respect to all x_i.

2.9.2 Moments

The moments are defined in the same manner as in sections 2.5 and 2.7.1. Of particular interest are the following:

1. The *process mean*, i.e. the expected value of the variable $X(t)$:

$$m(t) := E[X(t)] = \int_{-\infty}^{\infty} x f(x;t) dt \tag{2.75}$$

2. The process *autocovariance*, i.e. the covariance of the random variables $X(t_1)$ and $X(t_2)$:

$$C(t_1, t_2) := \mathrm{Cov}[X(t_1), X(t_2)] = E[(X(t_1) - m(t_1))(X(t_2) - m(t_2))] \tag{2.76}$$

The process variance (the variance of the variable $X(t)$), is $\mathrm{Var}[X(t)] = C(t, t)$. Consequently, the process autocorrelation (the correlation coefficient of the random variables $X(t_1)$ and $X(t_2)$) is

$$\rho(t_1, t_2) := \frac{\mathrm{Cov}[X(t_1), X(t_2)]}{\sqrt{Var[X(t_1)]Var[X(t_2)]}} = \frac{C(t_1, t_2)}{\sqrt{C(t_1, t_1)C(t_2, t_2)}} \tag{2.77}$$

2.9.3 Stationarity

As implied by the above notation, in the general setting, the statistics of a stochastic process, such as the mean and autocovariance, depend on time and thus vary with time. However, the case where these statistical properties remain constant in time is most interesting. A process with this property is called *stationary* process. More precisely, a process is called strict-sense stationary if all its statistical properties are invariant to a shift of time origin. That is, the distribution function of any order of $X(t + \tau)$ is identical to that of $X(t)$. A process is called *wide-sense stationary* if its mean is constant and its autocovariance depends only on time differences, i.e.

$$E[X(t)] = \mu, \quad E[(X(\tau) - \mu)\ (X(t + \tau) - \mu)] = C(\tau) \tag{2.78}$$

A strict-sense stationary process is also wide-sense stationary but the inverse is not true.

A process that is not stationary is called nonstationary. In a nonstationary process one or more statistical properties depend on time. A typical case of a nonstationary process is a cumulative process whose mean is proportional to time. For instance, let us assume that the rainfall intensity $\Xi(t)$ at a geographical location and time of the year is a stationary process, with a mean μ. Let us further denote $X(t)$ the rainfall depth collected in a large container (a cumulative raingauge) at time t and assume that at the time origin, $t = 0$, the container is empty. It is easy then to understand that $E[X(t)] = \mu\ t$. Thus $X(t)$ is a nonstationary process.

We should stress that stationarity and nonstationarity are properties of a process, not of a sample function or time series. There is some confusion in the literature about this, as a lot of studies assume that a time series is stationary or not, or can reveal whether the process is stationary or not. As a general rule, to characterise a process nonstationary, it suffices to show that some statistical property is a *deterministic* function of time (as in the above example of the raingauge), but this cannot be straightforwardly inferred merely from a time series.

Stochastic processes describing periodic phenomena, such as those affected by the annual cycle of Earth, are clearly nonstationary. For instance, the daily temperature at a mid-latitude location could not be regarded as a stationary process. It is a special kind of a nonstationary process, as its properties depend on time on a periodical manner (are periodic functions of time). Such processes are called *cyclostationary* processes.

2.9.4 Ergodicity

The concept of *ergodicity* (from the Greek words *ergon* – work – and *odos* – path) is central for the problem of the determination of the distribution function of a process from a single sample function (time series) of the process. A stationary stochastic process is ergodic if any statistical property can be determined from a sample function. Given that in practice, the statistical properties are determined as time averages of time series, the above definition can be stated alternatively as: a stationary stochastic process is ergodic if time averages equal ensemble averages (i.e. expected values). For example, a stationary stochastic process is *mean ergodic* if

$$E[X(t)] = \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N} X(t) \qquad \text{(for a discrete time process)}$$

$$E[X(t)] = \lim_{T\to\infty} \frac{1}{T} \int_0^T X(t)dt \qquad \text{(for a continuous time process)} \tag{2.79}$$

The left-hand side in the above equations represents the ensemble average whereas the right-hand side represents the time average, for the limiting case of infinite time. Whilst the left-hand side is a parameter, rather than a random variable, the right-hand side is a random variable (as a sum or integral of random variables). The equating of a parameter with a random variable implies that the random variable has zero variance. This is precisely the condition that makes a process ergodic, a condition that does not hold true for every stochastic process.

2.10 The central limit theorem and some common distribution functions

The *central limit theorem* is one of the most important in probability theory. It concerns the limit distribution function of a sum of random variables – components, which, under some conditions but irrespectively of the distribution functions of the components, is always the same, the celebrated *normal distribution*. This is the most commonly used distribution in probability theory as well as in all scientific disciplines and can be derived not only as a consequence of the central limit theorem but also from the principle of maximum entropy, a very powerful physical and mathematical principle (Papoulis, 1990, p. 422-430).

In this section we will present the central limit theorem, the normal distribution, and some other distributions closely connected to the normal (χ^2 and Student). All these distributions are fundamental in statistics (chapter 3) and are commonly used for statistical estimation and prediction. Besides, the normal distribution has several applications in hydrological statistics, which will be discussed in chapters 5 and 6.

2.10.1 The central limit theorem and its importance

Let X_i ($i = 1, \ldots, n$) be random variables and let $Z := X_1 + X_2 + \cdots + X_n$ be its sum with $E[Z] = m_Z$ and $\mathrm{Var}[Z] = s_Z^2$. The central limit theorem says that the distribution of *Z,* under some general conditions (see below) has a specific limit as *n* tends to infinity, i.e.,

$$F_Z(z) \xrightarrow[n \to \infty]{} \int_{-\infty}^{z} \frac{1}{\sigma_Z \sqrt{2\pi}} e^{-\frac{(\zeta - m_Z)^2}{2\sigma_Z^2}} d\zeta \tag{2.80}$$

and in addition, if X_i are continuous variables, the density function of Z has also a limit,

$$f_Z(z) \xrightarrow[n \to \infty]{} \frac{1}{\sigma_Z \sqrt{2\pi}} e^{-\frac{(z - m_Z)^2}{2\sigma_Z^2}} \tag{2.81}$$

The distribution function in the right-hand side of (2.80) is called the *normal* (or *Gauss*) *distribution* and, likewise, the function in the right-hand side of (2.81) is called the *normal probability density function*.

In practice, the convergence for $n \to \infty$ can be regarded as an approximation if n is sufficiently large. How large should n be so that the approximation be satisfactory, depends on the (joint) distribution of the components X_i. In most practical application, a value $n = 30$ is regarded to be satisfactory (with the condition that X_i are independent and identically distributed). Fig. 2.4 gives a graphical demonstration of the central limit theorem based on an example. Starting from independent random variables X_i with exponential distribution, which is positively skewed, we have calculated (using (2.62)) and depicted the distribution of the sum of 2, 4, 8, 16 and 32 variables. If the distribution of X_i were symmetric, the convergence would be much faster.

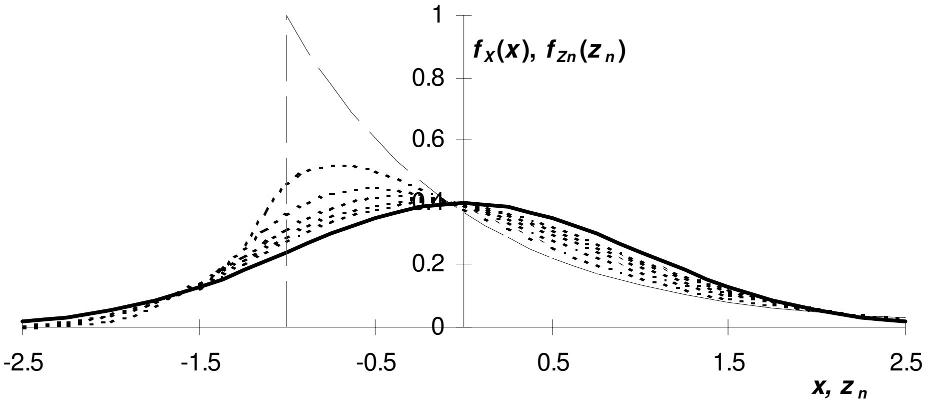

Fig. 2.4 Convergence of the sum of exponentially distributed random variables to the normal distribution (thick line). The dashed line with peak at $x = -1$ represents the probability density of the initial variables X_i, which is $f_X(x) = e^{-(x-1)}$ (mean 0, standard deviation 1). The dotted lines (going from the more peaked to the less peaked) represent the densities of the sums $Z_n = (X_1 + \ldots + X_n) / n$ for $n = 2, 4, 8, 16$ and 32. (The division of the sum by n helps for a better presentation of the curves, as all Z_i have the same mean and variance, 0 and 1, respectively, and does not affect the essentials of the central limit theorem.)

The conditions for the validity of the central limit theorem are general enough, so that they are met in many practical situations. Some sets of conditions (e.g. Papoulis, 1990, p. 215) with particular interest are the following: (a) the variables X_i are independent identically distributed with finite third moment; (b) the variables X_i are bounded from above and below with variance greater than zero; (c) the variables X_i are independent with finite third moment and the variance of Z tends to infinity as n tends to infinity. The theorem has been extended for variables X_i that are interdependent, but each one is effectively dependent on a finite number of other variables. Practically speaking, the central limit theorem gives satisfactory approximations for sums of variables unless the tail of the density functions of X_i are over-exponential (long, like in the Pareto example; see section 2.6.2) or the dependence of the variables is very strong and spans the entire sequence of X_i (long range dependence; see chapter 4). Note that the normal density function has an exponential tail (it can be approximated by an exponential decay for large x) and all its moments exist (are finite), whereas in over-exponential densities all moments beyond a certain order diverge. Since in hydrological processes the over-exponential tails, as well as the long-range dependence, are not uncommon, we must be attentive in the application of the theorem.

We observe in (2.80) and (2.81) that the limits of the functions $F_Z(z)$ and $f_Z(z)$ do not depend on the distribution functions of X_i, that is, the result is the same irrespectively of the distribution functions of X_i. Thus, provided that the conditions for the applicability of the theorem hold, (a) we can know the distribution function of the sum without knowing the distribution of the components, and (b) precisely the same distribution describes any variable that is a sum of a large number of components. Here lies the great importance of the normal distribution in all sciences (mathematical, physical, social, economical, etc.). Particularly, in statistics, as we will see in chapter 3, the central limit theorem implies that the sample average for any type of variables will have normal distribution (for a sample large enough).

In hydrological statistics, as we will see in chapters 5 and 6 in more detail, the normal distribution describes with satisfactory accuracy variables that refer to long time scales such as annual. Thus, the annual rainfall depth in a wet area is the sum of many (e.g. more than 30) independent rainfall events during the year (this, however, is not valid for rainfall in dry areas). Likewise, the annual runoff volume passing through a river section can be regarded as the sum of 365 daily volumes. These are not independent, but as an approximation, the central limit theorem can be applicable again.

2.10.2 The normal distribution

The random variable X is *normally distributed* or (Gauss distributed) with parameters μ and σ (symbolically $N(\mu, \sigma)$ if its probability density is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.82}$$

The corresponding distribution function is

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(\xi-\mu)^2}{2\sigma^2}} d\xi \tag{2.83}$$

The mean and standard deviation of X are μ and σ, respectively. The distribution is symmetric (Fig. 2.5) and thus its third central moment and its third L moment are zero. The fourth central moment is $3\sigma^4$ (hence $C_k = 3$) and the fourth L moment is $0.1226\lambda_X^{(2)}$ (hence $\tau_X^{(4)} = 0.1226$).

The integral in the right-hand side of (2.83) is not calculated analytically. Thus, the typical calculations ($x \to F_X(x)$ or $F_X(x) \to x$) are done either numerically or using tabulated values of the so-called *standard normal variate* Z, that is obtained from X with the transformation

$$Z = \frac{X-\mu}{\sigma} \leftrightarrow X = \mu + \sigma Z \tag{2.84}$$

and its distribution is $N(0,1)$. It is easy to obtain (see section 2.6.1) that

$$F_X(x) = F_Z(z) = F_Z\left(\frac{x-\mu}{\sigma}\right) \tag{2.85}$$

Such tables are included in all textbooks of probability and statistics, as well as in the Appendix of this text. However, nowadays all common numerical computer packages (including spreadsheet applications etc.) include functions for the direct calculation of the integral.*

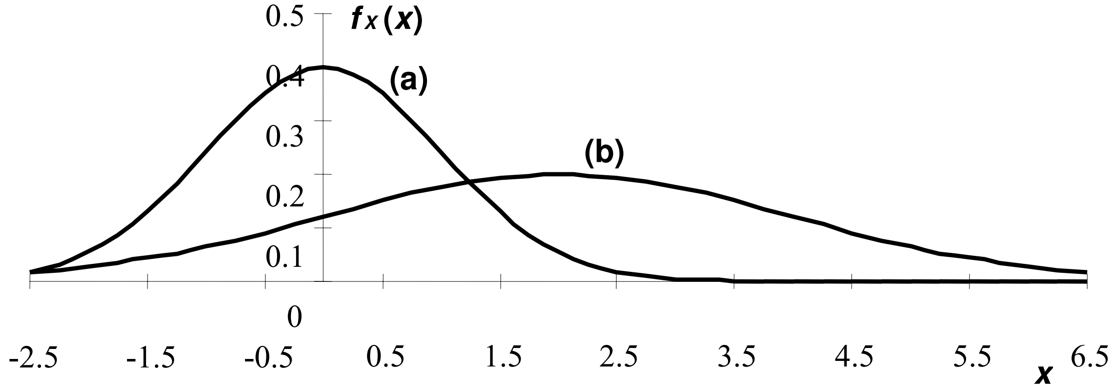

Fig. 2.5 Two examples of normal probability density function (a) $N(0,1)$ and (b) $N(2, 2)$.

2.10.3 A numerical example of the application of the normal distribution

We assume that in an area with wet climate the annual rainfall depth is normally distributed with $\mu = 1750$ mm and $\sigma = 410$ mm. To find the exceedence probability of the value 2500 mm we proceed with the following steps, using the traditional procedure with tabulated z values: $z = (2500 - 1750) / 410 = 1.83$. From normal distribution tables, $F_Z(z) = 0.9664$ ($= F_X(x)$). Hence, $F_X^*(x) = 1 - 0.9664 = 0.0336$.

* For instance, in Excel, the $x \to F_X(x)$ and $F_X(x) \to x$ calculations are done through the functions NormDist and NormInv, respectively (the functions NormSDist and NormSInv can be used for the calculations $z \to F_Z(z)$ and $F_Z(z) \to z$, respectively).

To find the rainfall value that corresponds to exceedence probability 2%, we we proceed with the following steps: $F_X(x) = F_Z(z) = 1 - 0.02 = 0.98$; from the table, $z = 2.05$ hence $x = 1750 + 410 \times 2.05 = 2590.5$ mm. The calculations are straightforward.

2.10.4 The χ^2 distribution

The chi-squared density with n degrees of freedom (symbolically $\chi^2(n)$) is

$$f_X(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \; x \geq 0, \; n = 1,2,\ldots \tag{2.86}$$

where $\Gamma(\,)$ is the gamma function (not to be confused with the gamma distribution function whose special case is the χ^2 distribution), defined from

$$\Gamma(a) = \int_0^\infty y^{a-1} e^{-y} dy \tag{2.87}$$

The gamma function has some interesting properties such as

$$\begin{array}{lll} \Gamma(1) = 1 & \Gamma(1/2) = \sqrt{\pi} & \Gamma(a+1) = a\Gamma(a) \\ \Gamma(n+1) = n! & \Gamma\left(n+\frac{1}{2}\right) = \left(n-\frac{1}{2}\right)\left(n-\frac{3}{2}\right)\cdots\frac{3}{2}\sqrt{\pi} & n = 1,2,\ldots \end{array} \tag{2.88}$$

The χ^2 distribution is a positively skewed distribution (Fig. 2.7) with a single parameter (n). Its mean and variance are n and $2n$, respectively. The coefficients of skewness and kurtosis are $C_s = 2\sqrt{2/n}$ and $C_k = 3 + 12/n$, respectively.

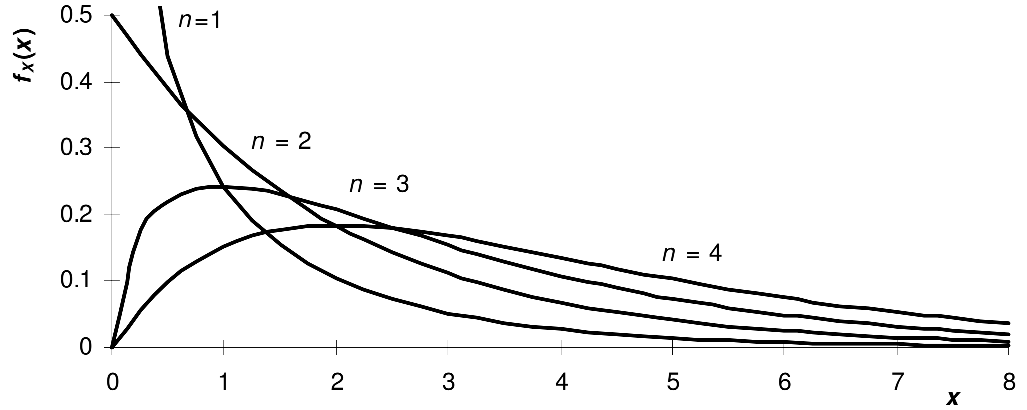

Fig. 2.6 Examples of the $\chi^2(n)$ density for several values of n.

The integral in (2.86) is not calculated analytically, so the typical calculations are based either on tabulated values (see Appendix) or on numerical packages.*

The χ^2 distribution is not directly used to represent hydrological variables; instead the more general gamma distribution (see chapter 6) is used. However, the χ^2 distribution has great importance in statistics (see chapter 3), because of the following theorem: If the random variables X_i ($i = 1, \ldots, n$) are distributed as $N(0, 1)$, then the sum of their squares,

* E.g. in Excel, the relative functions are ChiDist and ChiInv.

$$Q = \sum_{i=1}^{n} X_i^2 \tag{2.89}$$

is distributed as $\chi^2(n)$. Combining this theorem with the central limit theorem we find that for large n the $\chi^2(n)$ distribution tends to the normal distribution.

2.10.5 The Student (*t*) distribution

We shall say that the random variable X has a Student (or t) distribution with n degrees of freedom (symbolically $t(n)$) if its density is

$$f_X(x) = \frac{\Gamma[(n+1)/2]}{\sqrt{\pi n}\,\Gamma(n/2)} \frac{1}{\sqrt{(1+x^2/n)^{n+1}}}, \quad n = 1,2,\ldots \tag{2.90}$$

This is a symmetric distribution (Fig. 2.7) with a single parameter (n), mean zero and variance $n / (n - 2)$. In contrast to the normal distribution, it has an over-exponential tail but for large n (≥ 30) practically coincides with the normal distribution.

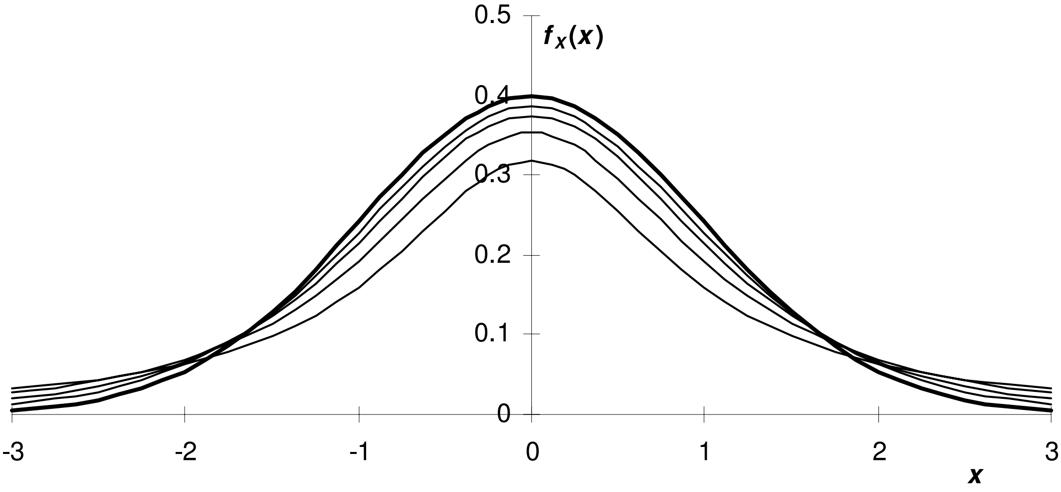

Fig. 2.7 Examples of the $t(n)$ probability density function for n = 1, 2, 4 and 8 (continuous thin lines from down to up), in comparison to the standard normal density $N(0, 1)$ (thick line).

The integral in (2.90) is not calculated analytically, so the typical calculations are based either on tabulated values (see Appendix) or on numerical packages.*

The t distribution is not directly used to represent hydrological variables but it has great importance in statistics (see chapter 3), because of the following theorem: If the random variables Z and W are independent and have $N(0, 1)$ and $\chi^2(n)$ distributions, respectively, then the ratio

$$T = \frac{Z}{\sqrt{W/n}} \tag{2.91}$$

has $t(n)$ distribution.

* E.g. in Excel, the relative functions are TDist and TInv.

References

Greenwood, J. A., J. M. Landwehr, N. C. Matalas, and J. R. Wallis, Probability-weighted moments: Definition and relation to parameters of several distributions expressable in inverse form, *Water Resources Research*, 15, 1049-1054, 1979.

Hosking, J. R. M., L-moments: Analysis and Estimation of Distribution using Linear Combinations of Order Statistics, *Journal of the Royal Statistical Society*, Series B, 52, 105-124, 1990.

Jaynes, E.T., *Probability Theory: The Logic of Science*, Cambridge University Press, 2003.

Kolmogorov, A. N., Foundations of the Theory of Probability, 2nd English Edition, 84 pp. Chelsea Publishing Company, New York, 1956.

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

Koutsoyiannis, D., Uncertainty, entropy, scaling and hydrological stochastics, 2, Time dependence of hydrological processes and time scaling, *Hydrological Sciences Journal*, 50(3), 405-426, 2005b.

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

Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, 3rd ed., McGraw-Hill, New York, 1991.

Chapter 3

Elementary statistical concepts

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

Summary

This chapter aims to serve as a reminder and synopsis of elementary statistical concepts, rather than a systematic and complete presentation of the concepts. Statistics is the applied branch of probability theory which deals with real world problems, trying to draw conclusions based on observations. Two major tasks in statistics are estimation and hypothesis testing. Statistical estimation can be distinguished in parameter estimation and prediction and can be performed either on a point basis (resulting in a single value, the expectation), or on an interval basis (resulting in an interval in which the quantity sought lies, associated with a certain probability or confidence). Uses of statistical estimation in engineering applications include the estimation of parameters of probability distributions, for which several methods exist, and the estimation of quantiles of distributions. Statistical hypothesis testing is also an important tool in engineering studies, not only in typical decision making processes, but also in more analytical tasks, such as in detecting relationships among different geophysical, and particularly hydrological, processes. All this concepts are briefly discussed both in a theoretical level, to clarify the concepts and avoid misuses, and a more practical level to demonstrate the application of the concepts.

3.1 Introductory comments

Statistics is the applied branch of probability theory which deals with real world problems, trying to draw conclusions based on observations. The conclusions are only inferences based on induction, not deductive mathematical proofs; however, if the associated probabilities tend to 1, they become almost certainties. The conclusions are attributed to a *population*, while they are drawn based on a *sample*. Although the content of term population is not strictly defined in the statistical literature, loosely speaking we consider that the term describes any collection of objects whose measurable attributes are of interest. It can be an abstraction of a real world population or of the repetition of a real experiment. The population can be finite (e.g. the population of the annual flows of the hydrologic year 1990-91 for all hydrologic basins of earth with size greater than 100 km^2) or infinite and abstractively defined (e.g. the population of all possible annual flows of a hydrologic basin). The term sample describes a collection of observations from the particular population (see definition in section 3.2.1).

An important concept of statistics is the *estimation*. It is distinguished in *parameter estimation* and *prediction*. In order to clarify these concepts, we consider a population that is represented by a random variable X with distribution function $F_X(x; \theta)$ where θ is a parameter. A parameter estimation problem is a problem in which the parameter is unknown and we seek an estimate of it. A prediction problem is a problem in which the parameter is known and we seek an estimate of the variable X or a function of X. As we will see below, these two problems are dealt with using similar methods of statistics, and thus they are both called estimation problems. The results of the estimation procedures are called *estimates*.

An estimate can be either a *point estimate*, i.e. a numerical value, or an *interval estimate*, i.e. one interval that contains the value sought with a given degree of certainty. Conversely, for a given interval, statistics can calculate the corresponding degree of certainty or, on the contrary, the degree of uncertainty, that the quantity sought lies within the interval.

Another important area of statistics is *hypothesis testing* that constitutes the basis of the decision theory. The process of hypothesis testing requires the formulation of two statements: the basic H_0, that is referred to as the null-hypothesis, and the alternative hypothesis H_1. We start the process of testing by considering that the null hypothesis is true and we use the observations to decide if this hypothesis should be rejected. This is done using of statistical methods. Although the hypothesis testing is based on the same theoretical background as the estimation, the difference lies in the examination of two alternative models, while in the estimation we use only one model.

The background for all these concepts is described in this chapter while in the next chapters several additional numerical examples are given. Of course, statistics include many other areas, such as the Bayesian analysis, but these are not covered in this text.

3.2 Concepts and definitions

3.2.1 Sample

We consider a random variable X with probability density function $f(x)$. The variable is defined based on a sample space Ω and is conceptualized with some population. A sample of X of size (or length) n of is a sequence of *n independent identically distributed* (IID random variables X_1, X_2, …, X_n (each having density $f(x)$) defined on the sample space $\Omega^n = \Omega \times \cdots \times \Omega$ (Papoulis, 1990, p. 238). Each one of the variables X_i corresponds to the possible results of a measurement or an observation of the variable X. After the observations are performed, to each variable there corresponds a numerical value. Consequently, we will have a numerical sequence $x_1, x_2, \dots, x_n$, called the *observed sample*.

The concept of a sample is, therefore, related to two types sequences: an abstractive sequence of random variables and the corresponding sequence of their numerical values. It is common in engineering application to use the term *sample* indistinguishably for both sequences, omitting the term *observed* from the second sequence. However, the two notions are fundamentally different and we should be attentive to distinguish each time in which of the two cases the term sample refers to.

In statistics it is assumed that the construction of a sample of size n or the *sampling* is done by performing n repetitions of an experiment. The repetitions should be independent to each other and be performed under virtually the same conditions. However, in dealing with natural phenomena and in engineering it is not possible to repeat the same experiment, and thus sampling is a process of multiple measurements of the a natural process at different times. As a consequence, it is not possible to ensure that independence and same conditions will hold. Nonetheless, for certain situations we can assume that the previous conditions are approximately valid (an assumption equivalent to simultaneously assuming independence, stationarity and ergodicity, cf. chapters 2 and 4) and thus we can use classical statistical methods of statistics to analyse them. However, there are cases where these conditions (the independence in particular) are far from holding and the use of classical statistics may become dangerous as the estimations and inferences may be totally wrong (see chapter 4).

3.2.2 Statistic

A *statistic* is defined to be a function of a sample's random variables, i.e. $\Theta = g(X_1, \dots, X_n)$ (in vector notation, $\Theta = g(\boldsymbol{X})$, where $\boldsymbol{X} := [X_1, \dots, X_n]^T$ is known as the *sample vector*; note that the superscript T denotes the transpose of a vector or matrix). From the observations we can calculate the numerical value of the statistic, i.e. $\theta = g(x_1, \dots, x_n)$. Clearly, the statistic Θ is not identical with its numerical value θ. In particular, the statistic, as a function of random variables, is a random variable itself, having a certain distribution function. Whereas the numerical value of the statistic is simply calculated from the mathematical expression $g(x_1, \dots, x_n)$ using the sample observations, its distribution function is deducted based on theoretical considerations as we will see in later sections. Typical examples of commonly used statistics are given below.

3.2.3 Estimators and estimates

A statistics is used to estimate a population parameter. For any population parameter η, there exists one or more statistic of the form $\Theta = g(X_1, \dots, X_n)$ suitable for the estimation of this parameter. In this case we say that $\Theta = g(X_1, \dots, X_n)$ is an estimator of the parameter η and that the numerical value $\theta = g(x_1, \dots, x_n)$ is an estimate of η.

There is not a unique criterion to decide if a statistic can be used for the estimation of a population parameter. Often the mathematical expression $g(X_1, \dots, X_n)$ is formulated as if η was a population parameter of a finite sample space identical with the available sample. For example, if we wish to find an estimator of the mean value $\eta \equiv m_X$ of a variable X, based on the sample $(X_1, \dots, X_n)$ with observations $(x_1, \dots, x_n)$, we can think of the case where X is a discrete variable taking values $(x_1, \dots, x_n)$, each with the same probability $P(X = x_i) = 1/n$. In this case, by definition of the mean (eq. (2.21) - (2.23)) we find that $\eta = (x_1 + \cdots + x_n)/n$. If in the latter equation we replace the numerical values with the corresponding variables, we obtain the statistic $\Theta = (X_1 + \cdots + X_n)/n$. As we will see, this is the estimator of the mean value of any random variable, it is named *sample mean* and it is typically denoted as $\overline{X}$. However, this empirical approach does not give always a good estimator.

Whereas an estimator is not defined by in a strict mathematical procedure in the general case, several estimator categories have rigorous definitions. Thus:

1. A statistic Θ is an *unbiased* estimator of the parameter η if $E[\Theta] = \eta$. Otherwise, it is a biased estimator and the difference $E[\Theta] - \eta$ is called *bias*.
2. An estimator is Θ is a *consistent* estimator of the parameter η if the estimation error $\Theta - \eta$ tends to zero with probability 1 as $n \to \infty$. Otherwise, the estimator is inconsistent.
3. A statistic Θ is *the best* estimator of the parameter η if the mean square error $E[(\Theta - \eta)^2]$ is minimum.
4. A statistic Θ is the *most efficient* estimator of the parameter η if it is unbiased and with minimum variance (where due to unbiasednees the variance equals the estimation error).

It is easy to show that the estimator $\overline{X}$ of the previous example is an unbiased and consistent estimator of the population mean m_X (see section 3.3.1). Moreover, for certain distributions functions, it is best and most efficient.

In practice, efforts are taken to use unbiased and consistent estimators, while the calculation of the best and most effective estimator is more of theoretical interest. For a certain parameter it is possible to find more than one unbiased or consistent estimator. Often, the determination of unbiased estimators is difficult or impossible, and thus we may content with the use of biased estimators.

3.2.4 Interval estimation and confidence intervals

An *interval estimate* of a parameter η is an interval of the form (θ_1, θ_2), where $\theta_1 = g_1(x_1, \dots, x_n)$ and $\theta_2 = g_2(x_1, \dots, x_n)$ are functions of the sample observations. The interval (Θ_1, Θ_2) defined by the corresponding statistics $\Theta_1 = g_1(X_1, \dots, X_n)$ and $\Theta_2 = g_2(X_1, \dots, X_n)$ is called the interval estimator of the parameter η.

We say that the interval (Θ_1, Θ_2) is a *γ-confidence interval* of the parameter η if

$$P\{\Theta_1 < \eta < \Theta_2\} = \gamma \tag{3.1}$$

where γ is a given constant ($0 < \gamma < 1$) called the *confidence coefficient*, and the limits Θ_1 and Θ_2 are called *confidence limits*. Usually we choose values of γ near 1 (e.g. 0.9, 0.95, 0.99, so as the inequality in (3.1) to become near certain). In practice the term confidence limits is often (loosely) used to describe the numerical values of the statistics θ_1 and θ_2, whereas the same happens for the term confidence interval.

In order to provide a general manner for the calculation of a confidence interval, we will assume that the statistic $\Theta = g(X_1, \dots, X_n)$ is an unbiased point estimator of the parameter η and that its distribution function is $F_\Theta(\theta)$. Based on this distribution function it is possible to calculate two positive numbers ξ_1 and ξ_2, so that the estimation error $\Theta - \eta$ lies in the interval $(-\xi_1, \xi_2)$ with probability γ, i.e.

$$P\{\eta - \xi_1 < \Theta < \eta + \xi_2\} = \gamma \tag{3.2}$$

and at the same time the interval $(-\xi_1, \xi_2)$ to be the as small as possible.[*] Equation (3.2) can be written as

$$P\{\Theta - \xi_2 < \eta < \Theta + \xi_1\} = \gamma \tag{3.3}$$

Consequently, the confidence limits we are looking for are $\Theta_1 = \Theta - \xi_2$ and $\Theta_2 = \Theta + \xi_1$.

Although equations (3.2) and (3.3) are equivalent, their statistical interpretation is different. The former is a *prediction*, i.e. it gives the confidence interval[†] of the random variable Θ. The latter is a *parameter estimation*, i.e. it gives the confidence limits of the unknown parameter η, which is not a random variable.

3.3 Typical point estimators

In this section we present the most typical point estimators referring to the population moments of a random variable X irrespectively of its distribution function $F(x)$. Particularly, we give the estimators of the mean, the variance and the third central moment of a variable. We will not extend to higher order moments, firstly because it is difficult to form unbiased estimators and secondly because for typical sample sizes the variance of estimators is very high, thus making the estimates extremely uncertain. This is also the reason why in engineering applications moments higher than third order are not used. Even the estimation of the third moment is inaccurate for a small size sample. However, the third moment is an important characteristic of the variable as it describes the skewness of its distribution. Moreover, hydrological variables are as a rule positively skewed and thus an estimate of the skewness is necessary.

Apart form the aforementioned moment estimators we will present the L-moment estimators as well as the covariance and correlation coefficient estimators of two variables that are useful for the simultaneous statistical analysis of two (or more) variables.

3.3.1 Moment estimators

The estimators of raw moments (moments about the origin) of one or two variables, i.e. the estimators of $m_X^{(r)}$ and $m_{XY}^{(rs)}$ (where r and s are chosen integers), formed according to the empirical method described in section 3.2.3, are given by the following relationships:

[*] If the distribution of Q is symmetric then the interval $(-\xi_1, \xi_2)$ has minimum length for $\xi_1 = \xi_2$. For non-symmetric distributions, it is difficult to calculate the minimum interval, thus we simplify the problem by splitting the (3.2) into the equations $P\{\Theta < \eta - \xi_1\} = P\{\Theta > \eta + \xi_2\} = (1 - \gamma) / 2$.

[†] The terms confidence limits, confidence interval, confidence coefficient etc. are also used for this prediction form of the equation.

$$\tilde{M}_X^{(r)} = \frac{\sum_{i=1}^{n} X_i^r}{n}, \quad \tilde{M}_{XY}^{(rs)} = \frac{\sum_{i=1}^{n} X_i^r Y_i^s}{n} \tag{3.4}$$

It can be proved (Kendall and Stewart, 1968, p. 229) that

$$E\left[\tilde{M}_X^{(r)}\right] = m_X^{(r)}, \quad E\left[\tilde{M}_{XY}^{(rs)}\right] = m_{XY}^{(rs)} \tag{3.5}$$

Consequently, the moment estimators are unbiased. The variances of these estimators are

$$\mathrm{Var}\left[\tilde{M}_X^{(r)}\right] = \frac{1}{n}\left[m_X^{(2r)} - \left(m_X^{(r)}\right)^2\right], \quad \mathrm{Var}\left[\tilde{M}_{XY}^{(rs)}\right] = \frac{1}{n}\left[m_{XY}^{(2r,2s)} - \left(m_{XY}^{(rs)}\right)^2\right] \tag{3.6}$$

It can be observed that if the population moments are finite, then the variances tend to zero as $n \to \infty$; therefore the estimators are consistent.

Typical central moment estimators, i.e. estimators of $\mu_X^{(r)}$ and $\mu_{XY}^{(rs)}$ of one and two variables, respectively, are those defined by the equations

$$\hat{M}_X^{(r)} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^r}{n}, \quad \hat{M}_{XY}^{(rs)} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^r \left(Y_i - \bar{Y}\right)^s}{n} \tag{3.7}$$

These have been formed based on the empirical method described in section 3.2.3.These estimators are biased (for $r + s > 1$).

3.3.2 Sample mean

The most common statistic is the sample mean. As we have seen in section 3.2.3, the sample mean is an estimator of the true (or population) mean $m_X = E[X]$ and is defined by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.8}$$

which is a special case of (3.4) for $r = 1$. Its numerical value

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.9}$$

is called the *observed sample mean* or simply the *average*. The symbols $\bar{X}$ and $\bar{x}$ should not be conceptually confused with each other nor with the true mean of the random variable X, i.e. $m_X = E[X]$, which is defined based on the equations (2.20) or (2.21) and (2.23). Nevertheless, these three quantities are closely related. Implementation of equations (3.5) and (3.6) gives

$$E\left[\bar{X}\right] = E[X], \quad \mathrm{Var}\left[\bar{X}\right] = \frac{\mathrm{Var}[X]}{n} \tag{3.10}$$

regardless of the distribution function of X*. Thus, the estimator is unbiased and consistent.

[*] However, Var[$\bar{X}$] depends on the dependence structure of the variables X_i; the formula given in (3.10) holds only if X_i are independent. On the other hand, the formula for E[$\bar{X}$] holds always.

3.3.3 Variance and standard deviation

A biased estimator of the true (population) variance σ_X^2 is:

$$S_X^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n} \tag{3.11}$$

It can be proved (Kendall and Stewart, 1968, p. 277) that

$$\begin{gathered} E[S_X^2] = \frac{n-1}{n} \sigma_X^2 \\ \operatorname{Var}[S_X^2] = \frac{\mu_X^{(4)} - \sigma_X^4}{n} - \frac{2(\mu_X^{(4)} - 2\sigma_X^4)}{n^2} + \frac{\mu_X^{(4)} - 3\sigma_X^4}{n^3} \end{gathered} \tag{3.12}$$

where $\mu_X^{(4)}$ is the fourth central population moment. The two last terms in the expression of $\operatorname{Var}[S_X^2]$can be omitted for large values of *n.* From the expression of $E[S_X^2]$ in (3.12) we observe that multiplication of S_X^2 by *n*/(*n* – 1) results in an unbiased estimator of σ_X^2, i.e.

$$S_X^{*2} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} \tag{3.13}$$

S_X^{*2} is known as *sample variance*. For large sample sizes, the two estimators S_X^2 and S_X^{*2} are practically the same. If the population is normally distributed, it can be shown that

$$\operatorname{Var}[S_X^2] = \frac{2(n-1)\sigma_X^4}{n^2}, \quad \operatorname{Var}[S_X^{*2}] = \frac{2\sigma_X^4}{(n-1)} \tag{3.14}$$

The standard deviation estimators in common use are the square roots of the variance estimators, namely the S_X and S_X^* and are not unbiased. Thus (Yevjevich, 1972, p. 193· Kendall and Stewart, 1968, p. 233),

$$E[S_X] = \sigma_X + O\left(\frac{1}{n}\right), \quad \operatorname{Var}[S_X] = \frac{\mu_X^{(4)} - \sigma_X^4}{4\sigma_X^2 n} + O\left(\frac{1}{n^2}\right) \tag{3.15}$$

where the terms $O(1/n)$ are $O(1/n^2)$ are quantities proportional to $1/n$ and $1/n^2$, respectively, and can be omitted if the sample size is large enough ($n \geq 20$).

If the population is normally distributed, the following approximate equations can be used for S_X

$$E[S_X] \approx \sigma_X \sqrt{\frac{n-1}{n}}, \quad \operatorname{Var}[S_X] \approx \frac{\sigma_X^2}{2n} \tag{3.16}$$

For $n \geq 10$ the errors of these equations are smaller than 2.5% and 2.7% respectively, while for $n \geq 100$ are practically zero. The corresponding equations for S_X^* are[*]

$$E\left[S_X^*\right] \approx \sigma_X, \quad \mathrm{Var}\left[S_X^*\right] \approx \frac{\sigma_X^2}{2(n-1)} \tag{3.17}$$

Finally, one of the two following estimators of the coefficient of variation can be used:

$$\hat{C}_{v_X} = \frac{S_X}{\overline{X}}, \quad \hat{C}^*_{v\,X} = \frac{S_X^*}{\overline{X}} \tag{3.18}$$

If the variable X is positive, then it can be shown that these estimators are bounded from above ($\hat{C}_{v_X} \leq \sqrt{n-1}$) while the same does not hold for the corresponding population parameters. Obviously, this introduces bias.[†]

3.3.4 Third central moment and skewness coefficient

A biased estimator of the true (population) third central moment $\mu_X^{(3)}$ is given by

$$\hat{M}_X^{(3)} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{n} \tag{3.19}$$

for which it can be shown (Kendall and Stewart, p. 278-281) that

$$E\left[\hat{M}_X^{(3)}\right] = \frac{(n-1)(n-2)}{n^2}\mu_X^{(3)} \tag{3.20}$$

It immediately follows that an unbiased (and consistent) estimator of $\mu_X{}^{(3)}$ is

$$\hat{M}_X^{*(3)} = \frac{n\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{(n-1)(n-2)} \tag{3.21}$$

For large sample size n, the two estimators are practically the same.

The estimation of the skewness coefficient C_{s_X} is done using the following estimator

$$\hat{C}_{s_X} = \frac{\hat{M}_X^{(3)}}{S_X^3} \qquad 22)$$

[*] More accurate approximations are given by

$$E\left[S_X^*\right] \approx \sigma_X \sqrt{\frac{n-\frac{5}{4}}{n-\frac{3}{4}}}, \quad Var\left[S_X^*\right] \approx \frac{\sigma_X^2}{2\left(n-\frac{3}{4}\right)}$$

the errors of which for $n \geq 10$ are less than 0.005% and 0.2%, respectively. The precise equations are

$$E\left[S_X^*\right] = \sigma_X \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\frac{n-1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)}, \quad Var\left[S_X^*\right] = \sigma_X^2\left[1 - \frac{\Gamma^2\left(\frac{n}{2}\right)}{\frac{n-1}{2}\Gamma^2\left(\frac{n-1}{2}\right)}\right]$$

[†] The expression of the estimator's variance is quite complex and is omitted (cf. Kendall and Stewart, 1968, p. 233). If X follows a normal distribution then

$$\mathrm{Var}\left[\hat{C}_{v_X}\right] \approx C_{v_X}^2 / 2n$$

which is not unbiased. The bias does not originate only from the fact that the two moment estimators (numerator and denominator) are not unbiased themselves, but also (mainly) from the fact that $\hat{C}_{s_X}$ is bounded both from above and from below, whilst the population C_{s_X} is not bounded. This is due to the finite sample size *n*, which determines the upper and lower limit. Thus, it has been shown (Kirby, 1974; Wallis *et al.*, 1974) that $|\hat{C}_{s_X}| \le (n-2) / \sqrt{n-1}$.

Several approximate bias correction coefficients have been proposed in the literature to be multiplied by $\hat{C}_{s_X}$ estimated from (3.22) to obtain a less biased estimate. None of them leads to a rigorous unbiased estimator of the coefficient of skewness. The four most common are:

$$\frac{\sqrt{n(n-1)}}{n-2},\ \frac{n^2}{(n-1)(n-2)},\ \frac{\sqrt{n(n-1)}}{n-2}\left(1+\frac{8.5}{n}\right),\ 1+\left(\frac{6.51}{n}+\frac{20.20}{n^2}\right)+\left(\frac{1.48}{n}+\frac{6.77}{n^2}\right)\hat{C}^2_{s_X} \quad (3.23)$$

The first is obtained if in (3.22) the biased moment estimators are replaced by the unbiased ones. The second results if in (3.22) we replace the biased third moment estimator with the unbiased one (Yevjevich, 1978, p. 110). The third one has been proposed by Hazen and the last one has been proposed by Bobée and Robitaille (1975), based on results by Wallis *et al.* (1974).

3.3.5 L-moments estimates

Unbiased estimates $b_X^{(r)}$ of the probability weighted moments $\beta_X^{(r)}$ are given by the following relationship (Landwehr *et al.*, 1979):

$$b_X^{(r)} = \frac{1}{n}\frac{\sum_{i=1}^{n-r}\binom{n-j}{r}x_{(i)}}{\binom{n-1}{r}} = \frac{1}{n}\sum_{i=1}^{n-r}\frac{(n-i)(n-i-1)\dots(n-i+r+1)}{(n-1)(n-2)\dots(n-r)}x_{(i)} \quad (3.24)$$

where *n* is the sample size, and $x_{(i)}$ the ordered observations so that $x_{(n)} \le \dots \le x_{(2)} \le x_{(1)}$[*]. The estimates[†] of the first four probability weighted moments are:

$$b_X^{(0)} = \frac{1}{n}\sum_{i=1}^{n}x_{(i)} = \bar{x}$$

$$b_X^{(1)} = \frac{1}{n}\sum_{i=1}^{n-2}\frac{n-i}{n-1}x_{(i)}$$

$$b_X^{(2)} = \frac{1}{n}\sum_{i=1}^{n-2}\frac{(n-i)(n-i-1)}{(n-1)(n-2)}x_{(i)} \quad (3.25)$$

[*] Notice that $x_{(1)}$ is the largest observation; the equations are somewhat simpler if the observations are ordered from smallest to largest but it has been the rule in engineering hydrology to put the observations in descending order.

[†] The estimators of the same quantities are obtained by replacing $x_{(i)}$ with the variable $X_{(i)}$, the so called *order statistic*.

$$b_X^{(3)} = \frac{1}{n}\sum_{i=1}^{n-3}\frac{(n-i)(n-i-1)(n-i-2)}{(n-1)(n-2)(n-3)}x_{(i)}$$

Accordingly, the estimates of the first four L moments are calculated by the equations relating L moments and probability weighted moments (see equation (2.32)), i.e.,

$$\begin{aligned} l_X^{(1)} &= b_X^{(0)} \ (=\bar{x}) \\ l_X^{(2)} &= 2\, b_X^{(1)} - b_X^{(0)} \\ l_X^{(3)} &= 6 b_X^{(2)} - 6 b_X^{(1)} + b_X^{(0)} \\ l_X^{(4)} &= 20 b_X^{(3)} - 30 b_X^{(2)} + 12\, b_X^{(1)} - b_X^{(0)} \end{aligned} \tag{3.26}$$

3.3.6 Covariance and correlation

A biased estimator of the covariance σ_{XY} of two variables X and Y is:

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n} \tag{3.27}$$

It can be shown (e.g. Papoulis, 1990, p. 295) that

$$E[S_{XY}] = \frac{n-1}{n}\sigma_{XY} \tag{3.28}$$

Therefore, an unbiased (and consistent) estimator of σ_{XY} is

$$S_{XY}^* = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \tag{3.29}$$

known as *sample covariance*[*].

The estimator of the correlation coefficient ρ_{XY} is given by the next relationship, known as the *sample correlation coefficient*:

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{S_{XY}^*}{S_X^* S_Y^*} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{3.30}$$

The precise distribution function of this estimator and its moments are difficult to determine analytically; however, this estimator is regarded approximately unbiased.

[*] In many books, the denominator of (3.29) has the term $n - 2$, which is not correct.

3.4 Typical confidence intervals

3.4.1 Mean – known population variance

Let X be a random variable with mean μ_X and standard deviation σ_X. According to the central limit theorem and equation (3.10), the sample mean $\overline{X}$ (the average of n random variables) will have normal distribution $N\left(\mu_X, \sigma_X / \sqrt{n}\right)$, if n is large enough. Moreover, it will have precisely this normal distribution irrespectively of the size n, if the random variable X is normal.

The problem we wish to study here is the determination of the confidence intervals of the mean μ_X for confidence coefficient γ. We denote $z_{(1+\gamma)/2}$ the ((1+γ)/2)-quantile of the standard normal distribution $N(0, 1)$ (that is the value z that corresponds to non-exceedence probability (1+γ)/2). Apparently, due to symmetry, $z_{(1-\gamma)/2} = -z_{(1+\gamma)/2}$ (see Fig. 3.1). Thus,

$$P\left\{\mu_X - \frac{z_{(1+\gamma)/2}\sigma_X}{\sqrt{n}} < \overline{X} < \mu_X + \frac{z_{(1+\gamma)/2}\sigma_X}{\sqrt{n}}\right\} = \gamma \tag{3.31}$$

or equivalently

$$P\left\{\overline{X} - \frac{z_{(1+\gamma)/2}\sigma_X}{\sqrt{n}} < \mu_X < \overline{X} + \frac{z_{(1+\gamma)/2}\sigma_X}{\sqrt{n}}\right\} = \gamma \tag{3.32}$$

Equation (3.32) gives the confidence intervals sought. For the numerical evaluation we simply replace the estimator $\overline{X}$ in (3.32) with its numerical value $\overline{x}$.

For convenience, Table 3.1 displays the most commonly used confidence coefficients and the corresponding normal quantiles $z_{(1+\gamma)/2}$. We observe that as the confidence coefficient tends to 1, which means that the reliability of the estimate increases, the confidence interval becomes larger so that the estimate becomes more vague. On the contrary, if we choose a smaller confidence coefficient, a more "compact" estimate will result. In this case, the confidence interval will be narrower but the uncertainty will be higher.

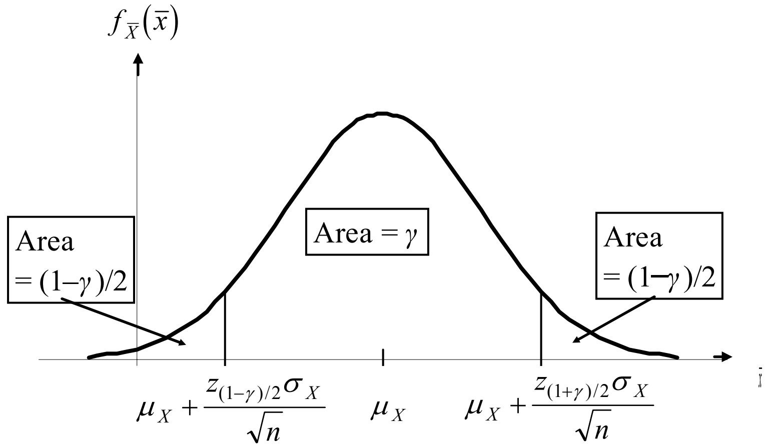
$f_{\overline{X}}(\overline{x})$
Area = (1−γ)/2
Area = γ
Area = (1−γ)/2
$\mu_X + \frac{z_{(1-\gamma)/2}\sigma_X}{\sqrt{n}}$
μ_X
$\mu_X + \frac{z_{(1+\gamma)/2}\sigma_X}{\sqrt{n}}$

Fig. 3.1 Explanatory sketch for the confidence intervals of the mean.

Table 3.1 Typical values of the normal quantiles $z_{(1+\gamma)/2}$ useful for the calculation of confidence intervals.

γ	0.90	0.95	0.99	0.999
$(1+\gamma)/2$	0.95	0.975	0.995	0.9995
$z_{(1+\gamma)/2}$	1.645	1.960	2.576	3.291

We observe from (3.32) that the only way to increase the accuracy without increasing the length of the confidence interval is to increase the sample size n by taking additional measurements.

The previous analysis was based on the assumption of known population variance, which in practice it is not realistic, because typically all our information comes from a sample. However, the results are of practical interest, since (3.32) provides a good approximation if the sample size n is large enough (> 30) and if we replace the population variance with its sample estimate.

3.4.2 Mean – unknown population variance

The analysis that we present here can be used for unknown population variance and for any sample size. However, this analysis has a restrictive condition, that the random variable X is normal, $N(\mu_X, \sigma_X)$. In this case the following conclusions can be drawn:

1. The sample mean has a normal distribution $N\left(\mu_X, \sigma_X / \sqrt{n}\right)$. This conclusion is a consequence of a basic property of the normal distribution, specifically the normal distribution is closed under addition or, else, a *stable* distribution.
2. The function of the sample variance $(n-1)S_X^{*2} / \sigma_X^2$ follows the $\chi^2(n-1)$ distribution. This is concluded by the theorem of section 2.10.4, according to which the sum of the squares of a number of standard normal variables follows the χ^2 distribution.
3. The random variables $\overline{X}$ and S_X^{*2} are independent. This results form a statistical theorem (see e.g. Papoulis, 1990, p. 222).
4. The ratio $(\overline{X} - \mu_X)/(S_X^* / \sqrt{n})$ follows the Student $t(n-1)$ distribution. This results by a theorem of 2.10.5.

We denote $t_{(1+\gamma)/2}$ the $[(1+\gamma)/2]$-quantile of the Student $t(n-1)$ distribution (that is the point t that corresponds to exceedence probability $(1+\gamma)/2$, for $n-1$ degrees of freedom). Because of the symmetry, $t_{(1-\gamma)/2} = -t_{(1+\gamma)/2}$. Thus,

$$P\left\{-t_{(1+\gamma)/2} < \frac{\overline{X} - \mu_X}{S_X^* / \sqrt{n}} < t_{(1+\gamma)/2}\right\} = \gamma \tag{3.33}$$

or equivalently

$$P\left\{\overline{X} - \frac{t_{(1+\gamma)/2}S_X^*}{\sqrt{n}} < \mu_X < \overline{X} + \frac{t_{(1+\gamma)/2}S_X^*}{\sqrt{n}}\right\} = \gamma \tag{3.34}$$

Equation (3.34) provides the confidence interval sought. For its numerical evaluation we simply replace in the interval estimators of (3.34) the estimators $\overline{X}$ and S_X^* with the corresponding sample estimates $\overline{x}$ and s_X^*.

Even though (3.32) and (3.34) are considerably different regarding their theoretical grounds and the assumptions they rely upon, from a computational perspective they are quite similar. Furthermore, for large *n* (>30) they practically coincide taking into account that $t_{(1+\gamma)/2} \approx z_{(1+\gamma)/2}$ (more precisely $t_{(1+\gamma)/2} \approx z_{(1+\gamma)/2}\sqrt{(n-1)/(n-3)}$, for $n - 1$ degrees of freedom).

The two previous analyses do not include the case of a small sample size, unknown variance and non-normal distribution. This case is not covered in statistics in a general and rigorous manner. However, as an approximation, often the same methodology is also used in these cases, provided that the population distribution is bell shaped and not too skewed. In general, the cases where precise confidence intervals can be determined based on a consistent theoretical procedure, are the exception rather than the rule. In most of the following problems we will use just approximations of the confidence intervals.

3.4.3 A numerical example of interval estimation of the mean

From a sample of annual inflows to a reservoir with length 15 (years), the sample mean is 10.05 hm^3 and the sample standard deviation 2.80 hm^3. We wish to determine (1) the 95% confidence interval of the annual inflow and (2) the sample size for 95% confidence coefficient that enables 10% precision in the estimation of the annual inflow.

(1) We assume that the annual inflows are IID with normal distribution (section 2.10.2) and we use the equation (3.34). Using the table of the Student distribution (Appendix A3) or any computational method (see section 2.10.5) we find that for $n - 1$ = 14 degrees of freedom $t_{(1+\gamma)/2} = t_{0.975}$ = 2.14. Consequently, the 95% confidence interval is*

$$10.05 - 2.14 \times 2.80/\sqrt{15} < \mu_X < 10.05 + 2.14 \times 2.80/\sqrt{15} \text{ (in hm}^3\text{)}$$

or

$$8.50 < \mu_X < 11.60 \text{ (in hm}^3\text{)}$$

For comparison, we will calculate the confidence interval using equation (3.32), even though this is not correct. From Table 3.1 we find $z_{(1+\gamma)/2} = z_{0.975}$ = 1.96. Thus, the 95% confidence interval is

* It would not be mathematically correct to write (3.34) replacing the estimators with their estimates, i.e.

$$P\{10.05 - 2.62 \times 2.80/\sqrt{15} < \mu_X < 10.05 - 2.62 \times 2.80/\sqrt{15}\} = 0.95$$

We note that μ_X is a (unknown) parameter (i.e. a number) and not a random variable, so it does not have a distribution function. Moreover, it is not correct to say e.g. that "with 95% probability the mean value lies in the interval (8.16, 11.94)". The correct expression would be "with 95% confidence".

$$10.05 - 1.96 \times 2.80/\sqrt{15} < \mu_X < 10.05 + 1.96 \times 2.80/\sqrt{15} \text{ (in hm}^3\text{)}$$

or

$$8.63 < \mu_X < 11.47 \text{ (in hm}^3\text{)}$$

The confidence interval is this case is a little smaller.

(2) Assuming that $n \geq 30$ we can use (3.32). The following equation must hold

$$1.96 \times 2.8 / \sqrt{n} = 10\% \times 10.05$$

so $n = 30$. We observe that the condition we have assumed ($n \geq 30$) is valid. (If it were not valid we should proceed with a trial-and-error procedure, using equation (3.34)).

3.4.4 Variance and standard deviation

As in the section 0, we will assume that the random variable X has a normal distribution $N(\mu_X, \sigma_X)$. As mentioned before, in this case the function of the sample variance $(n-1)S_X^{*2}/\sigma_X^2$ follows the $\chi^2(n-1)$ distribution.

We denote $\chi^2_{(1+\gamma)/2}$ and $\chi^2_{(1-\gamma)/2}$ the [(1+γ)/2]- and [(1–γ)/2]-quantiles, respectively, of the $\chi^2(n-1)$ distribution (the two are not equal because the χ^2 distribution is not symmetric). Thus, we have

$$P\left\{\chi^2_{(1-\gamma)/2} < \frac{(n-1)S_X^{*2}}{\sigma_X^2} < \chi^2_{(1+\gamma)/2}\right\} = \gamma \tag{3.35}$$

or equivalently

$$P\left\{\frac{(n-1)S_X^{*2}}{\chi^2_{(1+\gamma)/2}} < \sigma_X^2 < \frac{(n-1)S_X^{*2}}{\chi^2_{(1-\gamma)/2}}\right\} = \gamma \tag{3.36}$$

Equation (3.36) gives the confidence interval sought. It is easily obtained that confidence interval of the standard deviation is given by

$$P\left\{\frac{\sqrt{n-1}S_X^{*}}{\sqrt{\chi^2_{(1+\gamma)/2}}} < \sigma_X < \frac{\sqrt{n-1}S_X^{*}}{\sqrt{\chi^2_{(1-\gamma)/2}}}\right\} = \gamma \tag{3.37}$$

3.4.5 A numerical example of interval estimation of standard deviation

We wish to determine the 95% confidence interval of the standard deviation of annual inflow in the problem of section 3.4.3.

The sample standard deviation is 2.8 hm^3. With the assumption of normal distribution for the inflow, we utilize equation (3.37). Using the χ^2 distribution table (Appendix 2) or any computational method (see section 2.10.4) we find that for $n-1 = 14$ degrees of freedom $\chi^2_{(1+\gamma)/2} = \chi^2_{0.975} = 26.12$ and $\chi^2_{(1-\gamma)/2} = \chi^2_{0.025} = 5.63$. Thus, the 95% confidence interval is

$$\frac{\sqrt{14} * 2.80}{\sqrt{26.12}} < \sigma_X < \frac{\sqrt{14} * 2.80}{\sqrt{5.63}} \text{ (in hm}^3\text{)}$$

or

$$2.05 < \sigma_X < 4.41 \text{ (in hm}^3\text{)}$$

3.4.6 Normal distribution quantile – Standard error

In engineering design and management (in engineering hydrology in particular), the most frequent confidence interval problem that we face, concerns the estimation of design values for quantities that are modelled as random variables. For instance, in hydrological design we may wish to estimate the reservoir inflow that corresponds to a non-exceedence probability 1%, that is the 1% quantile of the inflow. Let X be a random variable with distribution $F_X(x)$ representing a natural quantity, e.g. a hydrological variable. Here we assume that $F_X(x)$ is a normal distribution $N(\mu_X, \sigma_X)$, which can be easily handled, whereas in Chapter 6 we will present similar methods for a repertoire of distributions being commonly used in engineering applications. For a given non-exceedence probability $u = F_X(x)$, the corresponding value of the variable X (symbolically x_u, the u-quantile) will be

$$x_u = \mu_X + z_u \sigma_X \tag{3.38}$$

where z_u the u-quantile of the standard normal distribution $N(0, 1)$. However, in this equation the population parameters μ_X and σ_X are unknown in practice. Using their point estimates, we obtain an estimate $\hat{x}_u = \bar{x} + z_u s_X$, that can be considered as a value of the random variable

$$\hat{X}_u = \bar{X} + z_u S_X \tag{3.39}$$

This latter equation can be used to determine the confidence interval of x_u. The precise determination is practically impossible, due to the complexity of the distribution function of $\hat{X}_u$. Here we will confine our analysis in seeking an approximate confidence interval, based on the assumption that $\hat{X}_u$ has normal distribution.

The mean of $\hat{X}_u$ is given from equation (2.59), which can be combined with (3.10) and (3.15) to give

$$E\left[\hat{X}_u\right] = E\left[\bar{X}\right] + z_u E\left[S_X\right] \approx \mu_X + z_u \sigma_X = x_u \tag{3.40}$$

assuming that n is large enough and omitting the term $O(1/n)$ in $E[S_X]$.* Likewise, the variance of $\hat{X}_u$ is given by equation (2.61), which can be written as

$$\operatorname{Var}\left[\hat{X}_u\right] = \operatorname{Var}\left[\bar{X}\right] + z_u^2 \operatorname{Var}\left[S_X\right] + 2 z_u \operatorname{Cov}\left[\bar{X}, S_X\right] \tag{3.41}$$

Given that X has normal distribution, the third term of (3.41) is zero (as mentioned before, the variables $\bar{X}$ and S_X are independent). Combining (3.10) and (3.16), we write (3.41) as

* The analysis here has an approximate character, thus we do not discriminate between the estimators S_X and S_X^*, because for n large enough the two estimators are virtually identical.

$$\varepsilon_u^2 := \operatorname{Var}\left[\hat{X}_u\right] \approx \frac{\sigma_X^2}{n} + z_u^2 \frac{\sigma_X^2}{2n} = \frac{\sigma_X^2}{n}\left(1 + \frac{z_u^2}{2}\right) \tag{3.42}$$

The quantity ε_u is known in literature as *standard quantile error* or simply as *standard error*.

Assuming that $\hat{X}_u$ has a normal distribution $N(x_u, \varepsilon_u)$ we can write

$$P\left\{-z_{(1+\gamma)/2} < \frac{\hat{X}_u - x_u}{\varepsilon_u} < z_{(1+\gamma)/2}\right\} = \gamma \tag{3.43}$$

where γ is the confidence coefficient. Equivalently,

$$P\left\{\hat{X}_u - z_{(1+\gamma)/2}\varepsilon_u < x_u < \hat{X}_u + z_{(1+\gamma)/2}\varepsilon_u\right\} = \gamma \tag{3.44}$$

Replacing in the previous equation the term ε_u from (3.42), and then the standard deviation σ_X with its estimator, we obtain the following final relationship

$$P\left\{\hat{X}_u - z_{(1+\gamma)/2}\sqrt{1 + \frac{z_u^2}{2}}\frac{S_X}{\sqrt{n}} < x_u < \hat{X}_u + z_{(1+\gamma)/2}\sqrt{1 + \frac{z_u^2}{2}}\frac{S_X}{\sqrt{n}}\right\} = \gamma \tag{3.45}$$

The latter equation is an approximation, whose accuracy is increased a n increases. Moreover, it is valid only in the case of normal distribution. However, (3.44) is also used for other distributions of the variable X, but with a different expression of the standard error ε_u and a different calculation method. The interested reader for a general expression of the standard error may consult Kite (1988, p. 33-38).

The estimates of the confidence limits are

$$\hat{x}_{u1} = \hat{x}_u - z_{(1+\gamma)/2}\sqrt{1 + \frac{z_u^2}{2}}\frac{s_X}{\sqrt{n}}, \quad \hat{x}_{u2} = \hat{x}_u + z_{(1+\gamma)/2}\sqrt{1 + \frac{z_u^2}{2}}\frac{s_X}{\sqrt{n}} \tag{3.46}$$

Clearly these estimates are functions of u or, equivalently, of the exceedence probability, 1 – u. The depictions of those functions in a probability plot placed on either side of the x_u curve are known as confidence curves of x_u.

3.4.7 A numerical example of interval estimation of distribution quantiles

Further to the numerical example of section 3.4.3, we wish to determine the 95% confidence interval of the annual inflow that has exceedence probability (a) 1% and (b) 99%. We note that, because of the small sample size, we will not expect a high degree of accuracy in our estimates (recall that the theoretical analysis assumed large sample size).

We will calculate first the point estimates (all units are hm^3). For the annual inflow with exceedence probability $F^* = 0.01$ we have $u = 1 - F^* = 0.99$ and $z_u = 2.326$. Thus, the point estimate of $\hat{x}_u = 10.05 + 2.326 \times 2.80 = 16.56$. Likewise, for the annual inflow with exceedence probability $F^* = 0.99$ we have $u = 1 - F^* = 0.01$ and $z_u = -2.326$, thus $\hat{x}_u = 10.05 - 2.326 \times 2.80 = 3.54$.

We can now proceed in the calculation of the confidence limits. For γ = 95% and $z_{(1+\gamma)/2}$ = 1.96, the limits for the inflow with exceedence probability 1% are:

$$\hat{x}_{u1} = 16.56 - 1.96\sqrt{1 + \frac{2.326^2}{2}}\frac{2.80}{\sqrt{15}} = 13.83$$

$$\hat{x}_{u2} = 16.56 + 1.96\sqrt{1 + \frac{2.326^2}{2}}\frac{2.80}{\sqrt{15}} = 19.29$$

Likewise, the limits for exceedence probability 99% are:

$$\hat{x}_{u1} = 3.54 - 1.96\sqrt{1 + \frac{2.326^2}{2}}\frac{2.80}{\sqrt{15}} = 0.81$$

$$\hat{x}_{u2} = 3.54 + 1.96\sqrt{1 + \frac{2.326^2}{2}}\frac{2.80}{\sqrt{15}} = 6.27$$

3.4.8 Correlation coefficient

To calculate the confidence limits of the correlation coefficient ρ of a population described by two variables X and Y, we use the auxiliary variable Z, defined by the so-called Fisher transformation:

$$Z = \frac{1}{2}\ln\frac{1+R}{1-R} \leftrightarrow R = \frac{e^{2Z}-1}{e^{2Z}+1} = \tanh Z \tag{3.47}$$

where R the sample correlation coefficient. We observe that for $-1 < R < 1$ the range of Z is $-\infty < Z < \infty$, while for $R = 0$, $Z = 0$. It can be shown that if X and Y are normally distributed, then Z has approximately normal distribution $N(\mu_Z, \sigma_Z)$ where

$$\mu_Z = E[Z] \approx \frac{1}{2}\ln\frac{1+\rho}{1-\rho}, \quad \sigma_Z^2 = \mathrm{Var}[Z] \approx \frac{1}{n-3} \tag{3.48}$$

As a consequence, if $\zeta_{(1+\gamma)/2}$ is the $(1+\gamma)/2$-quantile of the standard normal distribution, we obtain

$$P\left\{\mu_Z - \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}} < Z < \mu_Z + \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}}\right\} = \gamma \tag{3.49}$$

or equivalently

$$P\left\{Z - \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}} < \mu_Z < Z + \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}}\right\} = \gamma \tag{3.50}$$

Replacing μ_Z from (3.48) into (3.50) and solving for ρ, and also taking into account the monotonicity of the transformation (3.47), we obtain

$$P\{R_1 < \rho < R_2\} \tag{3.51}$$

where

$$R_1 = \frac{e^{2Z_1} - 1}{e^{2Z_1} + 1} \qquad R_2 = \frac{e^{2Z_2} - 1}{e^{2Z_2} + 1}$$
$$\left.\begin{matrix} Z_2 \\ Z_1 \end{matrix}\right\} = Z \pm \frac{\zeta_{(1+\gamma)/2}}{\sqrt{n-3}} \qquad Z = \frac{1}{2}\ln\frac{1+R}{1-R} \tag{3.52}$$

To numerically evaluate the confidence limits we implement equations (3.52) replacing the estimators with the corresponding estimates (e.g. $R = r$, etc.).

3.5 Parameter estimation of distribution functions

Assuming a random variable X with known distribution function and with unknown parameters θ_1, θ_2, …, θ_r we can denote the probability density function of X as a function $f_X(x; \theta_1, \theta_2, \ldots, \theta_r.)$. For known x, this, viewed as a function of the unknown parameters, is called the *likelihood function*. Here, we will examine the problem of the estimation of these parameters based on a sample X_1, X_2, …, X_n. Specifically, we will present the two most classical methods in statistics, namely the moments method and the maximum likelihood method. In addition, we will present a newer method that has become popular in hydrology, the method of L moments.

Several other general methods have been developed in statistics for parameter estimation, e.g. the maximum entropy method that has been also used in hydrology (the interested reader is referenced to Singh and Rajagopal, 1986). Moreover, in engineering hydrology in many cases, other types of methods like graphical, numerical, empirical and semi-empirical have been used. Examples of such methods will be given for certain distributions in chapter 6.

3.5.1 The method of moments

The method of moments is based on equating the theoretical moments of variable X with the corresponding sample moment estimates. Thus, if r is the number of the unknown parameters of the distribution, we can write r equations of the form

$$m_X^{(k)} = \hat{m}_X^{(k)}, \quad k = 1,2,\ldots,r \tag{3.53}$$

where $m_X^{(k)}$ are the theoretical raw moments, which are functions of the unknown parameters and are given by

$$m_X^{(k)} = \int_{-\infty}^{\infty} x^k f_X(x,\theta_1,\ldots,\theta_r)dx \tag{3.54}$$

whereas $\hat{m}_X^{(k)}$ are the estimates, calculated from the observed sample according to

$$\hat{m}_X^{(k)} = \frac{1}{n}\sum_{i=1}^{n} x_i^k \tag{3.55}$$

Thus, the solution of the resulting system of the r equations gives the unknown parameters θ_1, θ_2, …, θ_r. In general, the system of equations may not be linear and may not have an analytical solution. In this case the system can be solved only numerically.

Equivalently, we can use the central moments (for $k > 1$) instead of the raw moments. In this case, the system of equations is

$$\mu_X = \bar{x}, \quad \mu_X^{(k)} = \hat{\mu}_X^{(k)}, \quad k = 2, \dots, r \tag{3.56}$$

where $\mu_X = m_X^{(1)}$ is the population mean, $\bar{x} = \hat{m}_X^{(1)}$ the sample mean, $\mu_X^{(k)}$ the theoretical central moments given by the

$$\mu_X^{(k)} = \int_{-\infty}^{\infty} (x - \mu_X)^k f_X(x; \theta_1, \dots, \theta_r) dx \tag{3.57}$$

and $\hat{\mu}_X^{(k)}$ the corresponding sample estimates calculated by the relationship

$$\hat{\mu}_X^{(k)} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k \tag{3.58}$$

We recall that the raw moments in (3.55) are unbiased estimates, while the central moments in (3.58) are biased estimates. Nevertheless, unbiased central moment estimates are often used instead of the biased. Regardless of using biased or unbiased estimates for moments, in general the method of moments does not result in unbiased estimates of the parameters θ_1, θ_2, …, θ_r (except in special cases).

3.5.2 Demonstration of the method of moments for the normal distribution

As an example of the implementation of the method of moments, we will calculate the parameters of the normal distribution. The probability density function is:

$$f_X(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.59}$$

and has two parameter, μ and σ. Thus, we need two equations. Based on (3.56), these equations are

$$\mu_X = \bar{x}, \quad \sigma_X^2 = s_X^2 \tag{3.60}$$

where in the latter equation we have denoted the theoretical and sample variance (that is, the second central moment) of X, by the more common symbols σ_X^2 and s_X^2, respectively. We know (see section 2.10.2) that the theoretical moments are

$$\mu_X = \mu, \quad \sigma_X^2 = \sigma^2 \tag{3.61}$$

Consequently, the final estimates are

$$\mu = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \sigma = s_X = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3.62}$$

This estimation of σ biased, while that of μ is unbiased. If in the last equation we used the unbiased estimate of the variance, then we would have in the denominator (n-1) instead of n. Even in this case, the estimate of the σ would not be unbiased, for reasons explained in section 3.3.3.

As we have seen in this example, the application of the method of moments is very simple and this extends to other distribution functions.

3.5.3 The method of L moments

The logic of the method of L moments is the same as in the method of moments. If the distribution has r unknown parameters, we can write r equations of the form

$$\lambda_k = l_k,\ k = 1, 2, \ldots, r \tag{3.63}$$

where λ_k are the theoretical L-moments, which are functions of the unknown parameters, and l_k their sample estimates. Solving this system of equations we obtain the L-moment estimates of the unknown parameters of the distribution. Because L moments are linear combinations of the probability weighted moments, (3.63) can be written equivalently as

$$\beta_k = b_k,\ k = 0, 2, \ldots, r - 1 \tag{3.64}$$

where β_k is the probability weighted moment of order k and b_k is its estimate (see section 3.3.5).

Estimates based on L-moments are generally more reliable than those based on classical moments. Moreover, the L-moment estimators have some statistically desirable properties e.g. they are robust with respect to outliers, because contrary to standard moments, they do not involve squaring, cubing, etc., of the sample observations. In hydrology, the L moments have been widely used as descriptive statistics and in parameter estimation of several distributions. Examples of applications of the method can be found, among others, in Kjeldsen *et al.* (2002), Kroll and Vogel (2002), Lim and Lye (2003) and Zaidman *et al.* (2003).

3.5.4 The maximum likelihood method

Let X be a random variable with probability function $f_X(x, \theta_1, \theta_2, \ldots, \theta_r)$ where $\theta_1, \theta_2, \ldots, \theta_r$ are parameters, and $X_1, X_2, \ldots, X_n$ a sample of the variable. Let $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots\theta_r)$ be the joint distribution function of the sample vector $\boldsymbol{X} := [X_1, X_2, \ldots, X_n]^T$. Our entire observed sample can be thought of as single observation of the vector variable $\boldsymbol{X}$. The idea behind the maximum likelihood method is that the probability density $f_{X_1,\ldots,X_n}(\ )$ at this single point will be as high as possible (it is natural to expect an observation to lie in an area with high probability density). We can thus find $\theta_1, \theta_2, \ldots, \theta_r$, so that the function $f_{X_1,\ldots,X_n}(\ )$ have a value as high as possible at the point $(x_1, x_2, \ldots, x_n)$.

In a random sample, the variables $X_1, X_2, \ldots, X_n$ are independent and the joint probability density function is

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots\theta_r)=\prod_{i=1}^{n} f_X(x_i,\theta_1,\ldots\theta_r) \tag{3.65}$$

and, viewed as a function of the parameters θ_1, θ_2, …, θ_r (for values of random variables equal to the observations x_1, …, x_n) is the likelihood function of these parameters.

Assuming that $f_{X_1,\ldots,X_n}(\ )$ is differentiable with respect to its parameters, the condition that maximizes it is

$$\frac{\partial f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_r)}{\partial\theta_k}=0,\quad k=1,\ldots,r \tag{3.66}$$

Using these r equations, the r unknown parameters will result. However, the manipulation of these equations may be complicated and, instead of maximizing the likelihood, we may attempt to maximize its logarithm

$$L(x_1,\ldots,x_n;\theta_1,\ldots\theta_r):=\ln f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots\theta_r)=\sum_{i=1}^{n}\ln f_X(x_i;\theta_1,\ldots\theta_r) \tag{3.67}$$

The function $L(\ )$ is called the log-likelihood function of the parameters θ_1, θ_2, …, θ_r. In this case, the condition of maximum is

$$\frac{\partial L(x_1,\ldots,x_n;\theta_1,\ldots\theta_r)}{\partial\theta_k}=\sum_{i=1}^{n}\frac{1}{f_X(x_i;\theta_1,\ldots\theta_r)}\frac{\partial f_X(x_i;\theta_1,\ldots\theta_r)}{\partial\theta_k}=0,\quad k=1,\ldots,r \tag{3.68}$$

Solving these r equations we obtain the values of the r unknown parameters.

3.5.5 Demonstration of the maximum likelihood method for the normal distribution

We will calculate the parameters of the normal distribution using the maximum likelihood method. The probability density function of the normal distribution is

$$f_X(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.69}$$

Based on (3.65) we form the likelihood function

$$f_X(x_1,\ldots,x_n;\mu,\sigma)=\frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \tag{3.70}$$

and taking its logarithm we form the log-likelihood function:

$$L(x_1,\ldots,x_n;\mu,\sigma)=-\frac{n}{2}\ln(2\pi)-n\ln\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{3.71}$$

Taking the derivatives with respect of the unknown parameters μ and σ and equating them to 0 we have

$$\frac{\partial L}{\partial\mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-m)=0,\quad \frac{\partial L}{\partial\sigma}=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2=0 \tag{3.72}$$

and solving the system we obtain the final parameter estimates:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}, \qquad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = s_X \tag{3.73}$$

The results are precisely identical with those of the section 3.5.2, despite the fact that the two methods are fundamentally different. The application of the maximum likelihood method is more complex than that of the method of moments. The identical results found here are not the rule for all distribution functions. On the contrary, in most cases the two methods yield different results.

3.6 Hypothesis testing

A statistical hypothesis is a hypothesis related to the value of one or more parameters of a statistical model, which is described by a distribution function. The hypothesis testing is a process of establishing the validity of a hypothesis. The process has two possible outcomes: either the hypothesis is rejected or accepted (more precisely: not rejected).

In this section we present very briefly the related terminology and procedure, while in next chapters we will present some applications. The reader interested for a more detailed presentation of the theory should consult statistics books (e.g. Papoulis, 1990, p. 321-387, Freund *et al.*, 1988, p. 310-542), while for a presentation for hydrological applications is referenced to Hirsch *et al.* (1993, p. 17.11-29).

3.6.1 Terminology

- *Null hypothesis* is the hypothesis to be tested (symbolically H_0). Usually, it is a hypothesis of the form $\theta = \theta_0$, where θ is parameter related to a distribution function of a given variable and θ_0 is a numerical value.

- *Alternative hypothesis* is a second hypothesis that should not be true at the same time with the null hypothesis (symbolically H_1). It can be *simple*, such as $\theta = \theta_1$, or (more commonly) *composite*, such as $\theta \neq \theta_0$, $\theta > \theta_0$ or $\theta < \theta_0$.

- *Test statistic* is an appropriately chosen sample statistic, that is used for the test (symbolically Q).

- *Critical region* is an interval of real values. When the test statistic value lies in the critical region then the null hypothesis is rejected (symbolically R_c; see Fig. 3.2).

- *One-sided test* is a test where the alternative hypothesis is of the form $\theta > \theta_0$ or $\theta < \theta_0$. In this case the critical region is a half line of the form ($q > q_C$) or ($q < q_C$), respectively.

- *Two-side test* is a test where the alternative hypothesis if of the form $\theta \neq \theta_0$. In this case the critical region consists of two half lines ($q < q_L$) and ($q > q_U$).

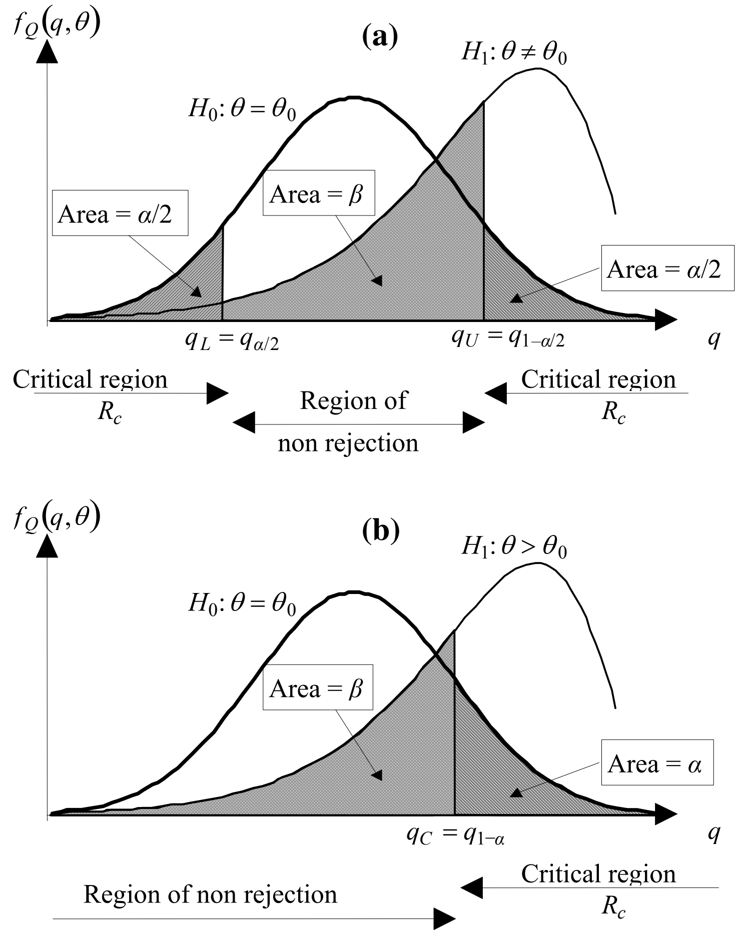

Fig. 3.2 Explanatory sketch of the concepts related to statistical testing: (a) a two-sided test, (b) an one-sided test.

- *Parametric* is a test whose hypotheses include specification of the population distribution function.
- *Non parametric* is a test valid for every population distribution function.
- *Decision rule* is the rule to reject or not the null hypothesis, expressed as:

$$\text{reject } H_0 \text{ if } q \in R_c$$

- *Type I error* is the rejection (based on the decision rule) of a true null hypothesis.
- *Type II error* is the acceptance (based on the decision rule) of a false null hypothesis
- *Significance level* of a test is the probability of type I error, namely the probability to reject a true null hypothesis. Symbolically

$$\alpha = P\{Q \in R_c \mid H_0\} \tag{3.74}$$

- *Power* of a test is the probability of rejecting a false null hypothesis. Symbolically,

$$p = 1 - \beta = P\{Q \in R_c \mid H_1\} \tag{3.75}$$

where β is the probability of type II error, that is

$$\beta = P\{Q \notin R_c \mid H_1\} \tag{3.76}$$

3.6.2 Testing procedure

The testing procedure consists of the following steps:

1. Formulation of the null hypothesis H_0 and of the alternative H_1.
2. Choice of the test statistic $Q = g(X_1, \ldots, X_n)$ and determination of the probability density function of the $f_Q(q; \theta)$.
3. Choice of the significance level α of the test and determination of the critical region R_c.
4. Calculation of the value $q = g(x_1, \ldots, x_n)$ of Q from the sample.
5. Application of the decision rule and rejection or acceptance of H_0.
6. Calculation of the power p of the test.

The last step is usually omitted in practice, due to its complexity. All remaining steps are clarified in the following section.

3.6.3 Demonstration of significance testing for the correlation coefficient

As an example of the above procedure we will present the significance testing of the correlation coefficient of two random variables X and Y, according to which we can decide whether or not the variables are linearly correlated.

If the variables are not linearly correlated then their correlation coefficient will be zero. Based on this observation, we proceed in the following steps of the statistical testing.

1. The null hypothesis H_0 is $\rho = 0$ and the alternative hypothesis H_1 is $\rho \neq 0$. As a consequence we will proceed with a two-sides test. (If we wanted to decide on the type of correlation, positive or negative, the alternative hypothesis would be formulated as $\rho > 0$ or $\rho < 0$, and we would perform an one-sided test).
2. We choose the test statistic as

   $$Q = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)\sqrt{n-3} = Z\sqrt{n-3} \tag{3.77}$$

   where R is the sample correlation coefficient and Z is the auxiliary Fisher variable (section 3.3.6), which, if H_0 is true, has approximately a normal distribution with mean 0 and standard deviation $1/\sqrt{n-3}$. Consequently, Q has standard normal distribution $N(0, 1)$.
3. We choose a significance level $\alpha = 0.05$. If $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the normal distribution, then the corresponding critical region R_c is the $|q| > z_{1-\alpha/2}$ or $|q| > z_{0.975}$, or finally $|q| > 1.96$, given that

$$P(|Q| > z_{1-\alpha/2}) = P(Q < -z_{1-\alpha/2}) + P(Q > z_{1-\alpha/2})$$
$$= 2P(Q < z_{\alpha/2}) = 2\ \alpha\ /\ 2 = \alpha$$

(We recall that, because of the symmetry of the normal probability density function, $z_{1-u} = z_u$. In the case of the one-side test with alternative hypothesis $\rho > 0$, the critical region would be $q > z_{1-\alpha}$).

4. The numerical value of q is determined from the observed sample by the following equations

$$q = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)\sqrt{n-3}, \quad r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.78}$$

5. The decision rule will be:

$$\text{reject } H_0 \text{ if } |q| > z_{1-\alpha/2}$$

and for $\alpha = 0.05$

$$\text{reject } H_0 \text{ if } |q| = \frac{1}{2}\left|\ln\left(\frac{1+r}{1-r}\right)\right|\sqrt{n-3} > 1.96$$

At the limit of this inequality, solving for r, we find the critical value r_c of the sample correlation coefficient, that determines the critical region R_c of the statistic R, that is,

$$r_c = \frac{e^{3.92/\sqrt{n-3}} - 1}{e^{3.92/\sqrt{n-3}} + 1} \tag{3.79}$$

A simple formula easy to remember that provides a very good approximation of (3.79) is:

$$r_c \approx \frac{2}{\sqrt{n}} \tag{3.80}$$

As a consequence, we can conduct the hypothesis testing in a more direct manner, by comparing the absolute value of r with the critical value r_c. If $|r| > r_\text{c}$ then we conclude that there is statistically significant correlation between the two variables.

3.6.4 A numerical example of significance testing for the correlation coefficient

From a 18-year-long record of measurements of concurrent annual rainfall and runoff at a catchment, we have calculated the correlation coefficient equal to 0.58. Is there a linear correlation between the annual rainfall and runoff?

We calculate the critical value r_c using one of (3.79) or (3.80). Here for comparison we use both:

$$r_c = \frac{e^{3.92/\sqrt{15}} - 1}{e^{3.92/\sqrt{15}} + 1} = 0.470, \quad r_c = \frac{2}{\sqrt{18}} = 0.471$$

Indeed, the two equations give practically the same result. Since 0.58 > 0.47 we conclude that there is statistically significant correlation between the annual rainfall and runoff.

Acknowledgement I thank Simon Papalexiou for translating into English an earlier Greek version of this text.

References

Bobée, B., and R. Robitaille, Correction of bias in the estimation of the coefficient of skewness, *Water Resour. Res.*, 11(6), 851-854, 1975.

Freund, J. E., F. J. Williams, and B. M. Perles, *Elementary Business Statistics, The Modern Approach*, Prentice-Hall, 1988.

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

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

Kite, G. W., *Frequency and Risk Analyses in Hydrology*, Water Resources Publications, Littleton, Colorado, 1988.

Kirby, W., Algebraic boundness of sample statistics, *Water Resour. Res.*, 10(2), 220-222, 1974.

Kjeldsen, T.R., J.C. Smithers and R.E. Schulze, Regional flood frequency analysis province, South Africa, using the index-flood method, *J. Hydrol.*, 255, 194–211, 2002.

Kroll, C.N., and R.M. Vogel, Probability distribution of low streamflow series in the United States, *J. Hydrol. Eng.*, 7, 137–146, 2002.

Landwehr, J.M., N.C. Matalas and J.R. Wallis, Probability weighted moments with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resour. Res.*, 15, 1055–1064, 1979.

Lim, Y.H., and L.M. Lye, Regional flood estimation for ungauged basins in Sarawak, Malaysia, *Hydrol. Sci. J.*, 48, 79–94, 2003.

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

Singh, V. P., and A. K. Rajagopal, A new method of parameter estimation for hydrologic frequency analysis, *Hydrological Science and Technology*, 2(3) 33-44, 1986.

Wallis, J. R., N. C. Matalas, and J. R. Slack, Just a moment!, *Water Resour. Res.*, 10(2), 211-219, 1974.

Yevjevich, V., *Probability and Statistics in Hydrology*, Water Resources Publications, Fort Collins, Colorado, 1972.

Zaidman, M.D., V. Keller, A.R. Young and D. Cadman, Flow-duration-frequency behaviour of British rivers based on annual minima data, *J. Hydrol.*, 277, 195–213, 2003.

Chapter 4

Special concepts of probability theory in geophysical applications

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

Summary

Geophysical processes (and hydrological processes in particular, which are the focus of this text) are usually modeled as stochastic processes. However, they exhibit several peculiarities, which make classical statistical tools inappropriate, unless several simplifications are done. Typical simplifications include time discetization at the annual time scale and selection of annual maxima and minima in a manner which eliminates the effect of the annual cycle and effectively reduces dependence, which always exists in geophysical processes evolving in continuous time. These simplifications allow us to treat certain geophysical quantities as independent random variables and observed time series as random samples, and then perform typical statistical tasks are using classical statistics. In turn, they allow convenient handling of concepts such as return period and risk, which are essential in engineering design. However, we should be aware that the independence assumption has certain limits and that dependence cannot be eliminated as natural processes are characterized by large-scale persistence, or more rarely antipersistence, which are manifestations of strong dependence in time.

4.1 General properties of probabilistic description of geophysical processes

In a probability theoretic approach, geophysical processes (and hydrological processes in particular, which are the focus of this text) are modeled as stochastic processes. For example, the river discharge $X(t)$ in a specific location at time t is represented as a random variable and thus, for varying time t, $X(t)$ makes a family of random variables, or a stochastic process, according to the definition given in chapter 2. More specifically, $X(t)$ is a continuous state and continuous time stochastic process, and a sequence of observations of the discharge at regular times is a time series.

Some clarifications are necessary to avoid misconceptions with regard to the introduction of the notion of a stochastic process to represent a natural process. The stochastic process is a mathematical model of the natural process and it is important to distinguish the two. For instance, once we have constructed the mathematical model, we can construct an ensemble of as many synthetic "realizations" (time series) of the stochastic process as we wish. In contrast, the natural process has a unique evolution and its observation can provide a single time series only.

In addition, the adoption of a probabilistic model, a stochastic process, does not mean that we refuse causality in the natural process or that we accept that natural phenomena happen spontaneously. We simply wish to describe the uncertainty, a feature intrinsic in natural processes, in an effective manner, which is provided by the probability theory. All deterministic controls that are present in the natural process are typically included in the stochastic description. For instance, most geophysical quantities display periodic fluctuations, which are caused by the annual cycle of earth, which affects all meteorological phenomena.

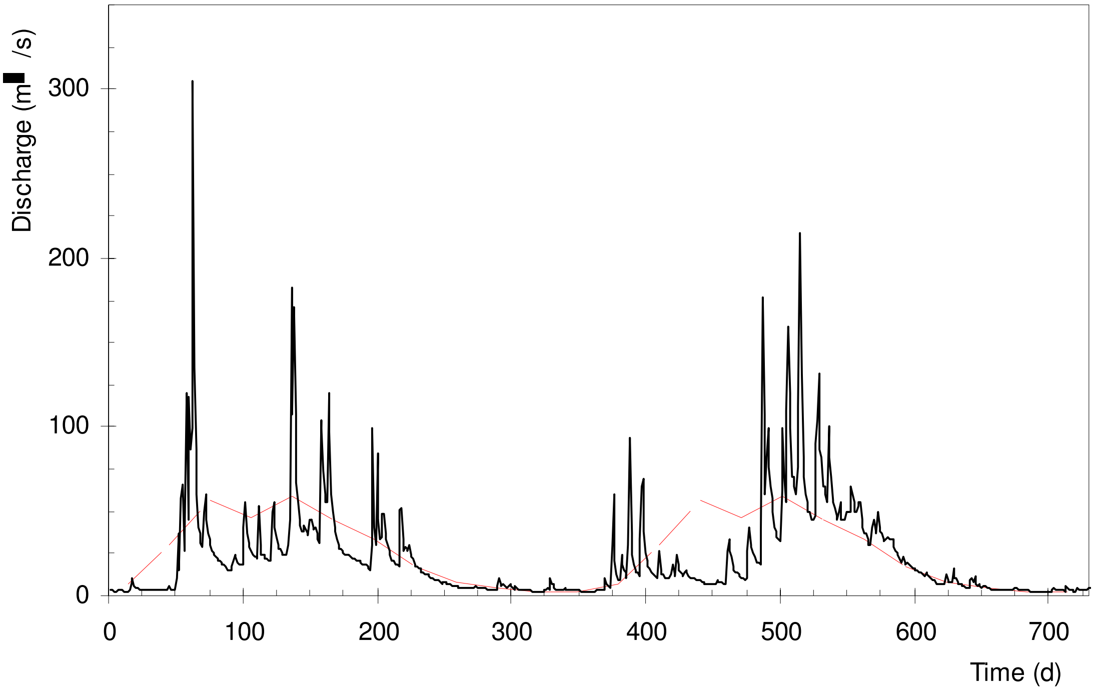

Fig. 4.1 Daily discharge of the Evinos River, Western Greece, at the Poros Reganiou gauge (hydrological years 1971-72 and 1972-73 – zero time is 1971/10/01). Dashed line shows the average monthly discharge of each month, estimated from a time series extending from 1970-71 to 1989-90.

An example is depicted Fig. 4.1, which shows the evolution of discharge of a river for a two-year period, where the annual cycle is apparent. A stochastic model can well incorporate the periodicity in an appropriate manner. This is typically done by constructing a *cyclostationary*, rather than a stationary, stochastic process (see chapter 2). Some authors have suggested that the process should be decomposed into two additive components, i.e. $X(t) = d(t) + \Xi(t)$, where $d(t)$ is a deterministic periodical function and $\Xi(t)$ is a stationary stochastic component. This, however, is a naïve approach, which adopts a simplistic view of natural phenomena of the type "actual" = "deterministic" + "stochastic". Stochastic theory provides much more powerful cyclostationary methodologies, whose presentation, however, are is of the scope of this text. Another common misconception (e.g. Haan, 1977; Kottegoda, 1980) is that deterministic components include the so-called "trends", which are either increasing or decreasing, typically linear, deterministic functions of time. Whilst it is true that geophysical (and hydrological in particular) time series display such "trends" for long periods, these are not deterministic components unless there exists a deterministic theory that could predict them in advance (not after their observation in a time series). Such "trends", after some time change direction (the increasing become decreasing and vice versa) in an irregular manner. In other words, typically they are parts of large-scale irregular fluctuations (Koutsoyiannis, 2006a).

We use the term "stochastic" instead of "random" in the mathematical process to stress the fact that our model does not assume pure randomness in the evolution of the natural process under study. In contrast, a stochastic model assumes that there is stochastic dependence between variables $X(t)$ that correspond to neighbouring times. Using the terminology of chapter 2, we say that the process has non negligible autocovariance or autocorrelation. Generally, these are decreasing functions of time lag but they sustain very high values for small lags. For example, if the discharge of a river at time t_0 is $X(t_0)$ = 500 m^3/s, it is very improbable that, after a small time interval Δt, say 1 hour, the discharge becomes $X(t_0 + \Delta t)$ = 0.5 m^3/s. On the contrary, it is very likely that this discharge will be close to 500 m^3/s and this is expressed by a high autocorrelation at a lag of 1 hour.

While the dependence of this type is easily understandable and is called *short-range dependence* or *short-term persistence*, hydrological and other geophysical processes (and not only) display another type of dependence, known as *long-range dependence* or *long-term persistence*. Thus, it is not uncommon that long time series of hydrological and other geophysical processes display significant autocorrelations for large time lags, e.g. 50 or 100 years. This property is related to the tendency of geophysical variables to stay above or below their mean for long periods (long period excursions from means), observed for the first time by Hurst (1951), and thus also known as the Hurst phenomenon. Another name for the same behaviour, inspired from the clustering of seven year drought or flood periods mentioned in the Bible, is the *Joseph effect* (Mandebrot, 1977). Koutsoyiannis (2002, 2006a) has demonstrated that this dependence is equivalent to the existence of multiple time scale fluctuations of geophysical processes, which, as mentioned above, were regarded earlier as deterministic trends. The long-term persistence will be further discussed in section 4.5.

Apart from periodicity (seasonality) and long-term persistence, geophysical processes have also other peculiarities that make classical statistical and stochastic models inappropriate for many modelling tasks. Among these are the intermittency and the long tails of distributions. Intermittency is visible in the flow time series of Fig. 4.1, where the flow alternates between two states, the regular flow (base flow) state and the flood state. In rainfall (as well as in the flow in ephemeral streams) this switch of states is even more apparent, as most of the time the processes are at a zero (dry) state. This is manifested in the marginal probability distribution of rainfall depth by a discontinuity at zero. Furthermore, the distribution functions **o**f geophysical processes are quite skewed on fine and intermediate time scales. The skewness is mainly caused by the fact that geophysical variables are non-negative and sometimes intermittent. This is not so common in other scientific fields whose processes are typically Gaussian. While at their lower end probability distributions of geophysical variables have a lower bound (usually zero), on the other end they are unbounded. Moreover, their densities $f(x)$ tend to zero, as state x tends to infinity, much more slowly than the typical exponential-type distributions, to which the normal distribution belongs. This gives rise to the long tails, which practically result in much more frequent extreme events than predicted by the typical exponential type models, a phenomenon sometimes called the Noah effect (Mandebrot, 1977).

4.2 Typical simplifications for geophysical applications

4.2.1 Processes in discrete time

The study of a geophysical process in continuous time is difficult and, in most practical engineering problems, not necessary. The continuous time description of geophysical processes is out of the scope of this text, which focuses on discrete time representation. However, the discrete time representation, requires consistency with the continuous time evolution of the actual processes. To establish this consistency we need two characteristic time steps. The first, D, is fixed to the duration of the year, in which a full cycle of geophysical phenomena is completed. In hydrology, the partitioning of the continuous time in years is done using the convention of a *hydrological year*, whose starting point does not generally coincide with that of a calendar year. Rather it is taken to be the beginning of the rainy period of the year. In Europe, this is typically regarded to be the 1st of October. The second time step, Δ, defines a time window, or *time scale*, within which we view the process. In contrast to the year, this is not fixed but depends on the specific problem we study. It can range from a few minutes, if we study storms and floods in an urban area, to one year, if we study the hydrological balance of a catchment, or to many years, if we study overannual fluctuations of water balance.

Now we can proceed in several simplifications of a continuous time stochastic process representing a geophysical (hydrological in particular) process, as demonstrated in Fig. 4.2, where time in horizontal axis is measured in (hydrological) years whereas for demonstration purposes it was assumed $\Delta = D/4$. The first simplification of the full continuous time process (Fig. 4.2(1)) is the formation of a discrete time process (Fig. 4.2(2)). To this aim we partition continuous time t in intervals of length Δ. The values $i = 1, 2, \ldots$, of discrete time correspond to continuous time intervals $[0, \Delta)$, $[\Delta, 2\Delta)$, and so no. The discrete time process $X_\Delta(i)$ in time i is defined to be the time average of $X(t)$ in the respective interval, i.e.

$$X_\Delta(i) := \frac{1}{\Delta} \int_{(i-1)\Delta}^{i\Delta} X(t) dt \tag{4.1}$$

For instance, if $X(t)$ represents the instantaneous discharge of a river and $\varDelta$ is taken to be one day or one month, then $X_{\varDelta}(i)$ represents the daily (more rigorously: the time averaged daily) or the monthly (more rigorously: the time averaged monthly) discharge, respectively. Sometimes, we wish to study the aggregated quantity, rather than the time average, in the corresponding time interval $\varDelta$, that is, the quantity

$$X_{\varDelta}^{*}(i) := \int_{(i-1)\varDelta}^{i\varDelta} X(t)dt \tag{4.2}$$

In this example, $X^{*}_{\varDelta}(i)$ represents the daily or the monthly runoff volume. Likewise, if $X(t)$ represents the instantaneous rainfall intensity in a specific point of a catchment and $\varDelta$ is taken as one day or one month, then $X^{*}_{\varDelta}(i)$ represents the daily or the monthly rainfall depth, respectively.

Even though time discretization is a step toward simplification of the study of a geophysical process, yet the mathematical description of $X_{\varDelta}(i)$ or $X^{*}_{\varDelta}(i)$ is complicated as it requires the analysis of periodicity and the autocorrelation of the process, for which the classical statistics, summarized in chapter 3, do not suffice. These issues are not covered in this text, except a few general discussions in the end of this chapter. The following simplifications, which are typical and useful in engineering problems, are more drastic and the resulting processes are easier to study using classical statistics.

If we construct the process $X_{\varDelta}(i)$ (or $X^{*}_{\varDelta}(i)$) assuming a time window equal to one year ($\varDelta = D$) then we obtain the *annual process*, $X_D(i)$ (or $X^{*}_{D}(i)$); now i denotes discrete time in years (Fig. 4.2(3)). Thus,

$$X_D(i) := \frac{1}{D}\int_{(i-1)D}^{iD} X(t)dt, \quad X_D^{*}(i) := \int_{(i-1)D}^{iD} X(t)dt \tag{4.3}$$

In this process the annual periodicity has been fully eliminated, because time intervals smaller than a year are not visible, and the process autocorrelation has been reduced significantly (but not eliminated), because of the large integration time step. This process, which represents the succession of an annual hydrological quantity (annual runoff, rainfall, evaporation, temperature) is very useful for problems of estimation of the water potential of an area.

One way to move to a time interval smaller than a year, simultaneously eliminating the annual periodicity and significantly reducing autocorrelation is shown in Fig. 4.2(4). In each hydrological year $i = 1, 2, \ldots$, we take an interval of length $\varDelta < D$, specifically the interval $[(i-1)D + (j-1)\varDelta, (i-1)D + j\varDelta)$. Here j is a specified integer with possible values $j = 1, 2, \ldots, D/\varDelta$ (in Fig. 4.2(4) it has been assumed $j = 1$). The process obtained is:

$$Y_{\varDelta}(i) := \frac{1}{\varDelta}\int_{(i-1)D+(j-1)\varDelta}^{(i-1)D+j\varDelta} X(t)dt, \quad Y_{\varDelta}^{*}(i) := \int_{(i-1)D+(j-1)\varDelta}^{(i-1)D+j\varDelta} X(t)dt \tag{4.4}$$

For instance if *X*(*t*) represents the instantaneous discharge of a river, *Δ* is taken as one month, and *j* = 1, then $Y_Δ(i)$ represents the average monthly discharge of the month of October of each hydrological year (assuming that it starts at the 1st of October) and $Y^*_Δ(i)$ is the corresponding runoff volume.

4.2.2 Processes of extreme quantities

In many problems, our interest is focused not of time averages, but on extreme quantities for a certain time interval, that is the maximum quantities (e.g. for flood studies) or the minimum quantities (e.g. for drought studies). For the study of these quantities we construct appropriate discrete time processes. Thus, Fig. 4.2(5) demonstrates the construction of the process of *instantaneous annual maxima*, $Z_0(i)$. In each year in a realization of the continuous time process *X*(*t*) we have taken only one value, the instantaneous maximum value that occurs during the entire year. We can extend this from the realization to the process and write

$$Z_0(i) := \max_{i-1 \le t < i} \{X(t)\} \tag{4.5}$$

Likewise, we can define the process of *instantaneous annual minima*. Again in these processes the annual periodicity has been fully eliminated and the process autocorrelation has been reduced significantly.

If, instead of instantaneous quantities, we are interested on an average during a time interval *Δ*, then we can construct and study the *process of annual maxima on a specified time scale*, i.e. (Fig. 4.2(6))

$$Z_Δ(i) := \max_{i-1 \le s < i-Δ} \left\{ \frac{1}{Δ} \int_s^{s+Δ} X(t) dt \right\}, \quad Z^*_Δ(i) := \max_{i-1 \le s < i-Δ} \left\{ \int_s^{s+Δ} X(t) dt \right\} \tag{4.6}$$

This definition was based on the continuous time process *X*(*t*). Alternatively – but with smaller precision – it can be based on the already time discretized process $X_Δ(i)$ (Fig. 4.2(7)):

$$Z'_Δ(i) := \max_{j_1 \le j \le j_2} \{X_Δ(j)\}, \quad Z'^*_Δ(i) := \max_{j_1 \le j \le j_2} \{X^*_Δ(j)\} \tag{4.7}$$

where $j_1 := (i-1)D/Δ + 1$, $j_2 := iD/Δ$. Comparing Fig. 4.2(6) and Fig. 4.2(7), it is apparent that $Z_Δ(i)$ και $Z'_Δ(i)$ are not identical in terms of the time position or their magnitude, but they do not differ much. Likewise, we construct the *process of annual minima on a specified time scale*. The typical values of the time interval *Δ* in flood and drought studies vary from a few minutes (e.g. in design storm studies of urban drainage networks) to a few months (in water quality studies of rivers in drought conditions).

A last series of maxima, known as *series above threshold* or *partial duration series* is demonstrated in Fig. 4.2(8), and can serve as a basis of the definition of the related processes. This is usually constructed from the discrete time process $X_Δ(i)$, as in Fig. 4.2(7). The difference here is that instead of taking the maximum over each year, we form the series of all values that exceed a threshold c, irrespectively of the location of these values in hydrological years, i.e.

$$\{W_\Delta(i),\, i=1,2,\ldots\} := \{X_\Delta(j) \,|\, X_\Delta(j) \ge c,\ j=1,2,\ldots\} \tag{4.8}$$

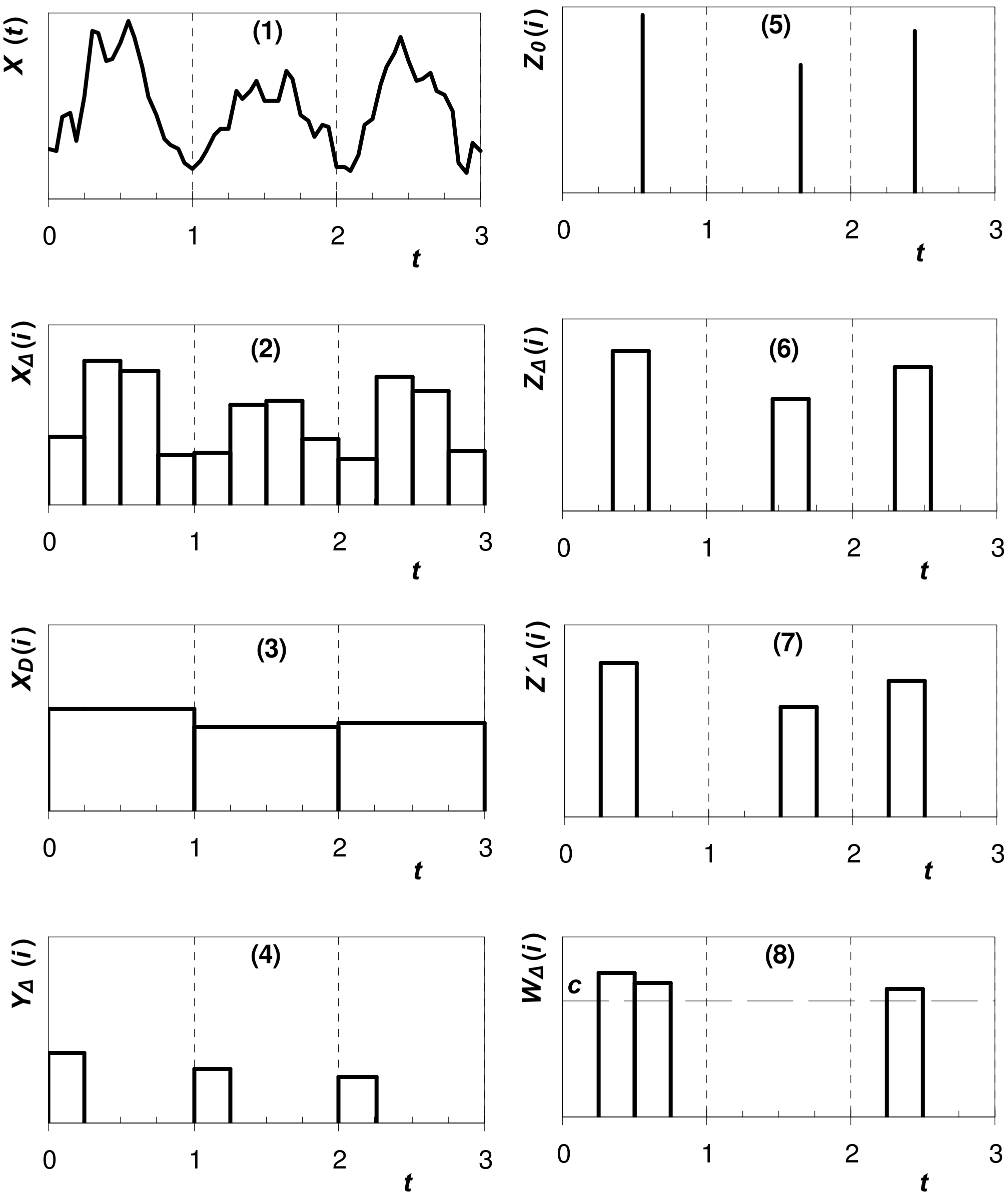

Fig. 4.2 Auxiliary sketch for the definition of the different types of stochastic processes; time t is in years.

Strictly speaking, the index i does not represent time, but it is just a the rank of the different variables $W_\Delta(i)$ in the time ordered series. The threshold c is usually chosen so that each year includes *on the average* one value greater than the threshold. Thus, in Fig. 4.2(8) the threshold, depicted as a horizontal dashed line, has been chosen so that it yields three values over three hydrological years. We observe that two values are located in the first year, none in the second year and one in the third year. With the above definition, it is possible that consecutive elements of the series correspond to adjacent time intervals, as in the two values of the first year in our example. This may introduce significant stochastic dependence in the series. To avoid this, we can introduce a second threshold of a minimum time distance between consecutive elements of the series.

4.2.3 From stochastic processes to random variables

As clarified above, this text does not cover the analysis of the complete geophysical processes either in continuous or discrete time. However, we have defined six other types of processes, in which the "time" index is discrete and may differ from actual time. Each of these processes includes one element per year, except of the process over threshold, which includes a variable number of elements per year, with an average of one per year. For our study, we shall make the following assumptions:

1. The processes are stationary: the distribution of each random variable remains the same from year to year.
2. The processes are ergodic: ensemble averages equal time averages.
3. The variables corresponding to different times are independent.

To clarify the meaning of these assumptions, we will discuss an example. Let $X(t)$ represent the instantaneous discharge of a river at time t, and (according to the above notation) $X_D(\tau)$ represent the mean annual discharge of (hydrological) year τ. Let us assume that 30 years of observations are available, so that we know the values $x_D(1)$, …, $x_D(30)$, which we regard as realizations of the random variables $X_D(1)$, …, $X_D(30)$. Obviously, for each of the variables $X_D(i)$ we can have (and we have) only one realization $x_D(i)$. In contrast to laboratory conditions, in nature we cannot repeat multiple experiments with different outcomes to acquire a sample for the same variable $X_D(i)$. Given the above observations, we can calculate time averages of certain quantities, for instance the standard sample average $\bar{x}_D$ = $[x_D(1) + … + x_D(30)]/30$. Does this quantity give any information for the variable $X_D(31)$? In general, the answer is negative. However, if assumptions 1 and 2 are valid, then $\bar{x}_D$ gives important information for $X_D(31)$ and, thus, it helps make a statistical prediction. Specifically, under the stationarity assumption all $X_D(i)$ have the same statistical properties, and this provides grounds to treat them collectively; otherwise a quantity such as $\bar{x}_D$ would not have any mathematical or physical meaning at all. Simultaneously, the stationarity assumption allows us to transfer any statistical property concerning the variables $X_D(1)$, …, $X_D(30)$ to the variable $X_D(31)$. The ergodicity assumption makes it possible to transform the time average $\bar{x}_D$ to an estimate of the unknown true average of each of the variables $X_D(i)$, i.e. to estimate $m = E[X_D(i)]$ as $\bar{x}_D$. So, both stationarity and ergodicity assumptions are fundamental and powerful and allow us to make predictions of future events, e.g. $E[X_D(31)] = m$. The third assumption, the independence, is not a fundamental one. It is just a simplification that makes possible the use of classical statistics. Otherwise, if the variables are dependent, the classical statistics need adaptations before they can be used (Koutsoyiannis, 2003).

It is important to stress that the stationarity assumption is not some property of the natural system under study, e.g. the river and its flow. It is an assumption about the mathematical model, i.e. the stochastic process, that we build for the natural system (Koutsoyiannis, 2006a). In this respect, it is related to the behaviour of the system but also to our knowledge of the system. For instance, let us assume that we have a reliable deterministic model of the evolution of the river discharge that predicts that at year 31 the average discharge will be $x_D(31) = 2m$. If we build an additional stochastic model for our system, in an attempt to describe the uncertainty of this deterministic prediction, it will be reasonable to assume that $E[X_D(31)] = 2m$, rather than $E[X_D(31)] = m$. Thus, our process will be not stationary. Even without this hypothetical deterministic model, we can arrive to a similar situation if there is stochastic dependence among the consecutive variables. In case of dependence, the past observations affect our future predictions. In this situation, it is natural to use the conditional mean $E[X_D(31)|x_D(1), \ldots, x_D(30)]$ instead of the unconditional mean $E[X_D(31)] = m$ as a prediction of the next year; the two quantities are different. Likewise, for year 32, given past information, the quantity $E[X_D(32)|x_D(1), \ldots, x_D(30)]$ will be different both from $E[X_D(31)|x_D(1), \ldots, x_D(30)]$ and m. In other words, dependence along with past observation make our model to behave as a nonstationary one in terms of conditional expectations, even though in an unconditional setting it is a stationary process (see also Koutsoyiannis et al., 2007).

Under the assumptions of stationarity, ergodicity and independence, we can replace the notion of a stochastic process $X_D(t)$ with a unique realization $x_D(t)$, with the notion of a single underlying random variable X_D with an ensemble of realizations $x_D(i)$, that are regarded as an observed sample of X_D. In the latter case the time ordering of $x_D(i)$ does not matter at all.

4.2.4 A numerical investigation of the limits of the independence assumption

We assume that, based on observational data of river discharge, we have concluded that the probability of the event of annual runoff volume smaller than 500 hm^3* is very small, equal to 10^{-2}. What is the probability that this event occurs for five consecutive years?

Assuming stationarity, ergodicity and independence, this probability is simply $(10^{-2})^5 = 10^{-10}$. This is an extremely low probability: it means that we have to wait *on the average* 10^{10} or 10 billion years to see this event happen (by the way, the age of earth is much smaller than this duration). However, such events (successive occurrences of extreme events for multiyear periods) have been observed in several historical samples (see section 4.5.3). This indicates that the independence assumption is not a justified assumption and yields erroneous results. Thus we should avoid such an assumption if our target is to estimate probabilities for multiyear periods. Methodologies admitting dependence, i.e. based on the theory of stochastic

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

processes, are more appropriate for such problems and will result in probabilities much greater than 10^{-10}; these however are out of the scope of this text.

Now let us assume that for four successive years our extreme event has already occurred, i.e. that the runoff volume was smaller than 500 hm^3 in all four years. What is the probability that this event will also occur in the fifth year?

Many people, based on an unrefined intuition, may answer that the occurrence of the event already for four years will decrease the probability of another consecutive occurrence, and would be inclined to give an answer in between 10^{-2} and 10^{-10}. This is totally incorrect. If we assume independence, then the correct answer is precisely 10^{-2}; the past does not influence the future. If we assume positive dependence, which is a more correct assumption for natural processes, then the desired probability becomes higher than 10^{-2}; it becomes more likely that a dry year will be followed by another dry year.

4.3 The concept of the return period

For a specific event A, which is a subset of some certain event Ω, we define the *return period*, T, as the mean time between consecutive occurrences of the event A. This is a standard term in engineering applications (in engineering hydrology in particular) but needs some clarification to avoid common misuses and frequent confusion. Under stationarity, if p is the probability of the event, then the return period T is related to p by

$$\frac{T}{\Delta} = \frac{1}{p} \tag{4.9}$$

where Δ is the time interval on which the certain event Ω is defined or, for events defined on varying time frame, the mean interarrival time of the event Ω. For instance in panel (2) of Fig. 4.2, $\Delta = D/4$, whereas in panels (3)-(8) $\Delta = D$, as by construction all these cases involve one event Ω per year (either one exactly or one on the average). In particular, in panel (8), as discussed above, we have chosen the threshold c that defines our event Ω so that each year includes *on the average* one event. Had we chosen a smaller threshold, so that each year include two events on the average, the mean interarrival time of Ω would be $\Delta = D/2$.

Apart from stationarity, no other conditions are needed for (4.9) to hold. To show this, we give the following general proof that is based on the simple identity $P(CA) = P(C) - P(CB)$, valid for any events A and C, with B denoting the opposite event of A. We assume a stationary process in discrete time with time interval Δ. At time i, we denote as A_i the occurrence of the event A and as B_i the non occurrence. Because of stationarity $P(A_1) = P(A_2) = \ldots = P(A) = p$ (also $P(B_1) = P(B_2) = \ldots = P(B) = 1 - p$). The time between consecutive occurrences of the event A is a random variable, say N, whose mean is the return period, T. Assuming that the event A has happened at time 0, if its next occurrence is at time n, we can be easily see that

$$P\{N = n\} = P(B_1, B_2, \ldots B_{n-1}, A_n|A_0) = P(A_0, B_1, B_2, \ldots B_{n-1}, A_n) \,/\, P(A_0) \tag{4.10}$$

or

$$P\{N = n\} = (1/p)\, P(A_0, B_1, B_2, \ldots B_{n-1}, A_n) \tag{4.11}$$

Obviously, $T = E[N]\, \mathit{\Delta}$, where the expected value of N is (by definition)

$$E[N] = 1\, P(N = 1) + 2\, P(N = 2) + \ldots \tag{4.12}$$

Combining the last two equations we obtain

$$p\, E[N] = 1\, P(A_0, A_1) + 2\, P(A_0, B_1, A_2) + 3\, P(A_0, B_1, B_2, A_3) + \ldots \tag{4.13}$$

and using the above mentioned identity,

$$\begin{aligned} p\, E[N] = {} & 1\, [P(A_0) - P(A_0, B_1)] + 2\, [P(A_0, B_1) - P(A_0, B_1, B_2)] + \\ & + 3\, [P(A_0, B_1, B_2) - P(A_0, B_1, B_2, B_3)] + \ldots \end{aligned} \tag{4.14}$$

or

$$p\, E[N] = P(A_0) + P(A_0, B_1) + P(A_0, B_1, B_2) + \ldots \tag{4.15}$$

Using once more the same identity, we find

$$p\, E[N] = [1 - P(B_0)] + [P(B_1) - P(B_0, B_1)] + [P(B_1, B_2) - P(B_0, B_1, B_2)] + \ldots \tag{4.16}$$

and observing that, because of stationarity, $P(B_0) = P(B_1)$, $P(B_0, B_1) = P(B_1, B_2)$, etc., we conclude that

$$p\, E[N] = 1 \tag{4.17}$$

which proves (4.9). From this general proof we conclude that (4.9) holds true either if the process is time independent or dependent, whatever the dependence is. (In most hydrological and engineering texts, e.g. Kottegoda, 1980, p. 213; Kottegoda and Rosso, 1998, p. 190; Koutsoyiannis, 1998, p. 96, independence has been put as a necessary condition for (4.9) to be valid). All this analysis is valid for processes in discrete time; as the time interval $\mathit{\Delta}$, on which the event A is defined, tends to zero, the return period will tend to zero too, provided that the probability of A is finite.

Extreme events that are of interest in geophysics (and hydrology) are usually of two types, highs (floods) or lows (droughts). In the former case the event we are interested is the exceedence of a certain value x, i.e. $\{X > x\}$, which is characterized by the probability of exceedence, $p = F_X^*(x) = P\{X > x\} = 1 - F_X(x)$. In the latter case the event we are interested is the non exceedence of a certain value x, i.e. $\{X \le x\}$, which is characterized by the probability of non exceedence, $p = F_X(x) = P\{X \le x\}$. As the processes that we deal with here are defined on the annual scale ($\mathit{\Delta} = D$), for an exceedence event (high flow) we have

$$\frac{T}{D} = \frac{1}{P\{X > x\}} = \frac{1}{F_X^*(x)} = \frac{1}{1 - F_X(x)} \tag{4.18}$$

whereas for a non exceedence event (drought) we have

$$\frac{T}{D} = \frac{1}{P\{X \le x\}} = \frac{1}{F_X(x)} \tag{4.19}$$

Sometimes we write the above relationships omitting D = 1 year, as it is very common to express the return period in years (essentially identifying T with $E[N]$). However, the correct (dimensionally consistent) forms are those written in equations (4.18)-(4.19). Sometimes (4.18) has been used as a definition of the return period, saying that the return period is the reciprocal of the exceedence probability. This again is not dimensionally consistent (given that return period should have dimensions of time) nor general enough (it does not cover the case of low flows).

The term *return period* should not trap us to imply that there is any periodic behaviour in consecutive occurrences of events such as in exceedence or nonexceedences of threshold values in nature. In a stochastic process the time between consecutive occurrences of the event is a random variable, N, whose mean is the return period, T. For example, if the value 500 m^3/s of the annual maximum discharge in a river has a return period of 50 years, this does not mean that this value would be exceeded periodically once every 50 years. Rather it means that the *average* time between consecutive exceedences will be 50 years. An alternative term that has been used to avoid “period” is *recurrence interval*. However, sometimes (e.g. in Chow *et al.*, 1988) this term has been given the meaning of the random variable N and not its mean T. Typical values used in engineering design of flood protection works are given in Table 4.1.

Table 4.1 Return periods most commonly used in hydrological design for high flows and corresponding exceedence and nonexceedence probabilities.

Return period T (years)	Exceedence probability F^* (%)	Nonexceedence probability F (%)
2	50	50
5	20	80
10	10	90
20	5	95
50	2	98
100	1	99
500	0.2	99.8
1000	0.1	99.9
5 000	0.02	99.98
10 000	0.001	99.99

Note: To adapt the table for low flow events we must interchange the columns *exceedence probability* and *nonexceedence probability*.

4.4 The concept of risk

Depending on the context, risk can be defined to be either a probability of failure or the product of probability of failure times the losses per failure. Here we use the former definition. A failure is an event that usually occurs when the load L, exceeds the capacity C of a construction. In the design phase, the design capacity is larger than the design load, so as to assure a certain *safety margin*

$$\mathrm{SM} := C - L > 0 \tag{4.20}$$

or a certain *safety factor*

$$\mathrm{SF} := \frac{C}{L} > 1 \tag{4.21}$$

In engineering hydrology, L could be, for instance, the design flood discharge of a dam spillway whereas C is the discharge capacity of the spillway, that is the discharge that can be routed through the spillway without overtopping of the dam.

In most empirical engineering methodologies both L and C are treated in a deterministic manner regarding them as fixed quantities. However, engineers are aware of the intrinsic natural uncertainty and therefore are not satisfied with a safety factor as low as, say, 1.01, even though in a deterministic approach this would suffice to avoid a failure. Rather, they may adopt a safety factor as high as 2, 3 or more, depending on empirical criteria about the specific type of structure. However, the empirical deterministic approach is more or less arbitrary, subjective and inconsistent. The probability theory can quantify the uncertainty and the risk and provide more design criteria. According to a probabilistic approach SM and SF are regarded as random variables and the risk is defined to be:

$$R := P\{\mathrm{SF} < 1\} = P\{\mathrm{SM} < 0\} \tag{4.22}$$

and its complement, $1 - R$ is known as *reliability*.

In the most typical problems of engineering hydrology, the design capacity (e.g. discharge capacity or storage capacity) could be specified with certainty ($C = c$), and is not regarded as a random variable. However, the load L should be taken as a random variable because of the intrinsic uncertainty of natural processes. In this case, the risk is

$$R = P\{L > c\} = 1 - P\{L \le c\} \tag{4.23}$$

The probability $P\{L \le c\}$ (the reliability) depends on the variability of the natural process (e.g. the river discharge, the fixed quantity c, and the life time of the project n D (n years). With the notation of section 4.2.2, assuming an appropriate time window $\varDelta$ for the phenomenon studied, the event $\{L \le c\}$ (which refers to the n year period) is equivalent to the event $\{Z_\varDelta(1) \le c, \ldots, Z_\varDelta(n) \le c\}$. Assuming independence of $Z_\varDelta(i)$ trough years, it is concluded that

$$R = 1 - [P\{Z_\Delta \le c\}]^n = 1 - [F_{Z_\Delta}(c)]^n \tag{4.24}$$

where $F_{Z_\Delta}(\ )$ is the distribution function of the annual flood. Expressing it in terms of the return period T from (4.18), we obtain the following relationship that relates the three basic quantities of engineering design, the risk R, the return period T and the life time n years:

$$R = 1 - \left(1 - \frac{D}{T}\right)^n \tag{4.25}$$

Graphical depiction of (4.25) is given in Fig. 4.3 for characteristic return periods. Given that $\ln(1-x)^n = n\ln(1-x) = n(-x - x^2/2 - \cdots) \approx -nx$, the following approximation of (4.25), with error < 1% for $T \ge 50$, is obtained:

$$R \approx 1 - e^{-nD/T} \tag{4.26}$$

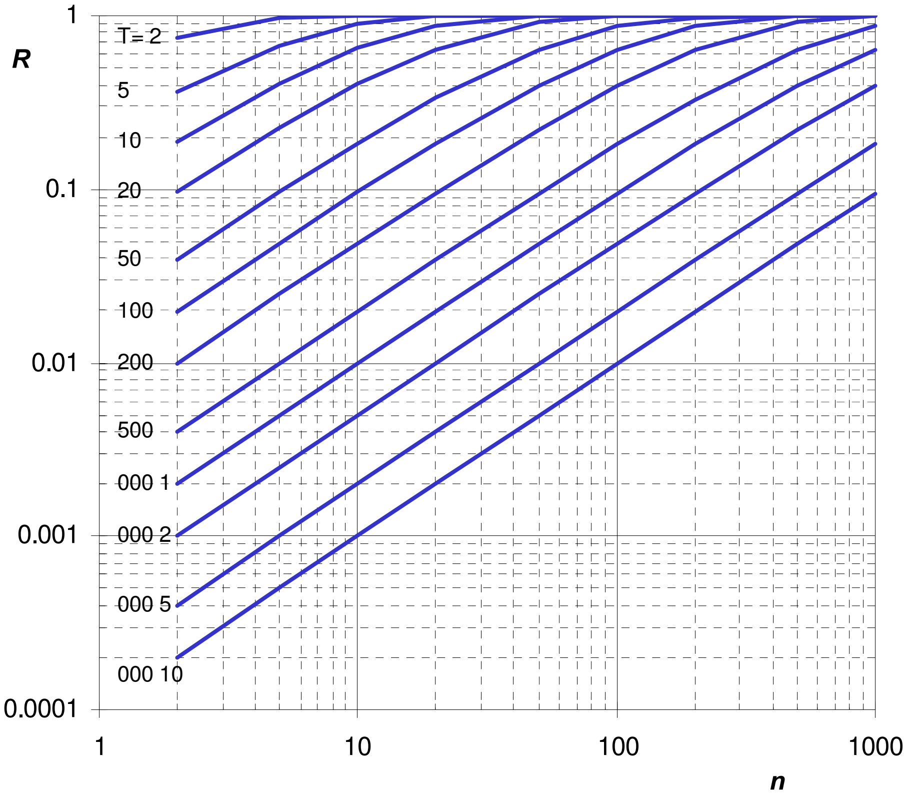

Fig. 4.3 Graphical depiction of the interrelationship of the characteristic quantities of engineering design (equation 4.25).

Solving (4.25) for T we obtain the following relationship that gives the required return period for given design risk and design life time:

$$T = \frac{D}{1 - (1 - R)^{1/n}} \tag{4.27}$$

All equations (4.24)-(4.27) are based on the assumption of independence and are not valid in case of dependence. To get an idea of the effect of dependence, let us examine the limiting case of complete dependence, in which the occurrence of a single event $Z_\Delta(1) \leq c$ entails that all $Z_\Delta(i) \leq c$. It is easy to see that in this case we should substitute 1 for n in all equations. Thus, (4.27) becomes $T = D/R$ so that it will yield a return period smaller than that estimated by (4.27) if the risk is specified. In other words, if we use (4.27) we are on the safe side in the case that there is dependence in the process.

4.4.1 Numerical examples

a. We assume that a diversion tunnel is planned to operate during the construction period of a dam, which has been estimated to be 5 years. What is the return period so that the risk be lower than an acceptable 10%?

From (4.27) we obtain

$$T = \frac{1}{1-(1-R)^{1/n}} = \frac{1}{1-(1-0.1)^{1/5}} = 47.9 \text{ years}$$

We round off the return period to 50 years.

b. What is the risk in a project, for which the return period was assumed equal to its design life time?

If the life time of the project is long enough (≥ 50 years), then from (4.26) we obtain $R = 1 - \mathrm{e}^{-1} = 0.632 = 63.2\%$. Otherwise from (4.25), we obtain

$$R = 1 - \left(1 - \frac{1}{n}\right)^n$$

which for values n = 5, 10 and 20 years results in R = 67.2%, 65.1% and 64.2%, respectively.

4.5 An introduction to dependence, persistence and antipersistence

4.5.1 Definitions and basic tools

Common random series like those observed for example in games of chance (dice, roulette, etc.) are obtained by repetitive experiments, each of which is independent of all other. In contrast, geophysical time series are not outcomes of separate experiments. The entire length of a geophysical time series could be thought of as equivalent to a single never ending experiment. It is like observing the whole trajectory of a die throughout its entire movement, assumed to be endless, rather than waiting to observe the outcome when the die goes to rest. While independence is well justified in a series of outcomes of separate experiments, it is totally unjustified when we are interested in the continuous trajectory of the die. Obviously, the state (position, momentum) of the die at time $t + \Delta t$ depends on the state at time t. The two states tend to be identical as Δt tends to zero.

Likewise, in all physical systems that evolve in continuous time, the autocorrelation coefficient for lag tending to zero tends to 1 (complete dependence). As lag increases, the autocorrelation decreases, generally tending to zero for lag tending to infinity. The positive autocorrelation is also termed persistence, as already discussed in section 4.1. The persistence is characterized as short-term persistence when the autocorrelation tends to zero as an exponential function of lag time and as log-term persistence when the autocorrelation tends to zero as a power-law function of lag time. The latter case corresponds to stronger or longer tail of the autocorrelation function. Sometimes, for intermediate lags negative autocorrelations may appear. The general behaviour corresponding to this case is known as *antipersistence*.

An easier means to explain antipersistence and persistence, short- or long-term, is provided by studying the variation of the standard deviation with time scale. To avoid the effect of seasonality, here we consider time scales Δ that are integer multiples of the annual time scale, i.e.,

$$\Delta = k\, D, \;\; k = 1, 2, \dots \tag{4.28}$$

By virtue of (4.1), which holds for any Δ (also for $\Delta > D$), we easily obtain that the process at scale Δ is related to that at scale D by

$$X_{kD}(i) = [X_D(ik - k + 1) + \dots + X_D(ik)]/k \tag{4.29}$$

This is nothing other than the time average of a number k of consecutive random variables. We can define similar time average processes for over-annual scales also for the other cases (Y and Z) that we discussed in section 4.2. For k sufficiently large (typically 30, even though sometimes $k = 10$ has been also used), such processes represent what we call *climate*; $\Delta = 30$ years is the typical climatic time scale. However, here we will regard Δ as varying and we will study the variation of the standard deviation of $X_\Delta(i)$ with $\Delta = kD$.

Let $\sigma_\Delta \equiv \sigma_{kD}$ denote the standard deviation at scale $\Delta = kD$, i.e. $\sigma_{kD} := \mathrm{StD}[X_{kD}(i)]$. According to (4.29) $X_{kD}(i)$ is the average of k random variables. If these variables are independent, then we know from chapter 3 that

$$\sigma_{kD} = \frac{\sigma_D}{\sqrt{k}} \text{ or } \sigma_\Delta = \sigma_D \sqrt{\frac{D}{\Delta}} \tag{4.30}$$

where σ_D is the standard deviation at scale 1. This provides a means to test whether or not the process at hand is independent in time. If it is independent, then the double logarithmic plot of σ_Δ *vs.* Δ will be a straight line with slope –0.5. Milder negative slopes (>–0.5) indicate persistence and steeper slopes (<–0.5) indicate antipersistence. Short-term persistence is manifested in the plot as a curve with mild slope for small k, which asymptotically tends to -0.5 for large k. In long-term persistence the slope remains milder than –0.5 even for large k. A more generalized law that asymptotically (for large k) holds in cases of long-term persistence and antipersistence is given by

$$\sigma_{kD} \propto \frac{\sigma_D}{k^{1-H}} \text{ or } \sigma_\Delta \propto \sigma_D \left(\frac{D}{\Delta}\right)^{1-H} \tag{4.31}$$

The coefficient H is termed the Hurst coefficient, after Hurst (1951) who discovered the long-term persistence in geophysical time series*. Clearly, $H = 1 + d$, where d is the slope of the plot of σ_Δ *vs.* Δ. Generally, for stationary processes, $0 < H < 1$ (Mandelbrot and van Ness, 1968). For independent processes, $H = 0.5$; for persistent processes, $0.5 < H < 1$ and for antipersistent processes, $0 < H < 0.5$. For persistent processes it is possible that the law (4.31) holds as an equality for any k. Mathematically, this is also possible for antipersistent processes ($H < 0.5$) but physically it is not realistic. To see the reason why this happens, we assume that the law (4.31) holds as an equality for any k; in this case it defines a stochastic process termed a simple scaling stochastic (SSS) process. It can be shown (e.g. Koutsoyiannis, 2002) that the autocorrelation $X_{kD}(i)$ of the process for scale kD and lag j, i.e. the quantity $\rho_{kD}(j) := \mathrm{Cov}[X_{kD}(i), X_{kD}(i + j)] / \mathrm{Var}[X_i^{(k)}])$, is given by

$$\rho_{kD}(j) = \rho(j) = (1/2)\ (|j + 1|^{2H} + |j - 1|^{2H}) - |j|^{2H} \tag{4.32}$$

This shows that the autocorrelation is independent of scale. Inspection shows that if $H > 0.5$ the autocorrelation for any lag is positive (persistence), whereas if $H < 0.5$ the autocorrelation for any lag is negative (antipersistence). In the latter case it takes the most negative values at lag $j = 1$, which is $\rho_{kD}(1) = \rho(1) = 2^{2H-1} - 1$. However, physical realism demands that for small scales and lags, the autocorrelation should be positive.

Given a time series of sufficient length n at time scale D, we can test in a simple way whether the law (4.30) is fulfilled or not, and if not, we can see whether the time series implies persistence or antipersistence. To this aim, we can estimate from the time series the standard deviation σ_{kD} for several values of k. Assuming $k = 1$, we estimate σ_D from the n data values available. For $k = 2$ (and assuming for simplicity that the series length n is even) we can construct a size (n/2) sample $X_{2D}(1) = [X_D(1) + X_D(2)]/2$, $X_{2D}(2) = [X_D(3) + X_D(4)]/2$, …, $X_{2D}(n/2) = [X_D(n - 1) + X_D(n)]/2$. From these we can estimate σ_{2D}. Proceeding this way (e.g. $X_{3D}(1) = [X_D(1) + X_D(2) + X_D(3)]/3$, etc.) we can estimate σ_{kD} for k up to, say, $n/10$ (in order to have 10 sample values for the estimation of standard deviation). Constructing a logarithmic plot of the estimate of standard deviation σ_{kD} versus k we can test graphically the validity of the statistical law (4.30) and estimate the coefficient H of law (4.31).

4.5.2 Synthetic examples

Now we will demonstrate the above concepts with the help of a few examples. We will start with the synthetic example that was already studied in chapter 1. Although this example is referred to a fully deterministic system, it is useful in understanding the behaviours discussed; besides, the statistical analyses outlined above can be applied also in time series that result from deterministic systems. It is reminded that the working example of chapter 1 examines a hypothetical plain with water stored in the soil, which sustains some vegetation. Each year a

[*] Hurst used a different formulation of this behaviour, based on the so-called rescaled range. The formulation in terms of standard deviation at the time scale kD, as in equation (4.31), is much simpler yet equivalent to Hurst's (see theoretical discussion by Beran, 1994, p. 83, and practical demonstration by Koutsoyiannis, 2002, 2003).

constant amount of water enters the soil and the potential evapotranspiration is also constant, but the actual evapotranspiration varies following the variation of the vegetation cover *f*. The vegetation cover and the soil water storage *s* are the two state variables of the system that vary in time *i*; the system dynamics are expressed by very simple equations.

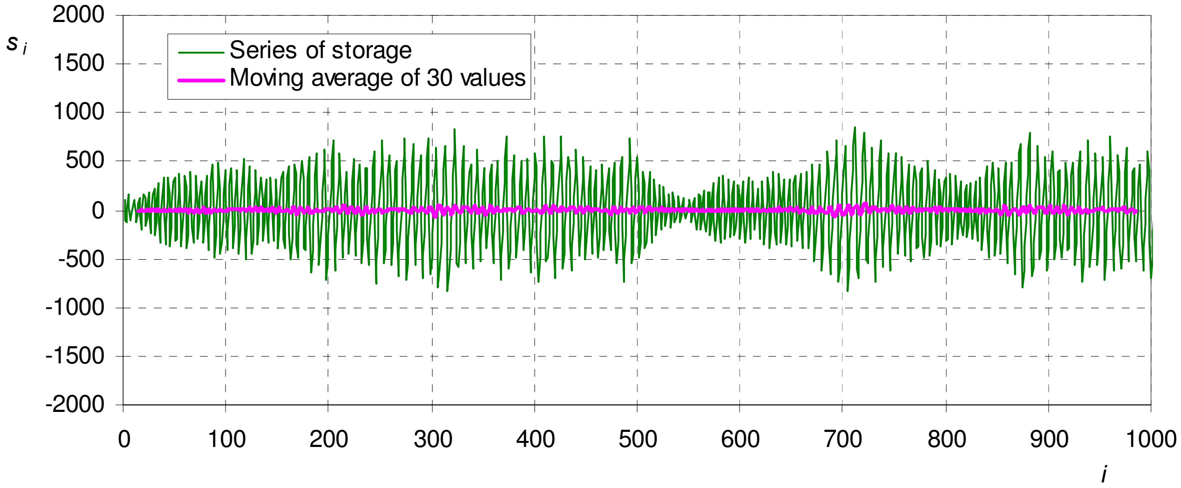

Fig. 4.4 Graphical depiction of the evolution of the system storage s_i (in mm) of the working example in chapter 1 for time up to 1000.

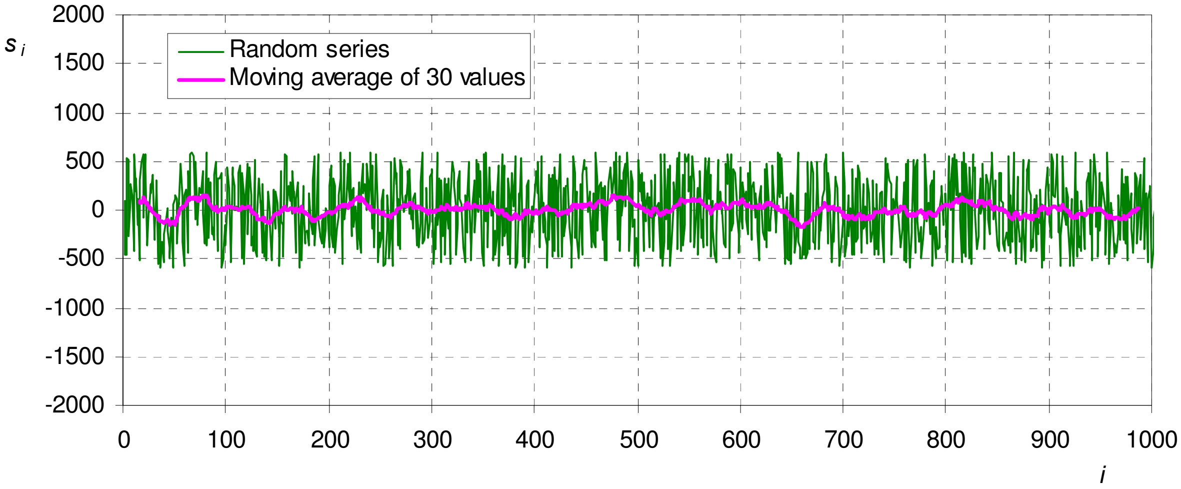

Fig. 4.5 Graphical depiction of a series of random numbers in the interval (-800, 800) having mean and standard deviation equal to those of the series in Fig. 4.4.

In chapter 1, Fig. 1.3, we have seen a graphical depiction of the system evolution for certain initial conditions that we called the “true” conditions. Now in Fig. 4.4 we depict the continuation of this evolution of the storage $s(i)$ (or s_i) for time up to 1000. In addition, we have given in Fig. 4.4 a plot of the 30-year moving average[*] of s_i, which shows that this

[*] The moving average is the average of *k* random variables consecutive in time, as in (4.29). However, for better illustration, here we used a slightly different definition, i.e., $X_{30D}(i) = [X_D(i - 15) + \ldots + X_D(i + 14)]/30$.

moving average is almost a horizontal straight line at $s = 0$. The experienced eye may recognize from this, without the need of further tools, a strongly antipersistent behaviour.

For comparison we have derived a random series* with mean and standard deviation equal to those of the original series, and we plotted it in Fig. 4.5. Comparing the plots of moving averages in Fig. 4.4 and Fig. 4.5 we see a clear difference. In the former case (antipersistence) the plot is virtually a horizontal straight line, whereas in the latter case (pure randomness) it is a curly line, which however does not depart very much from the line $s = 0$.

Sometimes antipersistence has been confused with periodicity or cyclic behaviour. However, periodicity would imply that the time between consecutive peaks in the time series would be constant, equal to the period of the phenomenon. To distinguish the two behaviours, we have calculated a series of times between peaks, τ, from the time series of our example, which for better accuracy we extended up to 10 000 items by the same algorithm. From this series we constructed an histogram of times between peaks, which is shown in Fig. 4.6. We see that the time between peaks varies from 4 to 22 years, with a mode of 6 years. Clearly, this behaviour is totally different from a periodic phenomenon, and is better described by the term antipersistence.

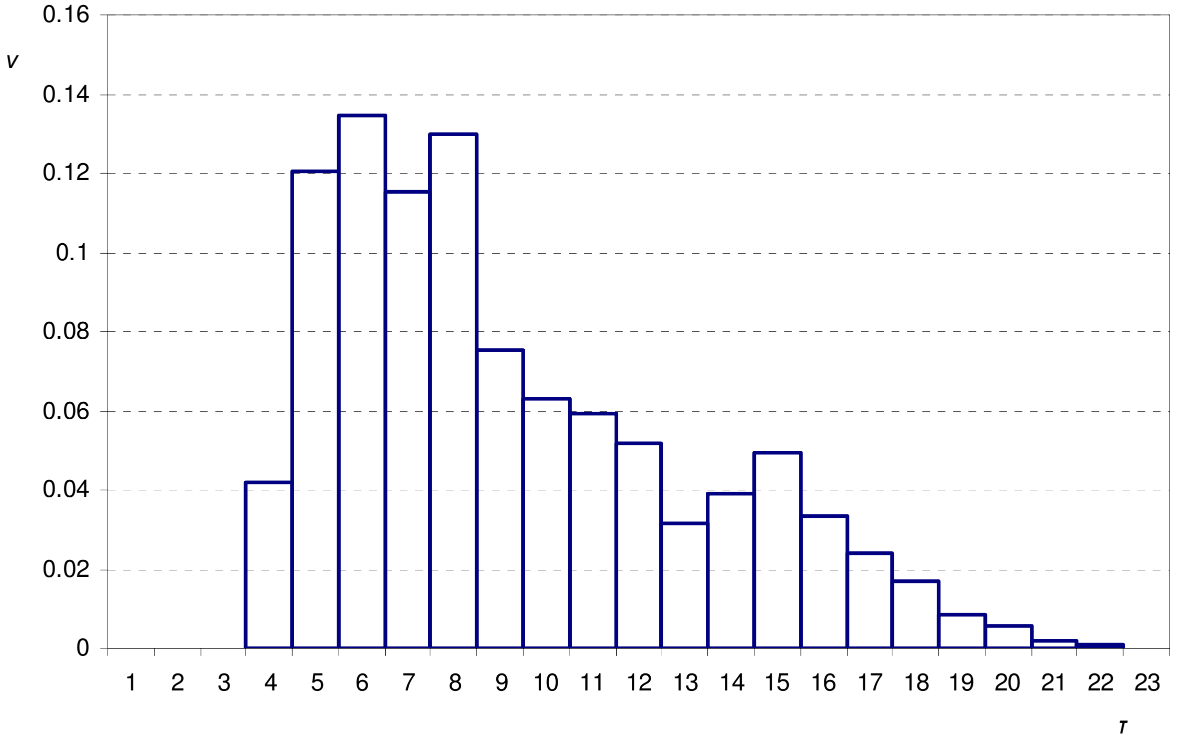

Fig. 4.6 Relative frequency v of the time τ between consecutive peaks, estimated from a series of 10 000 items of a series of s_i of the example system.

* This series has been produced as follows: First, we derive a series of integer random numbers q_i by the recursive relationship $q_i = \kappa\, q_{i-1} \bmod \lambda$, where $\kappa = 7^5$, $\lambda = 2^{31} - 1$, $q_0 = 78910785$ and mod is the modulo operator that finds the remainder of an integer division. Then, we derive a series of real numbers in the interval [0, 1) as $r_i = q_i / \lambda$. We obtain the final series s_i by $s_i = c(2q_i - 1)$, where $c = 600$.

The same example system helps us to acquire an idea of persistence. To this aim we have constructed and plotted in Fig. 4.7 1000 terms of the time series of the peaks p_j of the time series s_i. Now we see in Fig. 4.7 that the moving average of 30 values exhibits large and long excursions from the overall mean, which is about 800 (not plotted). These excursions are the obvious manifestation of persistence.

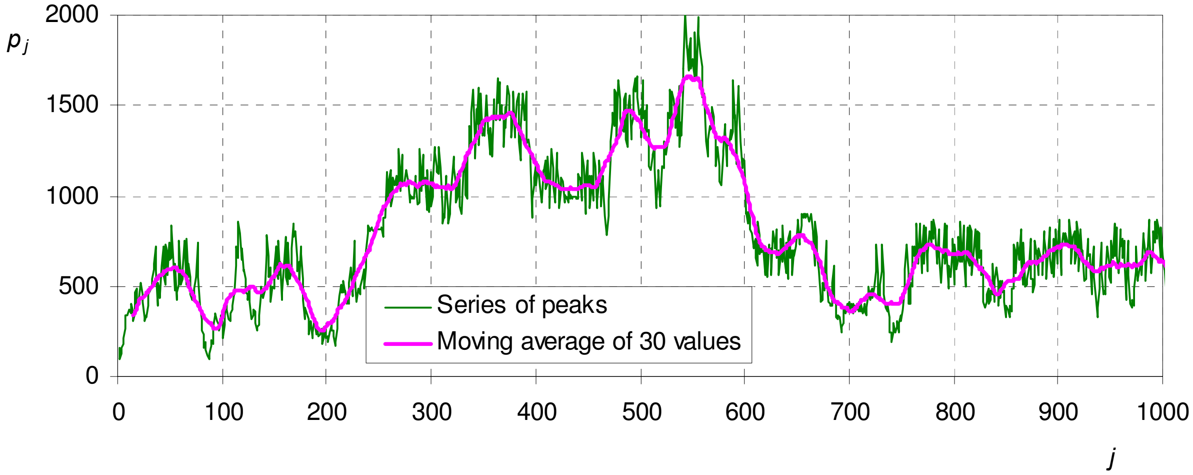

Fig. 4.7 Graphical depiction of a series of peaks p_j (in mm) of the soil storage s_i; here j does not denote time but the rank of each peak in time order.

However, a better depiction and quantification of the persistence and antipersistence is provided by the plot of standard deviation $\sigma_\varDelta$ vs. time scale $\varDelta$, as described in section 4.5.1. Fig. 4.8 gives such plots for the series of storage shown in Fig. 4.4 (but for 10 000 items) and for the random series of Fig. 4.5 (also for 10 000 items). Clearly, the plot of the random series shows a straight line arrangement with slope –0.5, which corresponds to a Hurst coefficient H = 1 – 0.5 = 0.5 (as expected). The plot of the storage time series is more interesting. For low scales ($\varDelta \leq 4$) the slope in the arrangement of the points is very low, indicating a positive dependence at small lags. However, for large scales ($\varDelta \geq 20$), a straight line arrangement of points appears, which has large slope, equal to –0.98. This corresponds to a Hurst coefficient H = 1 – 0.98 = 0.02, which indicates very strong antipersistence.

Likewise, Fig. 4.9 gives a similar plot for the series of peaks shown in Fig. 4.7, also in comparison with that of the random series. The plot of the series of peaks shows a straight line arrangement of points for low and high $\varDelta$, which has very low slope, equal to –0.02. This corresponds to a Hurst coefficient H = 1 – 0.02 = 0.98, which indicates very strong persistence.

We can observe in Fig. 4.8 that for scale $\varDelta$ = 1 the standard deviation of the series of storage is significantly greater than that of the random series, despite the fact that the latter was constructed so as to have the same mean and standard deviation with the former. This is because, after the expansion of the two series from 1000 to 10 000 items, σ_D of the storage series increased significantly whereas σ_D of the random series remained in the same level. The large persistence in the peaks results in high fluctuations of standard deviation and this was the reason for the increased σ_D of the storage time series.

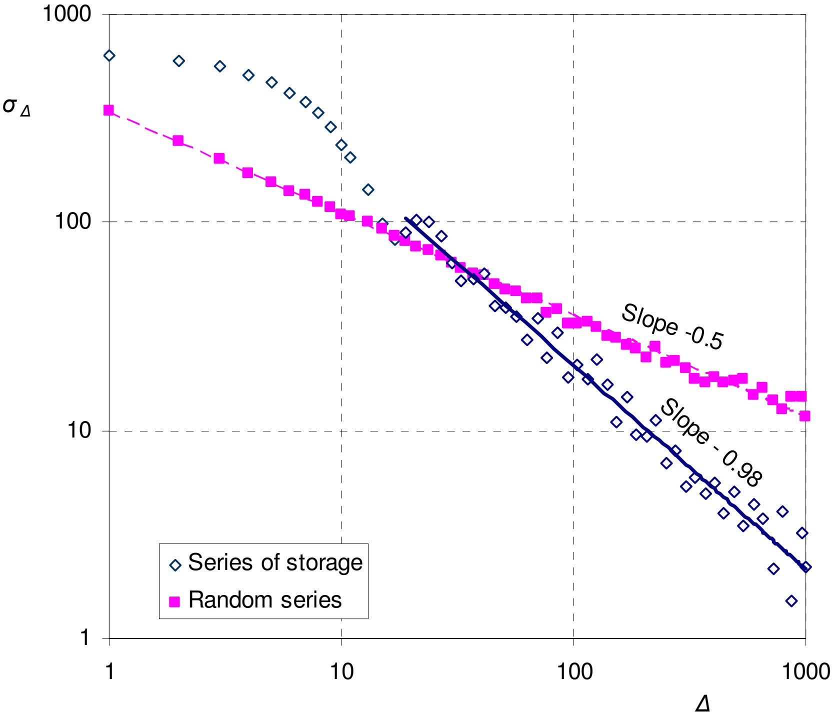

Fig. 4.8 Standard deviation σ_Δ (in mm) vs. time scale Δ plot of the series of storage shown in Fig. 4.4 (but for 10 000 items) in comparison with that of the random series of Fig. 4.5 (also for 10 000 items).

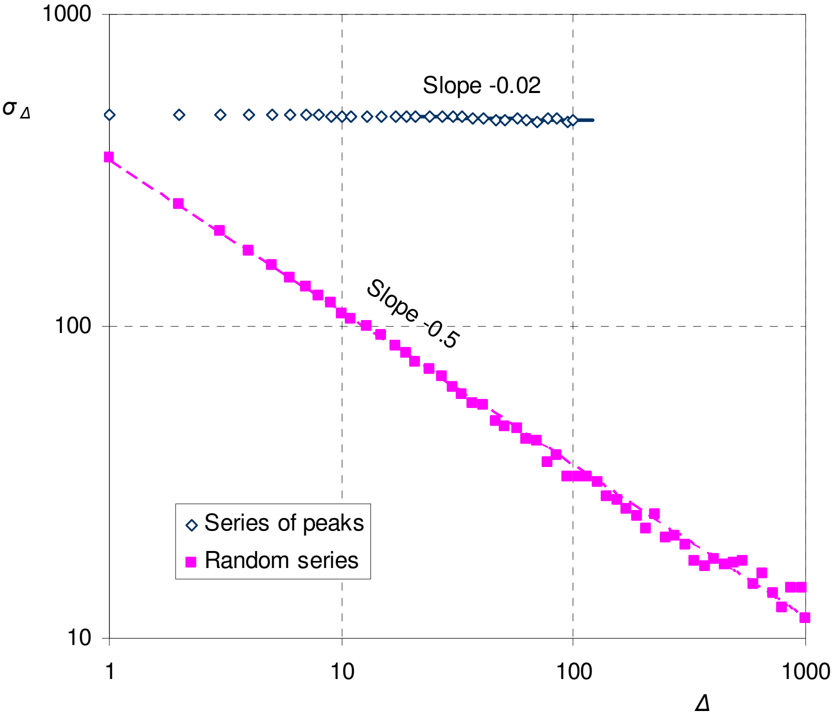

Fig. 4.9 Standard deviation σ_Δ (in mm) vs. scale Δ plot of the series of peaks shown in Fig. 4.7 in comparison with that of the random series of Fig. 4.5 (also shown in Fig. 4.8).

4.5.3 Real world examples

It is not easy to find real world examples with antipersistent behaviour. However, there are a few phenomena with such behaviour which are commonly called “oscillations”. The most widely known is the El Niño Southern Oscillation (ENSO), a fluctuation of air pressure and water temperature between the southeastern and southwestern Pacific. Typically it is quantified by the so-called Southern Oscillation Index (SOI), which expresses the difference in the air pressure between Tahiti (an island in French Polynesia) and Darwin (North Australia); this difference is typically standardized in monthly scale by monthly mean and standard deviation. Here, instead of SOI, we have used the raw time series of the air pressure in Tahiti*, to avoid the artificial effects of taking differences and standardizing, and we have averaged the monthly time series on annual basis to discard the effect of the annual cycle.

The annual series has been plotted in Fig. 4.10, where the antipersistent behaviour becomes apparent from the 30-year moving average, which is virtually an horizontal straight line. The same behaviour is also apparent in Fig. 4.11, which shows the plot of standard deviation σ_{Δ} vs. time scale Δ. For large scales ($\Delta \geq 2$ years), a straight line arrangement of points appears, which has high slope, equal to –0.8. This corresponds to a Hurst coefficient $H = 1 - 0.8 = 0.2$, which indicates strong antipersistence. The figure also shows a series of points that were derived from the monthly time series. For large scales, the monthly plot is virtually the same with the annual plot. For low time scales, the monthly plot clearly shows a low slope, which manifests the combined effect of the annual cycle and positive autocorrelation for small lags at the monthly scale (even for the annual scale, the lag one autocorrelation is positive, 0.18). Generally, the figure resembles Fig. 4.8.

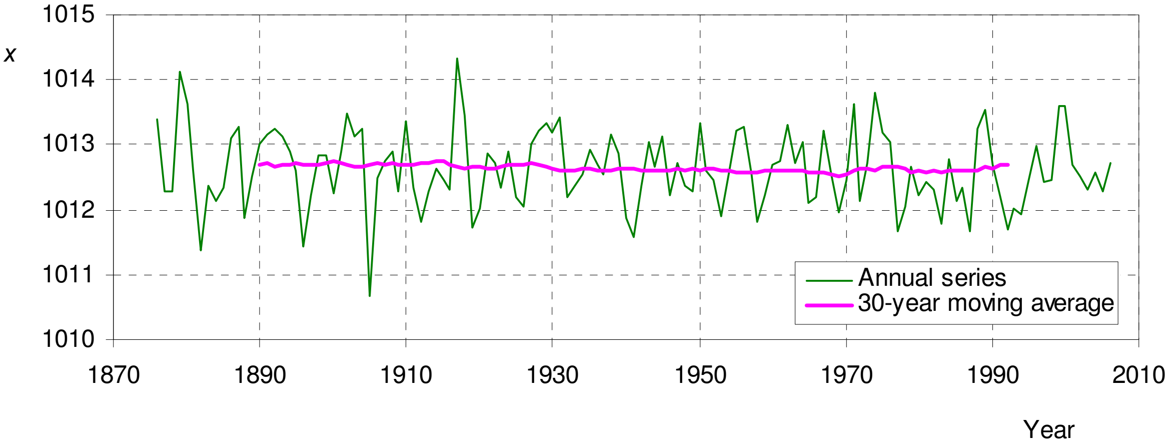

Fig. 4.10 Graphical depiction of the mean annual air pressure in Tahiti (in hPa), which is one of the two variables used to define the Southern Oscillation Index (SOI).

* The series is available online at ftp://ftp.bom.gov.au/anon/home/ncc/www/sco/soi/tahitimslp.html on a monthly scale.

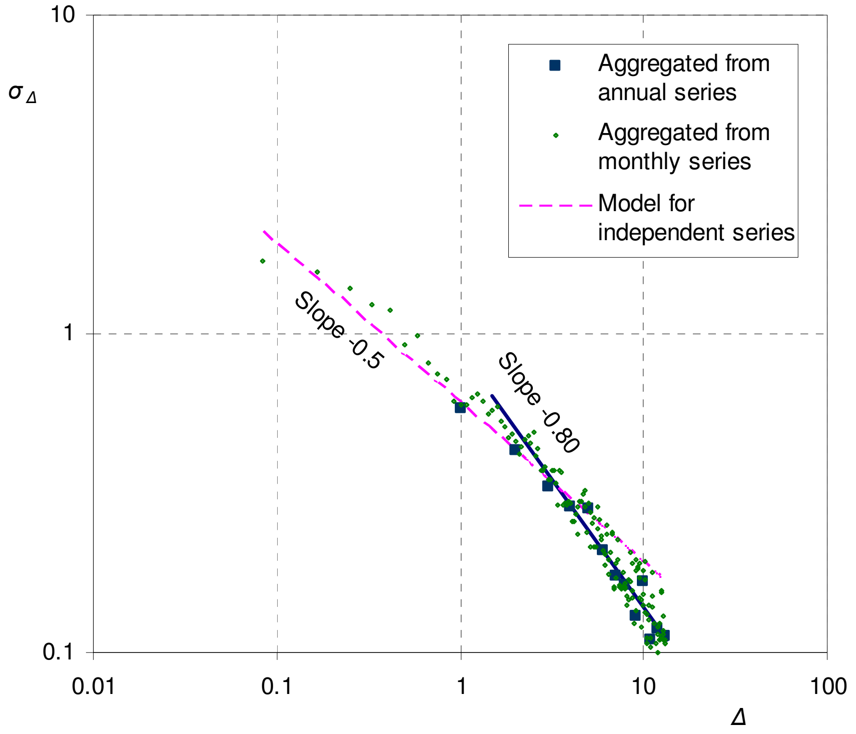

Fig. 4.11 Plot of standard deviation σ_Δ (in hPa) vs. scale Δ (in years) for the series of air pressure in Tahiti shown in Fig. 4.10.

While antipersistence is very rarely seen in nature, persistence is a very common behaviour, which however requires long time series to be observed. Long-term persistence has been found to be omnipresent in several long time series such as meteorological and climatological (temperature at a point, regional or global basis, wind power, proxy series such as tree ring widths or isotope concentrations) and hydrological (particularly river flows), but it has been also reported in diverse scientific fields such as biology, physiology, psychology, economics, politics and Internet computing (Koutsoyiannis and Montanari, 2007). Thus, it seems that in real world processes this behaviour is the rule rather than the exception. The omnipresence can be explained based either on dynamical systems with changing parameters (Koutsoyiannis, 2006b) or on the principle of maximum entropy applied to stochastic processes at all time scales simultaneously (Koutsoyiannis, 2005).

The example we study here is the most common one and refers to the longest available instrumental data series. This is the annual minimum water level of the Nile river for the years 622 to 1284 AD (663 observations), measured at the Roda Nilometer near Cairo (Beran, 1994)*. The time series is plotted in Fig. 4.12, where the long excursions of the 30-year moving average from the overall mean are apparent. As discussed above, the large fluctuations at large scales distinguishes the time series from random noise and is the signature of long-term persistence.

* The data are available from http://lib.stat.cmu.edu/S/beran.

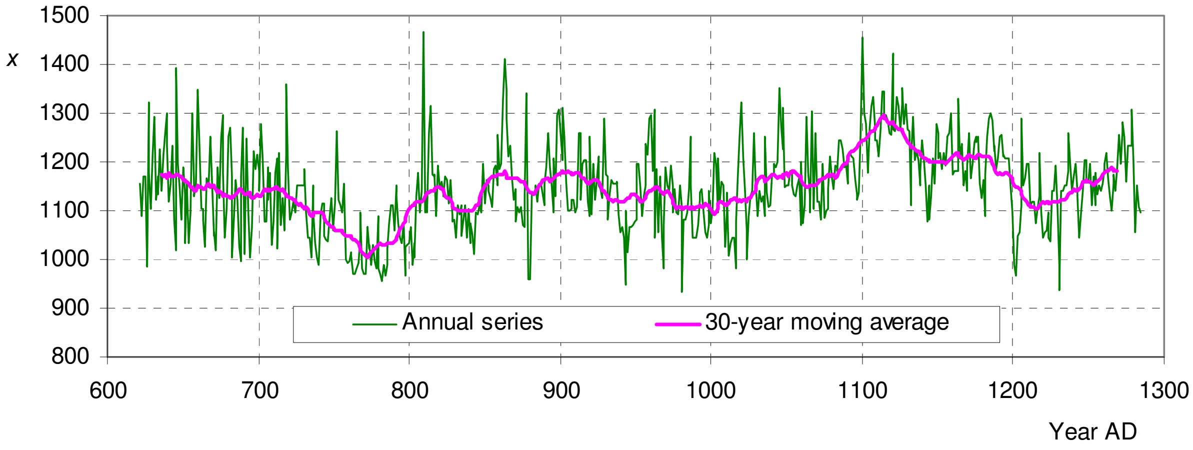

Fig. 4.12 Graphical depiction of the time series of the minimum annual water level at the Roda Nilometer (in cm) for the years 622 to 1284 AD (663 years).

The persistence is also apparent in Fig. 4.13, which shows the plot of standard deviation σ_Δ vs. time scale Δ. Here, the straight line arrangement of points appears on all scales, which makes the law (4.31) valid virtually on all scales. The slope equals –0.14 and it corresponds to a Hurst coefficient $H = 1 - 0.14 = 0.86$, which indicates strong persistence.

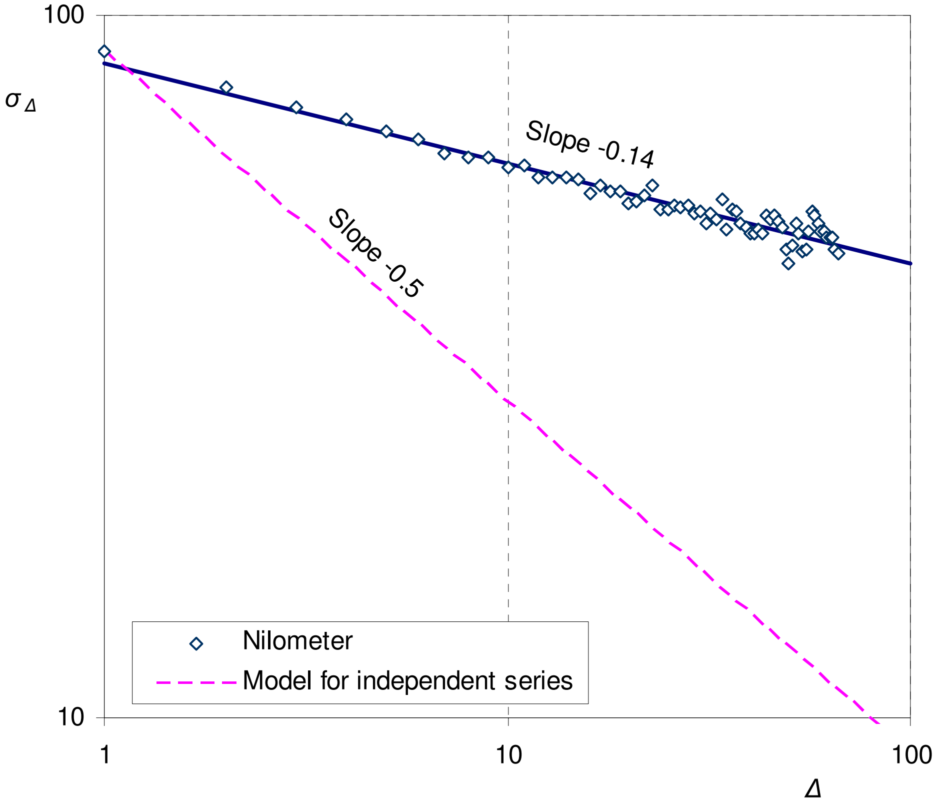

Fig. 4.13 Plot of standard deviation σ_Δ (in cm) vs. scale Δ (in years) for the Nilometer minimum annual water lever time series shown in Fig. 4.12.

References

Beran, J., *Statistics for Long-Memory Processes*, Monographs on Statistics and Applied Probability, vol. 61. Chapman & Hall, New York1994..

Chow, V. T., D. R. Maidment, and L. W. Mays, *Applied Hydrology*, McGraw-Hill, 1988.

Haan, C. T., *Statistical Methods in Hydrology*, The Iowa State University Press, USA, 1977.

Hurst, H.E., Long term storage capacities of reservoirs. *Transactions ASCE*, 116, 776–808, 1951.

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

Kottegoda, N. T., and R. Rosso, *Statistics, Probability and Reliability for Civil and Environmental Engineers*, McGraw-Hill, New York, 1998.

Koutsoyiannis, D., *Statistical Hydrology*, Edition 4, 312 pages, National Technical University of Athens, Athens, 1997 (in Greek).

Koutsoyiannis, D., The Hurst phenomenon and fractional Gaussian noise made easy, *Hydrological Sciences Journal*, 47(4), 573-595, 2002.

Koutsoyiannis, D., Climate change, the Hurst phenomenon, and hydrological statistics, *Hydrological Sciences Journal*, 48(1), 3-24, 2003.

Koutsoyiannis, D., Uncertainty, entropy, scaling and hydrological stochastics, 2, Time dependence of hydrological processes and time scaling, *Hydrological Sciences Journal*, 50(3), 405-426, 2005.

Koutsoyiannis, D., Nonstationarity versus scaling in hydrology, *Journal of Hydrology*, 324, 239-254, 2006a.

Koutsoyiannis, D., A toy model of climatic variability with scaling behaviour, *Journal of Hydrology*, 322, 25-48, 2006b.

Koutsoyiannis, D., A. Efstratiadis, and K. Georgakakos, Uncertainty assessment of future hydroclimatic predictions: A comparison of probabilistic and scenario-based approaches, *Journal of Hydrometeorology*, 8(3), 261-281, 2007.

Koutsoyiannis, D., and A. Montanari, Statistical analysis of hydroclimatic time series: Uncertainty and insights, *Water Resources Research*, 43(5), W05429.1-9, 2007.

Mandelbrot, B. B., *The Fractal Geometry of Nature*, 248 pages, Freeman, New York, USA, 1977.

Mandelbrot, B.B., and J.W. van Ness, Fractional Brownian Motion, fractional noises and applications, *SIAM Review*, 10, 422 – 437, 1968.

Chapter 5

Typical univariate statistical analysis in geophysical processes

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

Summary

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

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

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

5.1 Summary statistics

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

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

Table 5.1 Typical summary statistics and formulae for their calculation

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

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

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

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

5.1.1 Demonstration of summary statistics via a numerical example

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

a. Sample moments and derivative characteristics

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

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

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

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

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

Table 5.3 Traditional calculations of sample moments.

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

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

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

The sample variance is

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

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

the sample standard deviation

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

and the sample coefficient of variation

$$\hat{C}_{\mathrm{v}_X} = s_X / \bar{x} = 211.5 / 725.0 = 0.29$$

The third central moment is

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

and the coefficient of skewness

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

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

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

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

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

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

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

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

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

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

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

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

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

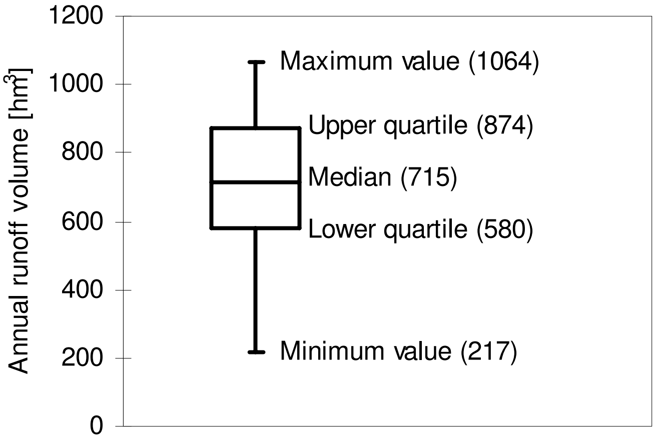

Fig. 5.1 Box plot of the annual runoff of the Evinos river basin at Poros Reganiou.

The results are summarized in Table 5.4.

b. Summary characteristics of the sorted sample

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

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

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

c. Box plot

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

5.2 Histograms

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

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

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

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

5.2.1 Demonstration of histogram

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

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

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

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

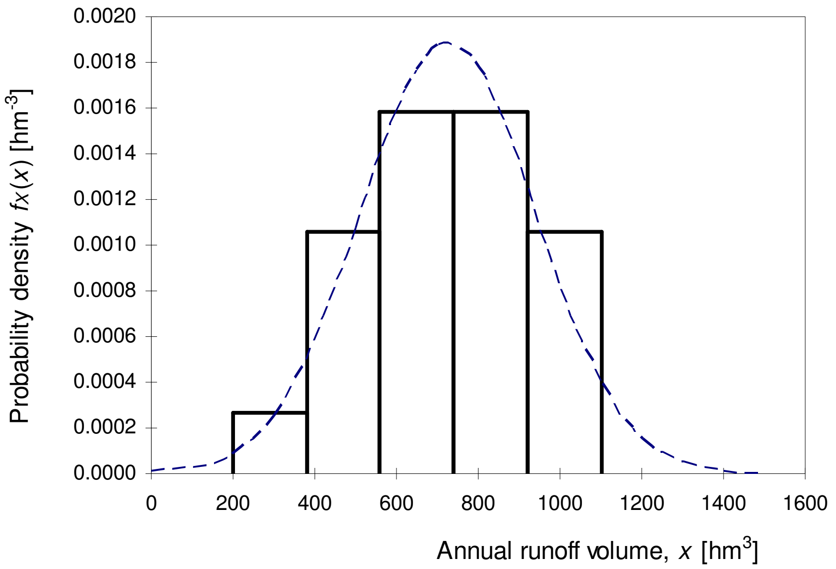

Fig. **5.**2 Histogram of the sample of Table 5.2. For comparison the probability density function of the normal distribution $N(725, 211.5)$ is plotted (dotted line).

5.3 Empirical distribution function

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

5.3.1 Order statistics

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

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

5.3.2 Classical empirical distribution function

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

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

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

5.3.3 Plotting position

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

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

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

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

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

To avoid the above problem we use the random variable

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

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

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

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

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

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

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

and variance

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

The simplest estimate of U_i is its mean, namely

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

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

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

Name	Formula $q_i =$	Constant $a =$	Return period of maximum value $T_1 =$	Applicability
Weibull	$\frac{i}{n+1}$	0	$n + 1$	All distributions, unbiased estimation of exceedance probability
Blom	$\frac{i-0.375}{n+0.25}$	0.375	$1.6\,n + 0.4$	Normal distribution, unbiased estimation of quantiles
Cunnane	$\frac{i-0.4}{n+0.2}$	0.4	$1.667\,n + 0.33$	Broad range of distributions, approx. unbiased estimation of quantiles
Gringorten	$\frac{i-0.44}{n+0.12}$	0.44	$1.786\,n + 0.21$	Gumbel distribution[†]
Hazen	$\frac{i-0.\ 5}{n}$	0.5	$2\,n$	The oldest proposed estimate; today it tends to be abandoned

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

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

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

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

[*] See also Stedinger et al. (1993) where additional formulae are also given.

[†] See chapter 6.

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

5.3.4 Probability plots

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

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

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

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

5.3.5 Demonstration of numerical probability plot

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

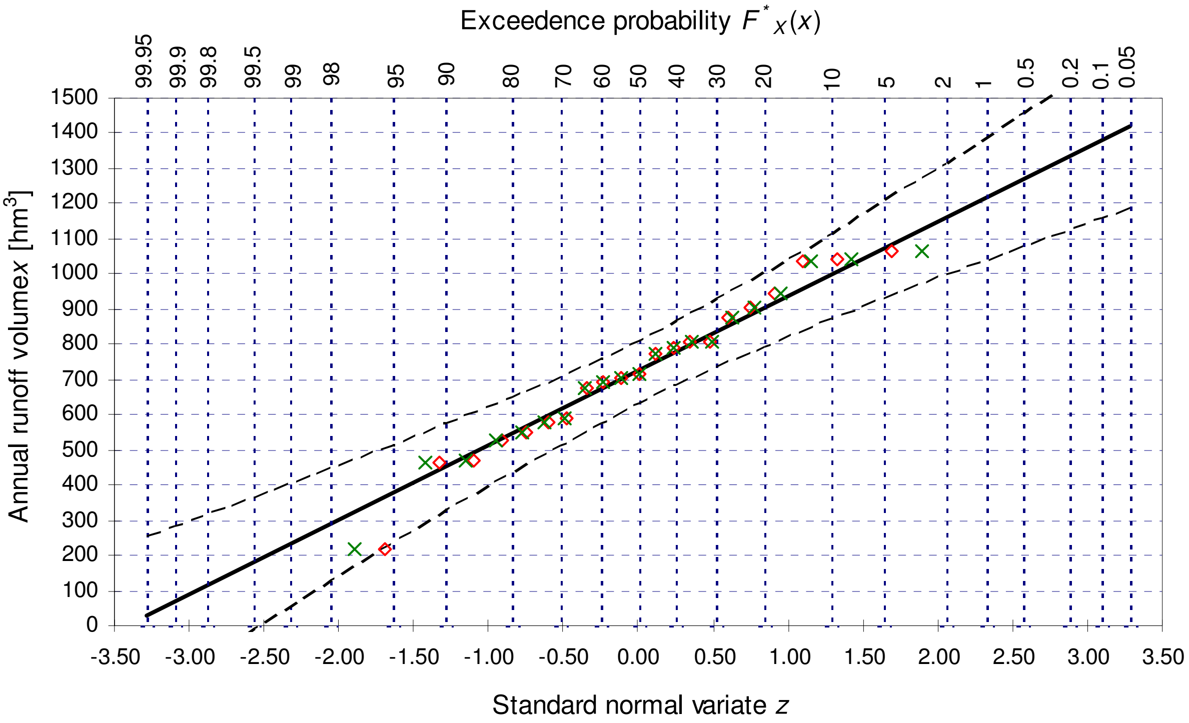

Fig. 5.3 Example of normal probability plot of the empirical distribution function using Weibull (diamonds) and Blom (symbols ×) plotting positions. For comparison, the theoretical normal distribution function N(725, 211.5) (see section 5.4.2) is also plotted (continuous line) along the corresponding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

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

Table 5.9 Demonstration of calculation of empirical exceedence probabilities.

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

5.4 Selection and fitting of the theoretical distribution function

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

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

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

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

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

5.4.1 Indications of suitability of the normal distribution for geophysical variables

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

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

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

5.4.2 Demonstration of fitting the normal distribution

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

5.5 Testing the goodness of fit of a probability model

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

5.5.1 The χ^2 test

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

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

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

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

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

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

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

where N_j is the random variable whose realization is n_j, is an aggregated measure of the differences between the actual and the theoretical number of points in all classes. If the null hypothesis is perfect, the distribution of Q is χ^2 with $k - 1$ degrees of freedom. In the most usual case of imperfect null hypothesis, the number of degrees of freedom is $k - r - 1$, where r is the number of parameters that are estimated from the sample.[*]

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

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

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

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

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

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

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

The algorithm is clarified in the following example.

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

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

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

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

5.5.2 Demonstration of testing the goodness of fit

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

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

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

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

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

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

Table 5.10 Elementary calculations demonstrating the χ^2 test.

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

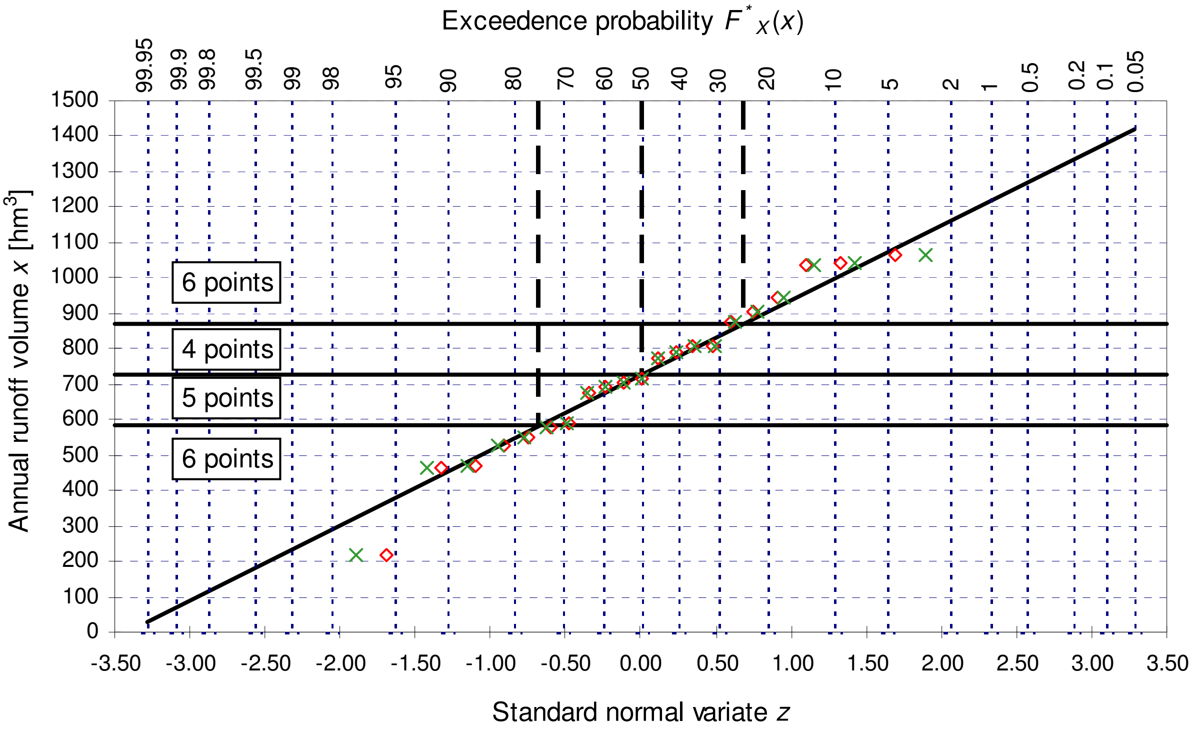

Fig. 5.4 Explanatory sketch for the numerical example of section 5.5.2.

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

From (5.14) we obtain

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

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

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

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

5.6 Statistical prediction

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

5.6.1 Demonstration of statistical prediction

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

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

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

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

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

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

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

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

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

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

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

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

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

References

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

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

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

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

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

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

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

Chapter 6

Typical distribution functions in geophysics, hydrology and water resources

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

Summary

In this chapter we describe four families of distribution functions that are used in geophysical and engineering applications, including engineering hydrology and water resources technology. The first includes the normal distribution and the distributions derived from this by the logarithmic transformation. The second is the gamma family and related distributions that includes the exponential distribution, the two- and three-parameter gamma distributions, the Log-Pearson III distribution derived from the last one by the logarithmic transformation and the beta distribution that is closely related to the gamma distribution. The third is the Pareto distribution, which in the last years tends to become popular due to its long tail that seems to be in accordance with natural behaviours. The fourth family includes the extreme value distributions represented by the generalized extreme value distributions of maxima and minima, special cases of which are the Gumbel and the Weibull distributions.

5.1 Normal Distribution and related transformations

5.1.1 Normal (Gaussian) Distribution

In the preceding chapters we have discussed extensively and in detail the normal distribution and its use in statistics and in engineering applications. Specifically, the normal distribution has been introduced in section 2.8, as a consequence of the central limit theorem, along with two closely related distributions, the χ^2 and the Student (or *t*), which are of great importance in statistical estimates, even though they are not used for the description of geophysical variables. The normal distribution has been used in chapter 3 to theoretically derive statistical estimates. In chapter 5 we have presented in detail the use of the normal distribution for the description of geophysical variables.

In summary, the normal distribution is a symmetric, two-parameter, bell shaped distribution. The fact that a normal variable X ranges from minus infinity to infinity contrasts the fact that hydrological variables are in general non-negative. This problem has been already discussed in detail in section 5.4.1. A basic characteristic of the normal distribution is that it is closed under addition or, else, a stable distribution. Consequently, the sum (and any linear combination) of normal variables is also a normal variable. Table 6.1 provides a concise summary of the basic mathematical properties and relations associated with the normal distribution, described in detail in previous chapters.

Table 6.1 Normal (Gaussian) distribution conspectus.

Probability density function	$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
Distribution function	$F_X(x) = \int_{-\infty}^{x} f_X(s)\, ds$
Range	$-\infty < x < \infty$ (continuous)
Parameters	μ: location parameter (= mean) $\sigma > 0$: scale parameter (= standard deviation)
Mean	$\mu_X = \mu$
Variance	$\sigma_X^2 = \sigma^2$
Third central moment	$\mu_X^{(3)} = 0$
Fourth central moment	$\mu_X^{(4)} = 3\sigma^4$
Coefficient of skewness	$C_{s_X} = 0$
Coefficient of kurtosis	$C_{k_X} = 3$
Mode	$x_\mathrm{p} = \mu$
Median	$x_{0.5} = \mu$
Second L moment	$\lambda_X^{(2)} = \frac{\sigma}{\sqrt{\pi}}$
Third L moment	$\lambda_X^{(3)} = 0$
L coefficient of variation	$\tau_X^{(2)} = \frac{\sigma}{\sqrt{\pi}\,\mu}$
L skewness	$\tau_X^{(3)} = 0$
L kurtosis	$\tau_X^{(4)} = 0.1226$

Typical calculations

The most typical calculations are the calculation of the value $u = F_X(x_u)$ of the distribution function for a given x_u, or inversely, the calculation of the u-quantile of the variable, i.e. the calculation of x_u, when the probability u is known. The fact that the integral defining the normal distribution function (Table 6.1) does not have an analytical expression, creates difficulties in the calculations. A simple solution is the use of tabulated values of the standardized normal variable $z = (x - \mu) / \sigma$, which is a normal variable with zero mean and standard deviation equal to 1 (section 2.6.1 and Table A1 in Appendix). Thus, the calculation of the u-quantile (x_u) becomes straightforward by

$$x_u = \mu + z_u \sigma \tag{6.1}$$

where z_u, corresponding to $u = F_Z(z_u)$, is taken from Table A1. Conversely, for a given x_u, z_u is calculated by (6.1) and $u = F_Z(z_u)$ is determined from Table A1.

Several numerical approximations of the normal distribution function are given in the literature, which can be utilized to avoid use of tables (Press *et al.*, 1987; Stedinger *et al.*, 1993; Koutsoyiannis, 1997), whereas most common computer applications (e.g. spreadsheets*) include ready to use functions.

Parameter estimation

As we have seen in section 3.5, both the method of moments and the maximum likelihood result in the same estimates of the parameters of normal distribution, i.e.,

$$\mu = \overline{x}\,, \quad \sigma = s_X \tag{6.2}$$

We notice that s_X in (6.2) is the biased estimate of the standard deviation. Alternatively, the unbiased estimation of standard deviation is preferred sometimes. The method of L moments can be used as an alternative (see Table 6.1) to estimate the parameters based on the mean and the second L moment.

Standard error and confidence intervals of quantiles

In section 3.4.6 we defined the standard error and the confidence intervals of the quantile estimation and we presented the corresponding equations for the normal distribution. Summarising, the point estimate of the normal distribution u-quantile is

$$\hat{x}_u = \overline{x} + z_u s_x \tag{6.3}$$

the standard error of the estimation is

$$\varepsilon_u = \frac{s_X}{\sqrt{n}}\sqrt{1 + \frac{z_u^2}{2}} \tag{6.4}$$

and the corresponding confidence limits for confidence coefficient γ are

$$\hat{x}_{u_{1,2}} \approx (\overline{x} + z_u\, s_X)\ \pm z_{(1+\gamma)/2}\frac{s_X}{\sqrt{n}}\sqrt{1 + \frac{z_u^2}{2}} = \hat{x}_u \pm z_{(1+\gamma)/2}\frac{s_X}{\sqrt{n}}\sqrt{1 + \frac{z_u^2}{2}} \tag{6.5}$$

Normal distribution probability plot

As described in section 5.3.4, the normal distribution is depicted as a straight line in a normal probability plot. This depiction is equivalent to plotting the values of the variable x (in the vertical axis) versus and the standardized normal variate z (in the horizontal axis).

5.1.2 Two-parameter log-normal distribution

The two-parameter log-normal distribution results from the normal distribution using the transformation

$$y = \ln x \leftrightarrow x = \mathrm{e}^y \tag{6.6}$$

* In Excel, these functions are NormDist, NormInv, NormSDist and NormSInv.

Thus, the variable X has a two-parameter log-normal distribution if the variable Y has normal distribution $N(\mu_Y, \sigma_Y)$. Table 6.2 summarizes the mathematical properties and relations associated with the two-parameter log-normal distribution.

Table 6.2 Two-parameter log-normal distribution conspectus.

Probability density function	$f_X(x) = \dfrac{1}{x\sqrt{2\pi}\sigma_Y} \mathrm{e}^{-\frac{1}{2}\left(\frac{\ln x - \mu_Y}{\sigma_Y}\right)^2}$
Distribution function	$F_X(x) = \int_0^x f_X(s)\,ds$
Range	$0 < x < \infty$ (continuous)
Parameters	μ_Y: scale parameter $\sigma_Y > 0$: shape parameter
Mean	$\mu_X = \mathrm{e}^{\mu_Y + \frac{\sigma_Y^2}{2}}$
Variance	$\sigma_X^2 = \mathrm{e}^{2\mu_Y + \sigma_Y^2}\left(\mathrm{e}^{\sigma_Y^2} - 1\right)$
Third central moment	$\mu_X^{(3)} = \mathrm{e}^{3\mu_Y + \frac{3\sigma_Y^2}{2}}\left(\mathrm{e}^{3\sigma_Y^2} - 3\mathrm{e}^{\sigma_Y^2} + 2\right)$
Coefficient of skewness	$C_{v_X} = \sqrt{\mathrm{e}^{\sigma_Y^2} - 1}$
Coefficient of kurtosis	$C_{s_X} = 3C_{v_X} + C_{v_X}^3$
Mode	$x_p = \mathrm{e}^{\mu_Y - \sigma_Y^2}$
Median	$x_{0.5} = \mathrm{e}^{\mu_Y}$

A direct consequence of the logarithmic transformation (6.6) is that the variable X is always positive. In addition, it results from Table 6.2 that the distribution has always positive skewness and that its mode is different from zero. Thus, the shape of the probability density function is always bell-shaped and positively skewed. These basic attributes of the log-normal distribution are compatible with observed properties of many geophysical variables, and therefore it is frequently used in geophysical applications. It can be easily shown that the product of two variables having a two-parameter log-normal distribution, has also a two-parameter log-normal distribution. This property, combined with the central limit theorem and taking into account that in many cases the variables can be considered as a product of several variables instead of a sum, has provided theoretical grounds for the frequent use of the distribution in geophysics.

Typical calculations

Typical calculations of the log-normal distribution are based on the corresponding calculations of the normal distribution. Thus, combining equations (6.1) and (6.6) we obtain

$$y_u = \mu_Y + z_u\sigma_Y \Leftrightarrow x_u = \mathrm{e}^{\mu_Y + z_u\sigma_Y} \tag{6.7}$$

where z_u is the u-quantile of the standardized normal variable.

Parameter estimation

Using the equations of Table 6.2, the method of moments results in:

$$\sigma_Y = \sqrt{\ln\left(1 + s_X^2 / \bar{x}^2\right)}, \quad \mu_Y = \ln\bar{x} - \sigma_Y^2 / 2 \tag{6.8}$$

Parameter estimation using the maximum likelihood method gives (e.g. Kite, 1988, p. 57)

$$\mu_Y = \sum_{i=1}^{n} \ln x_i / n = \bar{y}, \quad \sigma_Y = \sqrt{\sum_{i=1}^{n} (\ln x_i - \mu_Y)^2 / n} = s_Y \tag{6.9}$$

We observe that the two methods differ not only in the resulted estimates, but also in that they are based on different sample characteristics. Namely, the method of moments is based on the mean and the (biased) standard deviation of the variable *X* while the maximum likelihood method is based on the mean and the (biased) standard deviation of the logarithm of the variable *X*.

Standard error and confidence intervals of quantiles

Provided that the maximum likelihood method is used to estimate the parameters of the log-normal distribution, the point estimate of the *u*-quantile of *y* and *x* is then given by

$$\hat{y}_u = \ln(\hat{x}_u) = \bar{y} + z_u s_Y \Rightarrow \hat{x}_u = \mathrm{e}^{\bar{y} + z_u s_Y} \tag{6.10}$$

where z_u is the *u*-quantile of the standard normal distribution. The square of the standard error of the *Y* estimate is given by:

$$\varepsilon_Y^2 = \mathrm{Var}(\hat{Y}_u) = \mathrm{Var}(\ln \hat{X}_u) = \frac{s_Y^2}{n}\left(1 + \frac{z_u^2}{2}\right) \tag{6.11}$$

Combining these equations we obtain the following approximate relationship which gives the confidence intervals of x_u for confidence level γ

$$\hat{x}_{u_{1,2}} \approx \exp\left[(\bar{y} + z_u s_Y) \pm z_{(1+\gamma)/2} \frac{s_Y}{\sqrt{n}} \sqrt{1 + \frac{z_u^2}{2}}\right] = \hat{x}_u \exp\left[\pm z_{(1+\gamma)/2} \frac{s_Y}{\sqrt{n}} \sqrt{1 + \frac{z_u^2}{2}}\right] \tag{6.12}$$

where $z_{(1+\gamma)/2}$ is the $[(1+\gamma)/2]$-quantile of the standard normal distribution. When the parameter estimation is based in the method of moments, the standard error and the corresponding confidence intervals are different (see Kite 1988, p. 60).

Log-normal distribution probability plot

The normal distribution probability plot can be easily transformed in order for the log-normal distribution to be depicted as a straight line. Specifically, a logarithmic vertical axis has to be used. This depiction is equivalent to plotting the logarithm of the variable, ln *x*, (in the vertical axis) versus the standard normal variate (in the horizontal axis).

Numerical example

Table 6.3 lists the observations of monthly runoff of the Evinos river basin, central-western Greece, upstream of the hydrometric gauge at Poros Reganiou, for the month of January. We wish to fit the two-parameter log-normal distribution to the data and estimate the 50-year discharge.

Table 6.3 Observed sample of January runoff volume (in hm^3) at the hydrometric station of Poros Riganiou of the Evinos river.

Hydrological year	Runoff	Hydrological year	Runoff	Hydrological year	Runoff
1970-71	102	1977-78	121	1984-85	178
1971-72	74	1978-79	317	1985-86	185
1972-73	78	1979-80	213	1986-87	101
1973-74	48	1980-81	111	1987-88	57
1974-75	31	1981-82	82	1988-89	24
1975-76	48	1982-83	61	1989-90	22
1976-77	114	1983-84	133	1990-91	51

The sample mean is

$$\bar{x} = \sum x / n = 102.4 \text{ hm}^3$$

The standard deviation (biased estimate) is

$$s_X = \left(\sum x^2 / n - \bar{x}^2\right)^{1/2} = 70.4 \text{ hm}^3$$

and the coefficient of variation

$$\hat{C}_{\mathrm{v}_X} = s_X / \bar{x} = 70.4 / 102.4 = 0.69$$

The skewness coefficient (biased estimate) is

$$\hat{C}_{\mathrm{s}_X} = 1.4$$

These coefficients of variation and skewness suggest a large departure from the normal distribution.

The method of moments results in

$$\sigma_Y = \sqrt{\ln\left(1 + s_X^2 / \bar{x}^2\right)} = 0.622, \quad \mu_Y = \ln \bar{x} - \sigma_Y^2 / 2 = 4.435$$

whereas the maximum likelihood estimates are

$$\mu_Y = \sum \ln x / n = 4.404, \quad \sigma_Y = \sqrt{\sum (\ln x)^2 / n - \mu_Y^2} = 0.687$$

The 50-year discharge can be estimated from $x_u = \exp(\mu_Y + z_u\, \sigma_Y)$ where $u = 1 - 1/50 = 0.98$ and $z_u = 2.054$ (Table A1). Using the parameters estimated by the method of moments we obtain $x_{0.98} = 302.7$ hm^3, while using the maximum likelihood parameter estimates we get $x_{0.98} = 335.1$. In the latter case the 95% confidence interval for that value is (based on (6.12), for $z_u = 2.054$ and $z_{(1+\gamma)/2} = 1.96$):

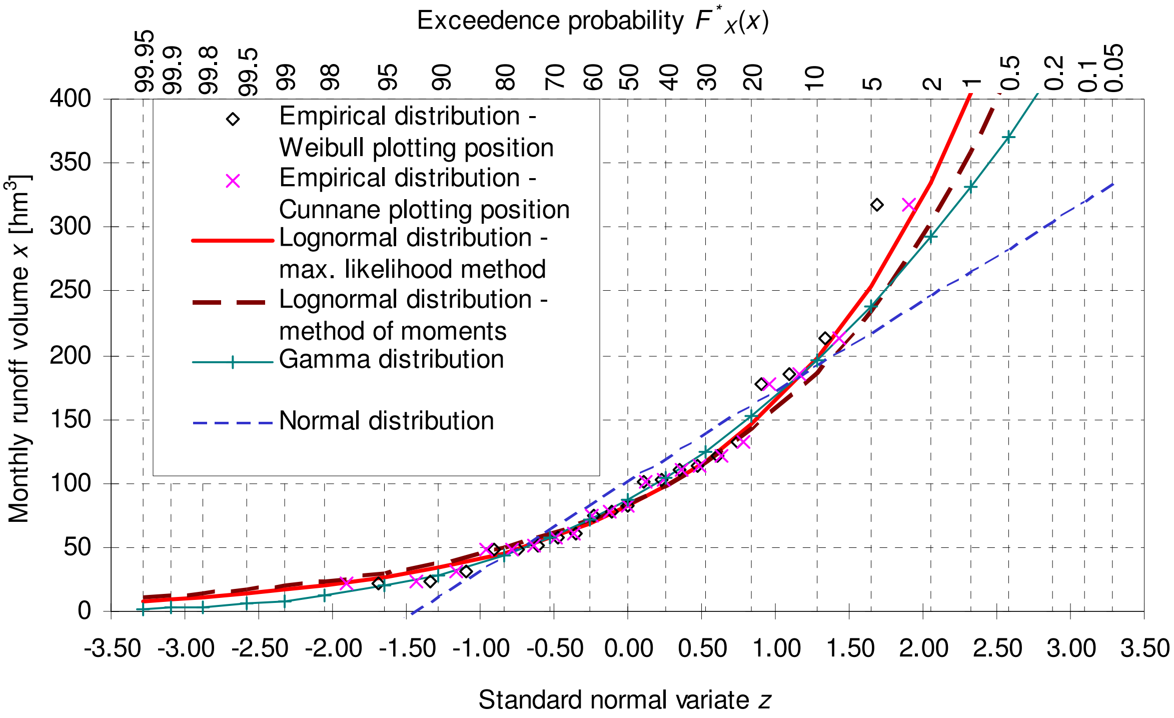

Fig. 6.1 Alternative empirical and theoretical distribution functions of the January runoff at Poros Riganiou (normal probability plot).

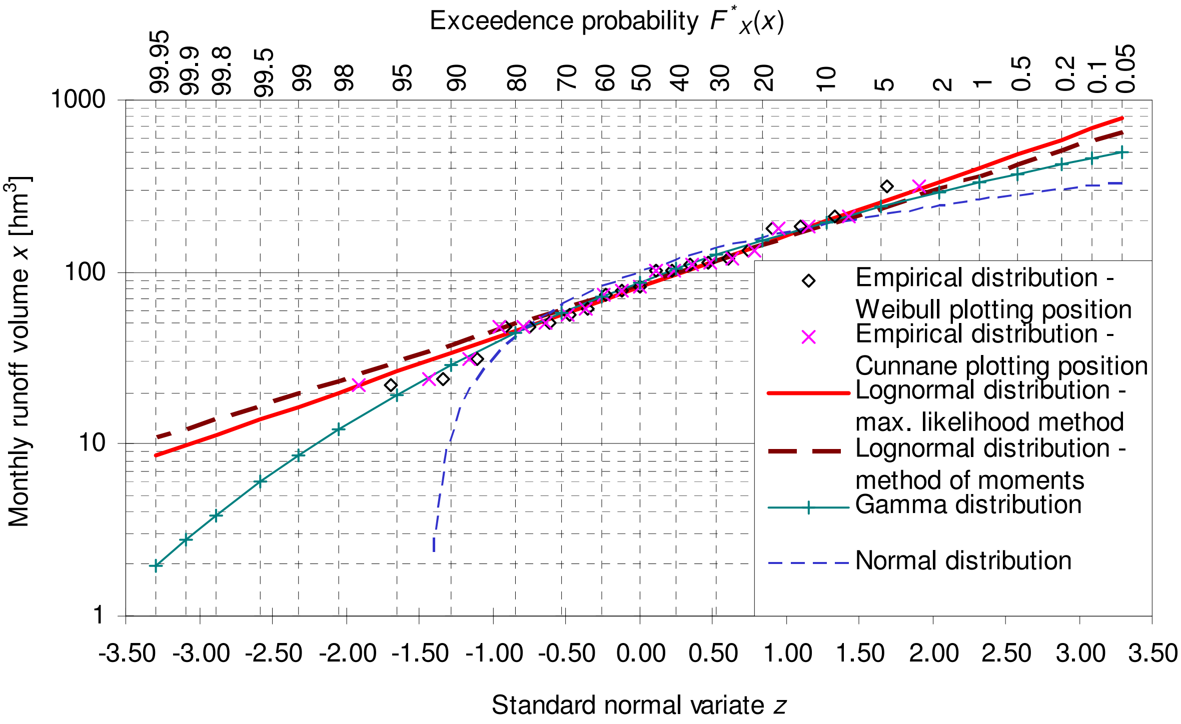

Fig. 6.2 Alternative empirical and theoretical distribution functions of the January runoff at Poros Riganiou (lognormal probability plot).

$$\hat{x}_{u_{1,2}} \approx \exp\left[4.404 + 2.054 \times 0.687 \pm 1.96 \times \frac{0.687}{\sqrt{21}} \times \sqrt{1 + \frac{2.054^2}{2}}\right]$$

$$= \exp(5.815 \pm 0.518) = \begin{cases} 562.8 \\ 199.7 \end{cases}$$

The huge width of the confidence interval reflects a poor reliability of the prediction of the 50-year January runoff. The reduction of the uncertainty would be made possible only by a substantially larger sample.

To test the appropriateness of the log-normal distribution we can use the χ^2 test (see section 5.5.1). As an empirical alternative, we depict in Fig. 6.1 and Fig. 6.2 comparisons of the empirical distribution function and the fitted log-normal theoretical distribution functions, on normal probability plot and on log-normal probability plot, respectively. For the empirical distribution we have used two plotting positions, the Weibull and the Cunnane (Table 5.8). Both log-normal distribution plots, resulted from the methods and the maximum likelihood are shown in the figures. Clearly, the maximum likelihood method results in a better fit in the region of small exceedence probabilities. For comparison we have also plotted the normal distribution, which apparently does not fit well to the data, and the Gamma distribution (see section 5.2.2).

5.1.3 Three-parameter log-normal (Galton) distribution

A combination of the normal distribution and the modified logarithmic transformation

$$y = \ln(x - \zeta) \Leftrightarrow x = \zeta + \mathrm{e}^{y} \tag{6.13}$$

results in the three-parameter log-normal distribution or the Galton distribution. This distribution has an additional parameter, compared to the two-parameter log-normal, the location parameter ζ, which is the lower limit of the variable. This third parameter results in a higher flexibility of the distribution fit. Specifically, if the method of moments is used to fit the distribution, the third parameter makes possible the preservation of the coefficient of skewness. Table 6.4 summarizes the basic mathematical properties and equations associated with the three-parameter log-normal distribution.

Typical calculations

The three-parameter log-normal distribution, can be handled in a similar manner with the two-parameter log-normal distribution according to the following relationship

$$y_u = \mu_Y + z_u \sigma_Y \Leftrightarrow x_u = \zeta + \mathrm{e}^{\mu_Y + z_u \sigma_Y} \tag{6.14}$$

where z_u is the u-quantile of the standard normal distribution.

Parameter estimation

Using the equations of Table 6.4 for the method of moments, and after algebraic manipulation we obtain the following relationships that estimate the parameter σ_Y.

$$\sigma_Y = \sqrt{\ln\left(1+\phi^2\right)} \tag{6.15}$$

where

$$\phi = \frac{1-\omega^{2/3}}{\omega^{1/3}}, \quad \omega = \frac{-\hat{C}_{s_X} + \sqrt{\hat{C}_{s_X}^2 + 4}}{2} \tag{6.16}$$

Table 6.4 Three-parameter log-normal distribution conspectus

Probability density function	$f_X(x) = \dfrac{1}{(x-\zeta)\sqrt{2\pi}\sigma_Y} \mathrm{e}^{-\frac{1}{2}\left(\frac{\ln(x-\zeta)-\mu_Y}{\sigma_Y}\right)^2}$
Distribution function	$F_X(x) = \int_c^x f_X(s)\,ds$
Range	$\zeta < x < \infty$ (continuous)
Parameters	ζ: location parameter μ_Y: scale parameter $\sigma_Y > 0$: shape parameter
Mean	$\mu_X = \zeta + \mathrm{e}^{\mu_Y + \frac{\sigma_Y^2}{2}}$
Variance	$\sigma_X^2 = \mathrm{e}^{2\mu_Y + \sigma_Y^2}\left(\mathrm{e}^{\sigma_Y^2} - 1\right)$
Third central moment	$\mu_X^{(3)} = \mathrm{e}^{3\mu_Y + \frac{3\sigma_Y^2}{2}}\left(\mathrm{e}^{3\sigma_Y^2} - 3\mathrm{e}^{\sigma_Y^2} + 2\right)$
Coefficient of skewness	$C_{s_X} = 3\left(\mathrm{e}^{\sigma_Y^2} - 1\right)^{1/2} + \left(\mathrm{e}^{\sigma_Y^2} - 1\right)^{3/2}$
Mode	$x_p = \zeta + \mathrm{e}^{\mu_Y - \sigma_Y^2}$
Median	$x_{0.5} = \zeta + \mathrm{e}^{\mu_Y}$

The other two parameters of the distribution can be calculated from

$$\mu_Y = \ln(s_X/\varphi) - \sigma_Y^2/2 \qquad \zeta = \bar{x} - \frac{s_X}{\varphi} \tag{6.17}$$

The maximum likelihood method is based on the following relationships (e.g. Kite, 1988, p. 74)

$$\mu_Y = \sum_{i=1}^{n} \ln\left(x_i - c\right)/n, \quad \sigma_Y^2 = \sum_{i=1}^{n}\left[\ln\left(x_i - c\right) - \mu_Y\right]^2/n \tag{6.18}$$

$$\left(\mu_Y - \sigma_Y^2\right)\sum_{i=1}^{n}\frac{1}{x_i - c} = \sum_{i=1}^{n}\frac{\ln\left(x_i - c\right)}{x_i - c} \tag{6.19}$$

that can be solved only numerically.

The estimation of confidence intervals for the three-parameter log-normal distribution is complicated. The reader can consult Kite (1988, p. 77).

5.2 The Gamma family and related distribution functions

5.2.1 Exponential distribution

A very simple yet useful distribution is the exponential. Its basic characteristics are summarized in Table 6.5.

Table 6.5 Exponential distribution conspectus

Probability density function	$f_X(x) = \dfrac{\mathrm{e}^{-\frac{x-\zeta}{\lambda}}}{\lambda}$
Distribution function	$F_X(x) = 1 - \mathrm{e}^{-\frac{x-\zeta}{\lambda}}$
Variable range	$\zeta < x < \infty$ (continuous)
Parameters	ζ: location parameter $\lambda > 0$: scale parameter
Mean	$\mu_X = \zeta + \lambda$
Variance	$\sigma_X^2 = \lambda^2$
Third central moment	$\mu_X^{(3)} = \lambda^3$
Fourth central moment	$\mu_X^{(4)} = 9\lambda^4$
Coefficient of variation	$C_{v_X} = \dfrac{\lambda}{\zeta + \lambda}$
Coefficient of skewness	$C_{s_X} = 2$
Coefficient of kurtosis	$C_{k_X} = 9$
Mode	$x_p = \zeta$
Median	$x_{0.5} = \zeta + \lambda \ln 2$
Second L moment	$\lambda_X^{(2)} = \lambda/2$
Third L moment	$\lambda_X^{(3)} = \lambda/6$
Fourth L moment	$\lambda_X^{(4)} = \lambda/12$
L coefficient of variation	$\tau_X^{(2)} = \dfrac{\lambda}{2(\lambda + \zeta)}$
L skewness	$\tau_X^{(3)} = 1/3$
L kurtosis	$\tau_X^{(4)} = 1/6$

In its simplesr form, as we have already seen in section 2.5.5, the exponential distribution has only one parameter, the location parameter λ (the second parameter ζ is 0). The probability density function of the exponential distribution is a monotonically decreasing function (it has an inverse J shape).

As we have already seen (section 2.5.5), the exponential distribution can be used to describe non-negative geophysical variables at a fine time scale (e.g. hourly or daily rainfall depths). In addition, a theorem in probability theory states that intervals between random points in time, have exponential distribution. Application of this theorem in geophysics suggests that, for instance, the time intervals between rainfall events have exponential distribution. This is verified only as a rough approximation. The starting times of rainfall events cannot be regarded as random points in time; rather, a clustering behaviour is evident, which is related to some dependence in time (Koutsoyiannis, 2006). Moreover, the duration of rainfall events and the total rainfall depth in an event have been frequently assumed to have exponential distribution. Again this is just a rough approximation (Koutsoyiannis, 2005).

5.2.2 Two-parameter Gamma distribution

The two-parameter Gamma distribution is one of the most commonly used in geophysics and engineering hydrology. Its basic characteristics are given in Table 6.6.

Table 6.6 Two-parameter Gamma distribution conspectus.

Probability density function	$f_X(x) = \frac{1}{\lambda^\kappa \Gamma(\kappa)} x^{\kappa-1} \mathrm{e}^{-x/\lambda}$
Distribution function	$F_X(x) = \int_0^x f_X(s)\,ds$
Range	$0 < x < \infty$ (continuous)
Parameters	$\lambda > 0$: scale parameter $\kappa > 0$: shape parameter
Mean	$\mu_X = \kappa\lambda$
Variance	$\sigma_X^2 = \kappa\lambda^2$
Third central moment	$\mu_X^{(3)} = 2\kappa\lambda^3$
Fourth central moment	$\mu_X^{(4)} = 3\kappa(\kappa+2)\lambda^4$
Coefficient of variation	$C_{v_X} = \frac{1}{\sqrt{\kappa}}$
Coefficient of skewness	$C_{s_X} = \frac{2}{\sqrt{\kappa}} = 2C_{v_X}$
Coefficient of kurtosis	$C_{k_X} = 3 + \frac{6}{\kappa} = 3 + 6C_{v_X}^2$
Mode	$x_\mathrm{p} = (\kappa - 1)\,\lambda$ (for $\kappa \geq 1$) $x_\mathrm{p} = 0$ (for $\kappa \leq 1$)

Similar to the two-parameter log-normal distribution, the Gamma distribution is positively skewed and is defined only for nonnegative values of the variable. These characteristics make the Gamma distribution compatible with several geophysical variables, including monthly and annual flows and precipitation depths.

The Gamma distribution has two parameters, the scale parameter λ and the shape parameter κ. For $\kappa = 1$ the distribution is identical with the exponential, which is a special case of Gamma. For $\kappa > 1$ the probability density function is bell-shaped, whereas for $\kappa < 1$ its shape becomes an inverse J, with an infinite ordinate at $x = 0$. For large κ values (above 15-30) the Gamma distribution approaches the normal.

The Gamma distribution, similar to the normal, is closed under addition, but only when the added variables are stochastically independent and have the same scale parameter. Thus, the sum of two independent variables that have Gamma distribution with common scale parameter λ, has also a Gamma distribution.

The χ^2 distribution, which has been discussed in section 2.10.4, is a special case of the Gamma distribution.

Typical calculations

Similar to the normal distribution, the integral in the Gamma distribution function does not have an analytical expression thus causing difficulties in calculations. A simple solution is to tabulate the values of the standardized variable $k = (x - \mu_X) / \sigma_X$, where μ_X and σ_X is the mean value and standard deviation of X, respectively. Such tabulations are very common in statistics books; one is provided in Table A4 in Appendix. Each column of this table corresponds to a certain value of κ (or, equivalently, to a certain skewness coefficient value $C_{\mathrm{s}X} = 2 / \sqrt{\kappa} = 2\sigma_X / \bar{x}$).The u-quantile (x_u) is then given by

$$x_u = \mu_X + k_u \sigma_X \tag{6.20}$$

where k_u is read from tables for the specified value of $u = F_K(k_u)$. Conversely, for given x_u, the k_u value can be calculated from (6.1) and then $u = F_K(k_u)$ is taken from tables (interpolation in a column or among adjacent columns may be necessary).

Several numerical approaches can be found in literature in order to avoid the use of tables (Press *et al.*, 1987; Stedinger *et al.*, 1993; Koutsoyiannis, 1997) whereas most common computer applications (e.g. spreadsheets[*]) include ready to use functions.

Parameter estimation

The implementation of the method of moments results in the following simple estimates of the two Gamma distribution parameters:

$$\kappa = \frac{\bar{x}^2}{s_X^2}, \quad \lambda = \frac{s_X^2}{\bar{x}} \tag{6.21}$$

Parameter estimation based on the maximum likelihood method is more complicated. It is based in the solution of the equations (cf. e.g. Bobée and Ashkar, 1991)

$$\ln \kappa - \psi(\kappa) = \ln \bar{x} - \frac{1}{n} \sum_{i=1}^{n} \ln x_i , \quad \lambda = \frac{\bar{x}}{\kappa} \tag{6.22}$$

[*] In Excel, these functions are GammaDist and GammaInv.

where $\psi(\kappa) = \mathrm{d} \ln \Gamma(\kappa) / \mathrm{d}\kappa$ is the so-called Digamma function (derivative of the logarithm of Gamma function).

Standard error and confidence intervals of quantiles

A point estimate of the *u*-quantile of Gamma distribution is given by

$$\hat{x}_u = \bar{x} + k_u s_X \tag{6.23}$$

If the method of moments is used to estimate the parameters the square of standard error of the estimate is (Bobée and Ashkar, 1991, p. 50)

$$\varepsilon_u^2 = \frac{s_X^2}{n}\left[\left(1 + k_u C_{v_X}\right)^2 + \frac{1}{2}\left(k_u + 2C_{v_X}\frac{\partial k_u}{\partial C_{s_X}}\right)^2\left(1 + C_{v_X}\right)^2\right] \tag{6.24}$$

In a first rough approximation, the term $\partial k_u / \partial C_{sX}$ can be omitted, leading to the simplification

$$\varepsilon_u^2 = \frac{s_X^2}{n}\left[1 + 2C_{v_X} k_u + \frac{1}{2}\left(1 + 3C_{v_X}^2\right)k_u^2\right] \tag{6.25}$$

Thus, an approximation of the confidence limits for confidence coefficient γ is

$$\hat{x}_{u_{1,2}} \approx (\bar{x} + k_u s_X) \pm z_{(1+\gamma)/2}\frac{s_X}{\sqrt{n}}\sqrt{1 + 2C_{v_X} k_u + \frac{1}{2}\left(1 + 3C_{v_X}^2\right)k_u^2} \tag{6.26}$$

The maximum likelihood method results in more complicated calculations of the confidence intervals. The interested reader may consult Bobée and Ashkar (1991, p. 46).

Gamma distribution probability plot

It is not possible to construct a probability paper that depicts any Gamma distribution as straight line. It is feasible, though, to create a Gamma probability paper for a specified shape parameter κ. Clearly, this is not practical, and thus the depiction of Gamma distribution is usually done on normal probability paper or on Weibull probability paper (see below). In that case obviously the distribution is not depicted as a straight line but as a curve.

Numerical example

We wish to fit a two-parameter Gamma distribution to the sample of January runoff of the river Evinos upstream of the hydrometric station of Poros Riganiou and to determine the 50-year runoff (sample in Table 6.3).

The sample mean value is 102.4 hm^3 and the sample standard deviation is 70.4 hm^3; using the method of moments we obtain the following parameter estimates:

$$\kappa = 102.4^2 / 70.4^2 = 2.11, \lambda = 70.4^2 / 102.4 = 48.4 \text{ hm}^3.$$

For return period $T = 50$ or equivalently for probability of non-exceedence $F = 0.98 = u$ we determine the quantile x_u either by an appropriate computer function or from tabulated standardized quantile values (Table A4); we find $k_{0.98} = 2.70$ and

$$x_u = 102.4 + 2.70 \times 70.4 = 292.5 \text{ hm}^3$$

Likewise, we can calculate a series of quantiles, thus enabling the depiction of the fitted Gamma distribution. This has been done in Fig. 6.1 (in normal probability plot) and in Fig. 6.2 (in log-normal probability plot) in comparison with other distributions. We observe that in general the Gamma distribution fit is close to those of the log-normal distribution; in the region of small exceedence probabilities the log-normal distribution provides a better fit.

To determine the 95% confidence intervals for the 50-year discharge we use the approximate relationship (6.26), which for $z_{(1+\gamma)/2} = 1.96$, $k_u = 2.70$ and $C_{vX} = 0.69$ results in

$$\begin{aligned}\hat{x}_{u_{1,2}} &\approx 292.5 \pm 1.96 \times \frac{70.4}{\sqrt{21}} \times \sqrt{1 + 2 \times 0.69 \times 2.70 + \frac{1}{2}\left(1 + 3 \times 0.69^2\right) \times 2.70^2} \\ &\approx 292.5 \pm 110.9 = \begin{cases} 403.4 \\ 181.6 \end{cases}\end{aligned}$$

5.2.3 Three-parameter Gamma distribution (Pearson III)

The addition of a location parameter (ζ) to the two-parameter Gamma distribution, results in the three-parameter Gamma distribution or the so-called Pearson type III (Table 6.7).

Table 6.7 Pearson type III distribution conspectus.

Probability density function	$f_X(x) = \frac{1}{\lambda^\kappa \Gamma(\kappa)} (x - \zeta)^{\kappa - 1} \mathrm{e}^{-(x-\zeta)/\lambda}$
Distribution function	$F_X(x) = \int_c^x f_X(s)\, ds$
Range	$\zeta < x < \infty$ (continuous)
Parameters	ζ: location parameter $\lambda > 0$: scale parameter $\kappa > 0$: shape parameter
Mean	$\mu_X = c + \kappa\lambda$
Variance	$\sigma_X^2 = \kappa\lambda^2$
Third central moment	$\mu_X^{(3)} = 2\kappa\lambda^3$
Fourth central moment	$\mu_X^{(4)} = 3\kappa(\kappa + 2)\lambda^4$
Coefficient of skewness	$C_{s_X} = \frac{2}{\sqrt{\kappa}}$
Coefficient of kurtosis	$C_{k_X} = 3 + \frac{6}{\kappa}$
Mode	$x_{\mathrm{p}} = \zeta + (\kappa - 1)\ \lambda$ (for $\kappa \geq 1$) $x_{\mathrm{p}} = \zeta$ (for $\kappa \leq 1$)

The location parameter ζ, which is the lower limit of the variable, enables a more flexible fit to the data. Thus, if we use the method of moments to fit the distribution, the third parameter permits the preservation of the coefficient of skewness.

The basic characteristics are similar to those of the two-parameter Gamma distribution. Typical calculations are also based in equation (6.20). In contrast, the equations used for parameter estimation differ. Thus, the method of moments results in

$$\kappa = \frac{4}{\hat{C}^2_{s_X}}, \quad \lambda = \frac{s_X}{\sqrt{\kappa}}, \quad \zeta = \bar{x} - \kappa\lambda \tag{6.27}$$

The maximum likelihood method results in more complicated equations. The interested reader may consult Bobée and Ashkar (1991, p. 59) and Kite (1988, p. 117) who also provide formulae to estimate the standard error and confidence intervals of distribution quantiles.

5.2.4 Log-Pearson III distribution

The Log-Pearson III results from the Pearson type III distribution and the transformation

$$y = \ln x \Leftrightarrow x = \mathrm{e}^y \tag{6.28}$$

Thus, the random variable *X* has Log-Pearson III distribution if the variable *Y* has Pearson III. Table 6.8 summarizes the basic mathematical relationships for the Log-Pearson III distribution.

Table 6.8 Log Pearson III distribution conspectus.

Probability density function	$f_X(x) = \frac{1}{x\lambda^\kappa \Gamma(\kappa)} (\ln x - \zeta)^{\kappa-1} \mathrm{e}^{-(\ln x - \zeta)/\lambda}$
Distribution function	$F_X(x) = \int_{e^\zeta}^{x} f_X(s) ds$
Range	$\mathrm{e}^\zeta < x < \infty$ (continuous)
Parameters	ζ: scale parameter $\lambda > 0$: shape parameter $\kappa > 0$: shape parameter
Mean	$\mu_X = \mathrm{e}^\zeta \left(\frac{1}{1-\lambda}\right)^\kappa, \quad \lambda < 1$
Variance	$\sigma_X^2 = \mathrm{e}^{2\zeta}\left[\left(\frac{1}{1-2\lambda}\right)^\kappa - \left(\frac{1}{1-\lambda}\right)^{2\kappa}\right], \quad \lambda < 1/2$
Raw moments of order *r*	$m_X^{(r)} = \mathrm{e}^{r\zeta}\left(\frac{1}{1-r\lambda}\right)^\kappa, \quad \lambda < 1/r$

The probability density function of the Log-Pearson III distribution can take several shapes like bell-, inverse-J-, U-shape and others. From Table 6.8 we can be conclude that the *r*th moment tends to infinity for $\lambda = 1/r$ and does not exist for greater λ. This shows that the distribution has a long tail (see section 2.5.6), which has made it a popular choice in engineering hydrology. Thus, it has been extensively used to describe flood discharges; in the USA the Log-Pearson III has been recommended by national authorities as the distribution of choice for floods.

Typical calculations

Typical calculations for the Log-Pearson III are based on those related to the Pearson III. Hence, a combination of the equations (6.20) and (6.28) gives

$$y_u = \mu_Y + k_u \sigma_Y \Leftrightarrow x_u = \mathrm{e}^{\mu_Y + k_u \sigma_Y} \tag{6.29}$$

where the standard Gamma variate k_u can be determined either from tables or numerically as described in section 5.2.2.

Parameter estimation

The parameter estimation by either the method of moments or the maximum likelihood is quite complicated (Bobée and Ashkar, 1991, p. 85· Kite, 1988, p. 138). Here we present a simpler *method of moments of logarithms*: According to this method we calculate the values $y_i = \ln x_i$ from the available sample and then we calculate the statistics of the values y_i. Finally, we apply the equations resulted from the method of moments for the variable Y, thus we have

$$\kappa = \frac{4}{\hat{C}_{sY}^2}, \quad \lambda = \frac{s_Y}{\sqrt{\kappa}}, \quad \zeta = \bar{y} - \kappa\lambda \tag{6.30}$$

As in the case of the Pearson III distribution, the estimation of the confidence intervals is pretty complicated.

Log-Pearson III probability plot

It is not possible to construct a probability paper that depicts any Log-Pearson III distribution as a straight line. Of course it is possible to make a probability paper for a specified value of the shape parameter κ but this is impractical. Thus, the depiction of the Log-Pearson III distribution is usually done on Log-normal probability paper or on Gumbel probability paper (see below). In that case the distribution is not depicted as a straight line but as a curve.

5.2.5 Two-parameter Beta distribution

The Beta distribution is an important distribution of the probability theory and has been extensively used as a conditional distribution and in Bayesian statistics. Moreover, the two-parameter Beta distribution is related to the Gamma distribution. Specifically, if X and Y are independent random variables with distributions Gamma(α, θ) and Gamma(β, θ) respectively (where Gamma(α, θ) denotes a Gamma distribution with shape parameter α and scale parameter θ), then the random variable $X / (X + Y)$ has Beta(α, β) distribution. A basic property of the Beta distribution is that the variable ranges from 0 to 1, contrary to the other distributions examined that are unbounded from above. The Beta distribution is frequently used in geophysics for doubly bounded variables, e.g. relative humidity.

The Beta distribution has two shape parameters, α and β whereas an additional scale parameter could be easily added. Depending on the parameter values, the probability density function of the Beta distribution can take a plethora of shapes. Specifically, for $\alpha = \beta = 1$ it becomes identical to the uniform distribution, while for $\alpha = 1$ and $\beta = 2$ (or $\alpha = 2$ and $\beta = 1$) it is identical to the negatively (positively) skewed triangular distribution. If $\alpha < 1$ (or $\beta < 1$) the probability density function is infinite at point $x = 0$ ($x = 1$). If $\alpha > 1$ and $\beta > 1$ the Beta probability density function is bell shaped. Table 6.9 summarizes the basic properties of the Beta distribution.

Table 6.9 Two-parameter Beta distribution conspectus.

Probability density function	$f_X(x) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
Distribution function	$F_X(x) = \int_0^x f_X(s)\,ds$
Variable range	$0 < x < 1$ (continuous)
Parameters	$\alpha, \beta > 0$: shape parameters
Mean	$\mu_X = \dfrac{\alpha}{\alpha+\beta}$
Variance	$\sigma_X^2 = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Third raw moment	$m_X^{(3)} = \dfrac{\alpha(\alpha+1)(\alpha+2)}{(\alpha+\beta)(\alpha+\beta+1)(\alpha+\beta+2)}$
Coefficient of variation	$C_{v_X} = \sqrt{\dfrac{\beta}{\alpha(\alpha+\beta+1)}}$
Mode	$x_\mathrm{p} = \dfrac{\alpha-1}{\alpha+\beta-2}$ (for $\alpha, \beta > 1$)

5.3 Generalized Pareto distribution

The Pareto distribution was introduced by the Italian economist Vilfredo Pareto to describe the allocation of wealth among individuals since it seemed to describe well the fact that a larger portion of the wealth of a society is owned by a smaller percentage of the people. Its original form is expressed by the power-law equation

$$P\{X > x\} = \left(\frac{x}{\lambda}\right)^{-\frac{1}{\kappa}} \tag{6.31}$$

where λ is a (necessarily positive) minimum value of x ($x > \lambda$) and κ is a (positive) shape parameter. A generalized form, the so-called generalized Pareto distribution, in which a location parameter ζ independent of the scale parameter λ has been added, has been used in geophysics. Its basic characteristics are summarized in Table 6.10. Similar to the Log-Pearson III, the generalized Pareto distribution has a long tail. Indeed, as can be observed in Table 6.10, its third, second and first moments diverge (become infinite) for $\kappa \geq 1/3$, $\kappa \geq 1/2$ and $\kappa \geq 1$, respectively. For its long tail the distribution recently tends to replace short-tail distributions such as the Gamma distribution in modelling fine-time-scale rainfall and river discharge (Koutsoyiannis, 2004a,b, 2005). Since the analytical expression of the distribution function is very simple (Table 6.10) no tables or complicated numerical procedures are needed to handle it. Application of l'Hôpital's rule for $\kappa = 0$ results precisely in the exponential distribution, which thus can be derived as a special case of the Pareto distribution.

Table 6.10 Generalized Pareto distribution conspectus.

Probability density function	$f_X(x) = \frac{1}{\lambda}\left(1+\kappa\frac{x-\zeta}{\lambda}\right)^{-\frac{1}{\kappa}-1}$
Distribution function	$F_X(x) = 1-\left(1+\kappa\frac{x-\zeta}{\lambda}\right)^{-\frac{1}{\kappa}}$
Range	For $\kappa > 0$, $\zeta \le x < \infty$ For $\kappa < 0$, $\zeta \le x < \zeta - \lambda / \kappa$ (continuous)
Parameters	ζ: location parameter $\lambda > 0$: scale parameter κ: shape parameter
Mean	$\mu_X = \zeta + \frac{\lambda}{1-\kappa}$
Variance	$\sigma_X^2 = \frac{\lambda^2}{(1-\kappa)^2(1-2\kappa)}$
Third central moment	$\mu_X^{(3)} = \frac{2\lambda^3(1+\kappa)}{(1-\kappa)^3(1-2\kappa)(1-3\kappa)}$
Skewness coefficient	$C_{s_X} = \frac{2(1+\kappa)\sqrt{1-2\kappa}}{1-3\kappa}$
Mode	$x_{\mathrm{p}} = \zeta$
Median	$x_{0.5} = \zeta + \frac{\lambda}{\kappa}\left(1-0.5^{-\kappa}\right)$
Second L moment	$\lambda_X^{(2)} = \frac{\lambda}{(1-\kappa)(2-\kappa)}$
Third L moment	$\lambda_X^{(3)} = \frac{\lambda(1+\kappa)}{(1-\kappa)(2-\kappa)(3-\kappa)}$
Fourth L moment	$\lambda_X^{(4)} = \frac{\lambda(1+\kappa)(2+\kappa)}{(1-\kappa)(2-\kappa)(3-\kappa)(4-\kappa)}$
L coefficient of variation	$\tau_X^{(2)} = \frac{\lambda}{[\zeta(1-\kappa)+\lambda](2-\kappa)}$
L skewness	$\tau_X^{(3)} = \frac{1+\kappa}{3-\kappa}$
L kurtosis	$\tau_X^{(4)} = \frac{(1+\kappa)(2+\kappa)}{(3-\kappa)(4-\kappa)}$

5.4 Extreme value distributions

It can be easily shown that, given a number n of independent identically distributed random variables $Y_1,\ldots,Y_n$, the largest (in the sense of a specific realization) of them (more precisely, the largest order statistic), i.e.:

$$X_n = \max(Y_1, \ldots, Y_n) \tag{6.32}$$

has probability distribution function:

$$H_n(x) = [F(x)]^n \tag{6.33}$$

where $F(x) := P\{Y_i \le x\}$ is the common probability distribution function (referred to as the parent distribution) of each Y_i.

The evaluation of the exact distribution (6.33) requires the parent distribution to be known. For n tending to infinity, the limiting distribution $H(x) := H_\infty(x)$ becomes independent of $F(x)$. This has been utilised in several geophysical applications, thus trying to fit (justifiably or not) limiting extreme value distributions, or asymptotes, to extremes of various phenomena, and bypassing the study of the parent distribution. According to Gumbel (1958), as n tends to infinity, $H_n(x)$ converges to one of three possible asymptotes, depending on the mathematical form of $F(x)$. However, all three asymptotes can be described by a single mathematical expression, known as the generalized extreme value (GEV) distribution of maxima.

The logic behind the use of the extreme value distributions is this. Let us assume that the variable Y_i denotes the daily average discharge of a river of the day i. From (6.33), X_{365} will be then the maximum daily average discharge within a year. In practical problems of flood protection designs we are interested on the distribution of the variable X_{365} instead of that of Y_i. It is usually assumed that the distribution of X_{365} (the maximum of 365 variables) is well approximated by one of the asymptotes. Nevertheless, the strict conditions that make the theoretical extreme value distributions valid are rarely satisfied in real world processes. In the previous example the variables Y_i can neither be considered independent nor identically distributed. Moreover, the convergence to the asymptotic distribution in general is very slow, so that a good approximation may require that the maximum is taken over millions of variables (Koutsoyiannis, 2004a). For these reasons, the use of the asymptotic distributions should be done with attentiveness.

If are interested about minima, rather than maxima, i.e.:

$$X_n = \min(Y_1, \ldots, Y_n) \tag{6.34}$$

then the probability distribution function of X_n is:

$$G_n(x) = 1 - [1 - F(x)]^n \tag{6.35}$$

As n tends to infinity we obtain the generalized extreme value distribution of minima, a distribution symmetric to the generalized extreme value distribution of maxima.

These two generalized distributions and their special cases are analysed below. Nevertheless, several other distributions have been used in geophysics to describe extremes, e.g. the log-normal, the two and three-parameter Gamma and the log-Pearson III distributions.

5.4.1 Generalized extreme value distribution of maxima

The mathematical expression that comprises all three asymptotes is known as the generalized extreme value (GEV) distribution. Its basic characteristics are summarized in Table 6.11.

Table 6.11 Generalized extreme value distribution of maxima conspectus.

Probability density function	$f_X(x) = \frac{1}{\lambda}\left(1 + \kappa\frac{x-\zeta}{\lambda}\right)^{-1/\kappa-1} \exp\left[-\left(1 + \kappa\frac{x-\zeta}{\lambda}\right)^{-1/\kappa}\right]$
Distribution function	$F_X(x) = \exp\left[-\left(1 + \kappa\frac{x-\zeta}{\lambda}\right)^{-1/\kappa}\right]$
Range	In general: $\kappa x \geq \kappa\zeta - \lambda$ For $\kappa > 0$ (Extreme value of maxima type II): $\zeta - \lambda/\kappa \leq x < \infty$ For $\kappa < 0$ (Extreme value of maxima type III): $-\infty < x \leq \zeta - \lambda/\kappa$
Parameters	ζ: location parameter $\lambda > 0$: scale parameter κ: shape parameter
Mean	$\mu_X = \zeta - \frac{\lambda}{\kappa}[1 - \Gamma(1-\kappa)]$
Variance	$\sigma_X^2 = \left(\frac{\lambda}{\kappa}\right)^2 [\Gamma(1-2\kappa) - \Gamma^2(1-\kappa)]$
Third central moment	$\mu_X^{(3)} = \left(\frac{\lambda}{\kappa}\right)^3 [\Gamma(1-3\kappa) - 3\,\Gamma(1-2\kappa)\,\Gamma(1-\kappa) + 2\Gamma^3(1-\kappa)]$
Coefficient of skewness	$C_{s_X} = \mathrm{sgn}(\kappa)\frac{\Gamma(1-3\kappa) - 3\,\Gamma(1-2\kappa)\,\Gamma(1-\kappa) + 2\,\Gamma^3(1-\kappa)}{[\Gamma(1-2\kappa) - \Gamma^2(1-\kappa)]^{3/2}}$
Second L moment	$\lambda_X^{(2)} = -\Gamma(-\kappa)\,(2^\kappa - 1)\,\lambda$
Third L moment	$\lambda_X^{(3)} = -\Gamma(-\kappa)\,[2(3^\kappa - 1) - 3(2^\kappa - 1)]\,\lambda$
Fourth L moment	$\lambda_X^{(4)} = -\Gamma(-\kappa)\,[5(4^\kappa - 1) - 10(3^\kappa - 1) + 6(2^\kappa - 1)]\,\lambda$
L Coefficient of variation	$\tau_X^{(2)} = \frac{\Gamma(1-\kappa)\,(2^\kappa - 1)\,\lambda}{\lambda\,\Gamma(1-\kappa) + \zeta\kappa - \lambda}$
L Skewness	$\tau_X^{(3)} = 2\,\frac{3^\kappa - 1}{2^\kappa - 1} - 3$
L Kurtosis	$\tau_X^{(4)} = 6 + \frac{5(4^\kappa - 1) - 10(3^\kappa - 1)}{2^\kappa - 1}$

The shape parameter κ determines the general behaviour of the GEV distribution. For $\kappa > 0$ the distribution is bounded from below, has long right tail, and is known as the type II extreme value distribution of maxima or the Fréchet distribution. For $\kappa < 0$ it is bounded from above and is known as the type III extreme value distribution of maxima; this is not of practical interest in most real world problems because a bound from above is unrealistic. The limiting case where $\kappa = 0$, derived by application of l'Hôpital's rule, corresponds to the so-called extreme value distribution of type I or the Gumbel distribution (see section 5.4.2), which is unbounded both from above and below.

Typical calculations

The simplicity of the mathematical expression of the distribution function, permits typical calculations to be made directly without the need of tables or numerical approximations. The value of the distribution function can be calculated if the variable value is known. Also, the inverse distribution function has an analytical expression, namely the *u*-quantile of the distribution is

$$x_u = \zeta + \frac{\lambda\left[(-\ln u)^{-\kappa} - 1\right]}{\kappa} \tag{6.36}$$

Parameter estimation

As shown in Table 6.11, both coefficients of skewness and L skewness are functions of the shape parameter κ only, which enables the estimation of κ from either of the two expressions using the samples estimates of these coefficients. However the expressions are complicated and need to be solved numerically. Instead, the following explicit equations (Koutsoyiannis, 2004b) can be used, which are approximations of the exact (but implicit) equations of Table 6.11:

$$\kappa = \frac{1}{3} - \frac{1}{0.31 + 0.91\hat{C}_{s_X} + \sqrt{(0.91\,\hat{C}_{s_X})^2 + 1.8}} \tag{6.37}$$

$$\kappa = 8c - 3c^2, \;\; c := \frac{\ln 2}{\ln 3} - \frac{2}{3 + \hat{\tau}_X^{(3)}} \tag{6.38}$$

The former corresponds to the method of moments and the resulting error is smaller than ±0.01 for $-1 < \kappa < 1/3$ ($-2 < C_{s_X} < \infty$). The latter corresponds to the method of L moments and the resulting error is smaller than ±0.008 for $-1 < \kappa < 1$ ($-1/3 < \tau_X^{(3)} < 1$).

Once the shape parameter is calculated, the estimation of the remaining two parameters becomes very simple. The scale parameter can be estimated by the method of moments from:

$$\lambda = c_1 s_X, \;\; c_1 = |\kappa| \,/\, \sqrt{\Gamma(1 - 2\kappa) - \Gamma^2(1 - \kappa)} \tag{6.39}$$

or by the method of L moments from:

$$\lambda = c_2\, l_X^{(2)}, \;\; c_2 = \kappa/[\Gamma(1 - \kappa)(2^\kappa - 1)] \tag{6.40}$$

The estimate of the location parameter for both the method of moments and L moments is:

$$\zeta = \bar{x} - c_3\,\lambda, \;\; c_3 = [\Gamma(1 - \kappa) - 1]/\kappa \tag{6.41}$$

5.4.2 Extreme value distribution of maxima of type I (Gumbel)

As we have explained in the previous section, the type I or the Gumbel distribution is a special case of the generalized extreme value distribution of maxima for $\kappa = 0$. Its basic characteristics are summarized in Table 6.12, where the constant γ_{E} that appears in some equations is the Euler* constant.

Table 6.12 Type I or Gumbel distribution of maxima conspectus.

Probability density function	$f_X(x) = \frac{1}{\lambda}\exp\left(-\frac{x-\zeta}{\lambda}\right)\exp\left[-\exp\left(-\frac{x-\zeta}{\lambda}\right)\right]$
Distribution function	$F_X(x) = \exp\left[-\exp\left(-\frac{x-\zeta}{\lambda}\right)\right]$
Range	$-\infty < x < \infty$ (continuous)
Parameters	ζ: location parameter $\lambda > 0$: scale parameter
Mean	$\mu_X = \zeta + \gamma_{\mathrm{E}}\lambda = \zeta + 0.5772\lambda$
Variance	$\sigma_X^2 = \frac{\pi^2}{6}\lambda^2 = 1.645\lambda^2$
Third central moment	$\mu_X^{(3)} = 2.404\lambda^3$
Fourth central moment	$\mu_X^{(4)} = 14.6\lambda^4$
Coefficient of skewness	$C_{s_X} = 1.1396$
Coefficient of kurtosis	$C_{k_X} = 5.4$
Mode	$x_{\mathrm{p}} = \zeta$
Median	$x_{0.5} = \zeta - \lambda\ln(-\ln 0.5) = \zeta + 0.3665\lambda$
Second L moment	$\lambda_X^{(2)} = \lambda \ln 2$
Third L moment	$\lambda_X^{(3)} = (2\ln 3 - 3\ln 2)\,\lambda$
Fourth L moment	$\lambda_X^{(4)} = 2(8\ln 2 - 5\ln 3)\,\lambda$
L coefficient of variation	$\tau_X^{(2)} = \frac{\ln 2\,\lambda}{\zeta + \gamma_{\mathrm{E}}\,\lambda}$
L skewness	$\tau_X^{(3)} = 2\frac{\ln 3}{\ln 2} - 3 \approx 0.1699$
L kurtosis	$\tau_X^{(4)} = 16 - 10\frac{\ln 3}{\ln 2} \approx 0.1504$

[*] The Euler constant is defined as the limit

$$\gamma_{\mathrm{E}} := \lim_{n\to\infty}\left(1 + \frac{1}{2} + \dots + \frac{1}{n} - \ln n\right) \approx 0.5772156649\ \dots$$

Typical calculations

Due to the simplicity of the mathematical expression of the distribution function, typical calculations can be done explicitly without the need of tables or numerical approximations. The value of the distribution function can be calculated easily if the value of the variable is known. Moreover, the inverse distribution function has an analytical expression, namely the *u*-quantile of the distribution is

$$x_u = \zeta - \lambda \ln(-\ln u) \tag{6.42}$$

Parameter estimation

Since the Gumbel distribution is a special case of the GEV distribution, the parameter estimation procedures of the latter can be applied also in this case (except for the estimation of κ which by definition is zero). Specifically equations (6.39)-(6.41) for the method of moments and L moments still hold, and the constants c_i have the following values: $c_1 = \sqrt{6}/\pi = 0.78$, $c_2 = 1/\ln 2 = 1.443$ and $c_3 = \gamma_E = 0.577$.

Another method that results in similar expressions is the Gumbel method (Gumbel, 1958, p. 227). The method is based in the least square fit of the theoretical distribution function to the empirical distribution. For the empirical distribution function the Weibull plotting position must be used. The expressions of this method depend on the sample size *n*. The original Gumbel method is based on tabulated constants. To avoid the use of tables we give the following expressions that are good approximations of the original method:

$$\lambda = \frac{s_X}{\frac{1}{0.78} - \frac{1.57}{(n+1)^{0.65}}}, \quad \zeta = \bar{x} - \left[0.577 - \frac{0.53}{(n+2.5)^{0.74}}\right]\lambda \tag{6.43}$$

The approximation error is smaller than 0.25% for the former equation and smaller than 0.10% for the latter (for $n \geq 10$). For small exceedence probabilities, the Gumbel method results in safer predictions in comparison to the method of moments. The maximum likelihood method is more complicated; the interested reader may consult Kite (1988, p. 96).

Standard error and confidence intervals of quantiles

If the method of moments is used to estimate the parameters, then the point estimate of the *u*-quantile can be written in the following form that is equivalent to (6.42):

$$\hat{x}_u = \bar{x} - 0.5772\lambda - \lambda \ln(-\ln u) = \bar{x} + k_u s_X \tag{6.44}$$

where,

$$k_u = \lambda \frac{-0.5772 - \ln(-\ln u)}{s_X} = -0.45 - 0.78 \ln(-\ln u) \tag{6.45}$$

In this case it can be shown (Gumbel, 1958, p. 228· Kite, 1988, p. 103) that the square of the standard error of the estimate is

$$\varepsilon_X^2 = \mathrm{Var}(\hat{X}_u) = \frac{s_X^2}{n}\left(1 + 1.1396k_u + 1.1k_u^2\right) \tag{6.46}$$

Consequently, the confidence intervals of the u-quantile for confidence coefficient γ is approximately

$$\hat{x}_{u,2} = \left(\bar{x} + k_u s_X\right) \pm z_{(1+\gamma)/2}\frac{s_X}{\sqrt{n}}\sqrt{1 + 1.1396k_u + 1.1k_u^2} \tag{6.47}$$

Gumbel probability plot

The Gumbel distribution can be depicted as a straight line on a Gumbel probability plot. This plot can be easily constructed with horizontal probability axis $h = -\ln(-\ln F)$ (sometimes called Gumbel reduced variate) and vertical axis the variable of interest. Clearly, equation (6.42) is a straight line in this probability plot.

Numerical example

Table 6.13 lists a sample of the annual maximum daily discharge of the Evinos river upstream of the hydrometric station of Poros Reganiou. We wish to fit the Gumbel distribution of maxima and to determine the 100-year maximum discharge.

Table 6.13 Sample of annual maximum daily discharge (in m^3/s) of the river Evinos upstream of the hydrometric station of Poros Reganiou.

Hydrological year	Maximum discharge	Hydrological year	Maximum discharge	Hydrological year	Maximum discharge
1970-71	884	1977-78	365	1984-85	317
1971-72	305	1978-79	502	1985-86	374
1972-73	215	1979-80	381	1986-87	188
1973-74	378	1980-81	387	1987-88	192
1974-75	176	1981-82	525	1988-89	448
1975-76	430	1982-83	412	1989-90	70
1976-77	713	1983-84	439		

The sample average is

$$\bar{x} = \sum x / n = 385.1 \text{ m}^3\text{/s}$$

The standard deviation is

$$s_X = \sqrt{\sum x^2 / n - \bar{x}^2} = 181.5 \text{ m}^3\text{/s}$$

and the coefficient of variation is

$$\hat{C}_{\mathrm{v}_X} = s_X / \bar{x} = 181.5 / 385.1 = 0.47$$

The skewness coefficient is

$$\hat{C}_{\mathrm{s}_X} = 0.94$$

a value close to the theoretical value of the Gumbel distribution (1.14).

The method of moments results in

$$\lambda = 0.78 \times 181.5 = 141.57 \text{ m}^3\text{/s}, \zeta = 385.1 - 0.577 \times 141.57 = 303.4 \text{ m}^3\text{/s}$$

The maximum daily discharge for $T = 100$, or equivalently for $u = 1 - 1/100 = 0.99$, is

$$x_{0.99} = 303.4 - 141.57 \times \ln[-\ln(0.99)] = 955.0 \text{ m}^3\text{/s}$$

Based on (6.47), for

$$k_u = (955.0 - 385.1) / 181.5 = 3.16, z_{(1+\gamma)/2} = 1.96$$

we determine the 95% confidence intervals of the 100-year maximum daily discharge:

$$\hat{x}_{u1,2} \approx 955.0 \pm 1.96 \times \frac{181.5}{\sqrt{20}} \sqrt{1 + 1.1396 \times 3.16 + 1.1 \times 3.16^2}$$

$$\approx 955.0 \pm 313.1 = \begin{cases} 1268.1 \text{ m}^3\text{/s} \\ 641.9 \text{ m}^3\text{/s} \end{cases}$$

The Gumbel method using the equations (6.43), for $n = 20$, gives

$$\lambda = 170.36 \text{ m}^3\text{/s}, \zeta = 295.7 \text{ m}^3\text{/s}$$

and the 100-year maximum discharge estimation is

$$x_{0.99} = 295.7 - 170.36 \times \ln[-\ln(0.99)] = 1079.4 \text{ m}^3\text{/s}$$

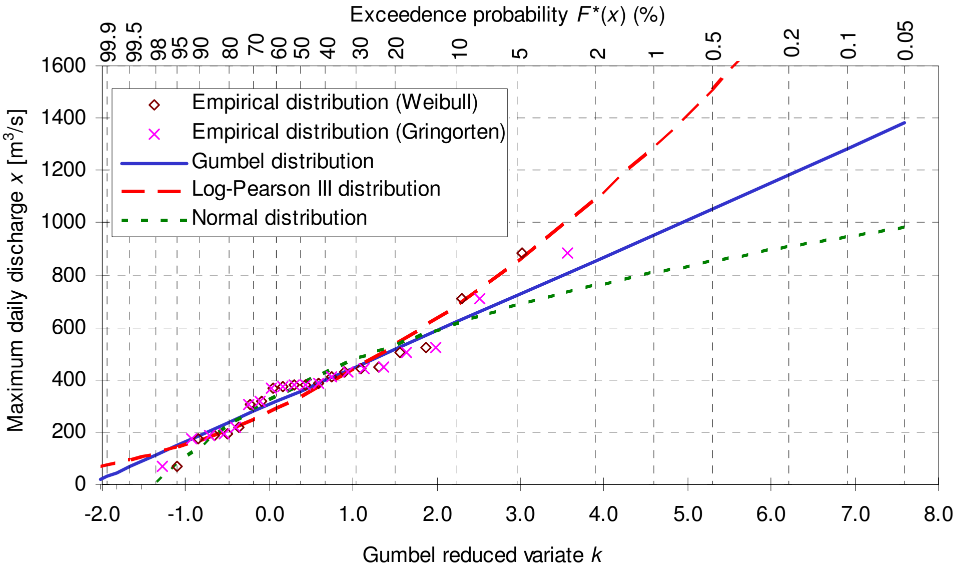

Fig. 6.3 Empirical and theoretical distribution of the daily maximum discharge of the river Evinos at station of Poros Riganiou plotted in Gumbel of maxima probability paper.

Fig. 6.3 depicts a comparison of the empirical distribution function and the theoretical Gumbel distribution of maxima on a Gumbel probability plot. For the empirical distribution function we have used the Weibull and the Gringorten plotting positions. For comparison we have also plotted the normal and log-Pearson III distributions. Clearly, the normal distribution is inappropriate (as expected) but even the Gumbel distribution does not fit well in the area of small exceedence probabilities that are of more interest, and seems to underestimate the highest discharges. The log-Person III distribution seems to be the most appropriate for the highest values of discharge. This seems to be a general problem for the Gumbel distribution. For more than have a century it has been the prevailing model for quantifying risk associated with extreme geophysical events. Newer evidence and theoretical studies (Koutsoyiannis, 2004a,b, 2005) have shown that the Gumbel distribution is quite unlikely to apply to hydrological extremes and its application may misjudge the risk, as it underestimates seriously the largest extremes. Besides, it has been shown that observed samples of typical length (like the one of this example) may display a distorted picture of the actual distribution, suggesting that the Gumbel distribution is an appropriate model for geophysical extremes while it is not. Therefore, it can be recommended to avoid the Gumbel distribution for the description of extreme rainfall and river discharge and use long-tail distributions such as the extreme value distribution of type II or log-Pearson III.

5.4.3 Generalized extreme value distribution of minima

If $H(x)$ is the generalized extreme value distribution of maxima then the distribution function $G(x) = 1 - H(-x)$ is the generalized extreme value distribution of minima. Its general characteristics are summarized in Table 6.14, where we have changed the sign convention in the parameter κ so that the distribution be unbounded from above for $\kappa > 0$ (bounded from below). This is similar to the generalized extreme value distribution of maxima where again $\kappa > 0$ corresponds to a distribution be unbounded from above. However, they are termed, respectively, type II extreme value distribution of maxima and type III extreme value distribution of minima (or the Weibull distribution). For $\kappa < 0$ the distribution of minima (similar to that of maxima) is bounded from above and is known as the type II extreme value distribution of minima; this is not of practical interest as in most real world problems a bound from above is unrealistic. The limiting case where $\kappa = 0$, derived by application of l'Hôpital's rule, corresponds to the so-called type I extreme value distribution of minima or the Gumbel distribution of minima, which is unbounded both from above and below.

Typical calculations

The mathematical expression of the generalized extreme value distribution of minima is similar to that of maxima. Thus, typical calculations can be done explicitly. The value of the distribution function can be calculated directly from the value of the variable. Also, the inverse distribution function has an analytical expression, namely the u-quantile of the distribution is given by

$$x_u = \zeta + \frac{\lambda}{\kappa}\{\, [-\ln (1 - u)]^{\kappa} - 1\} \tag{6.48}$$

Table 6.14 Generalized extreme value distribution of minima conspectus.

Probability density function	$f_X(x) = \frac{1}{\lambda}\left[1 + \kappa\left(\frac{x-\zeta}{\lambda}\right)\right]^{1/\kappa-1} \exp\left\{-\left[1 + \kappa\left(\frac{x-\zeta}{\lambda}\right)\right]^{1/\kappa}\right\}$
Distribution function	$F_X(x) = 1 - \exp\left[-\left(1 + \kappa\frac{x-\zeta}{\lambda}\right)^{1/\kappa}\right]$
Range	In general: $\kappa x \ge \kappa \zeta - \lambda$ For $\kappa > 0$ (Extreme value of minima type III): $\zeta - \lambda / \kappa \le x < \infty$ For $\kappa < 0$ (Extreme value of minima type II): $-\infty < x \le \zeta - \lambda / \kappa$
Parameters	ζ: location parameter $\lambda > 0$: scale parameter κ: shape parameter
Mean	$\mu_X = \zeta + \frac{\lambda}{\kappa}[\Gamma(1+\kappa) - 1]$
Variance	$\sigma_X^2 = \left(\frac{\lambda}{\kappa}\right)^2 [\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)]$
Third central moment	$\mu_X^{(3)} = \left(\frac{\lambda}{\kappa}\right)^3 [\Gamma(1+3\kappa) - 3\,\Gamma(1+2\kappa)\,\Gamma(1+\kappa) + 2\Gamma^3(1+\kappa)]$
Coefficient of skewness	$C_{s_X} = \mathrm{sgn}(\kappa)\frac{\Gamma(1+3\kappa) - 3\,\Gamma(1+2\kappa)\,\Gamma(1+\kappa) + 2\,\Gamma^3(1+\kappa)}{[\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)]^{3/2}}$
Second L moment	$\lambda_X^{(2)} = \Gamma(\kappa)\,(1 - 2^{-\kappa})\,\lambda$
Third L moment	$\lambda_X^{(3)} = \Gamma(\kappa)\,[3(1 - 2^{-\kappa}) - 2(1 - 3^{-\kappa})]\,\lambda$
Fourth L moment	$\lambda_X^{(4)} = \Gamma(\kappa)\,[5(1 - 4^{-\kappa}) - 10(1 - 3^{-\kappa}) + 6(1 - 2^{-\kappa})]\,\lambda$
L coefficient of variation	$\tau_X^{(2)} = \frac{\Gamma(1+\kappa)\,(1 - 2^{-\kappa})\,\lambda}{\lambda\,\Gamma(1+\kappa) + \zeta\kappa - \lambda}$
L skewness	$\tau_X^{(3)} = 3 - 2\,\frac{1 - 3^{-\kappa}}{1 - 2^{-\kappa}}$
L kurtosis	$\tau_X^{(4)} = 6 + \frac{5(1 - 4^{-\kappa}) - 10(1 - 3^{-\kappa})}{1 - 2^{-\kappa}}$

Parameter estimation

As shown in Table 6.14, both coefficients of skewness and L skewness are functions of the shape parameter κ only, which enables the estimation of κ from either of the two expressions using the sample estimates of these coefficients. However the expressions are complicated and need to be solved numerically. Instead, the following explicit equations (Koutsoyiannis, 2004b) can be used, which are approximations of the exact (but implicit) equations of Table 6.11:

$$\kappa = \frac{1}{0.28 - 0.9\hat{C}_{s_X} + 0.998\sqrt{(0.9\,\hat{C}_{s_X})^2 + 1.93}} - \frac{1}{3} \tag{6.49}$$

$$\kappa = 7.8c + 4.71\, c^2,\ \ c := \frac{2}{3 - \hat{\tau}_X^{(3)}} - \frac{\ln 2}{\ln 3} \tag{6.50}$$

The former corresponds to the method of moments and the resulting error is smaller than ±0.01 for $-1/3 < \kappa < 3$ ($-\infty < C_s < 20$). The latter corresponds to the method of L moments and the resulting error is even smaller.

Once the shape parameter is known, the scale parameter can be estimated by the method of moments from:

$$\lambda = c_1 s_X,\ \ c_1 = |\kappa| \,/\, \sqrt{\Gamma(1 + 2\kappa) - \Gamma^2(1 + \kappa)} \tag{6.51}$$

or by the method of L moments from:

$$\lambda = c_2\, \hat{l}_X^{(2)},\ \ c_2 = \kappa/[\Gamma(1 - \kappa)(2^\kappa - 1)] \tag{6.52}$$

The estimate of the location parameter for both the method of moments and L moments is:

$$\zeta = \bar{x} + c_3 \lambda,\ \ c_3 = [1 - \Gamma(1 + \kappa)]/\kappa \tag{6.53}$$

5.4.4 Extreme value distribution minima of type I (Gumbel)

As shown in Table 6.15, the type I distribution of minima resembles the type I distribution of maxima. The typical calculations are also similar. The inverse distribution function has an analytical expression and thus the u-quantile is given by:

$$x_u = \zeta + \lambda \ln\,[-\ln(1 - u)] \tag{6.54}$$

Since the Gumbel distribution is a special case of the GEV distribution, the parameter estimation procedures of the latter is based on equations (6.51)-(6.53) but with constants c_i as follows: $c_1 = \sqrt{6}/\pi = 0.78$, $c_2 = 1/\ln 2 = 1.443$ and $c_3 = \gamma_{\mathrm{E}} = 0.577$.

We can plot the Gumbel distribution of minima on a Gumbel-of-maxima probability paper if we replace the probability of exceedence with the probability of non-exceedence. Further, we can construct a Gumbel-of-minima probability plot if we use as horizontal axis the variate $h = \ln[-\ln\,(1-F)]$.

5.4.5 Two-parameter Weibull distribution

If in the generalized extreme value distribution of minima we assume that the lower bound ($\zeta - \lambda/\kappa$) is zero, we obtain the special case known as the two two-parameter Weibull distribution. Its main characteristics are shown in Table 6.15, where for convenience we have performed a change of the scale parameter replacing λ/κ with α.

Typical calculations

The related calculations are simple as in all previous cases and the inverse distribution function, from which quantiles are estimated, is

$$x_u = \alpha\{\,[-\ln\,(1 - u)]^{\,\kappa}\} \tag{6.55}$$

Table 6.15 Type I or Gumbel distribution of minima conspectus.

Probability density function	$f_X(x) = \frac{1}{\lambda} \exp\left(\frac{x-\zeta}{\lambda}\right) \exp\left[-\exp\left(\frac{x-\zeta}{\lambda}\right)\right]$
Distribution function	$F_X(x) = 1 - \exp\left[-\exp\left(\frac{x-\zeta}{\lambda}\right)\right]$
Variable range	$-\infty < x < \infty$ (continuous)
Parameters	ζ: location parameter $\lambda > 0$: scale parameter
Mean	$\mu_X = \zeta - \gamma_E \lambda = \zeta - 0.5772\lambda$
Variance	$\sigma_X^2 = \frac{\pi^2}{6}\lambda^2 = 1.645\lambda^2$
Third central moment	$\mu_X^{(3)} = -2.404\lambda^3$
Fourth central moment	$\mu_X^{(4)} = 14.6\lambda^4$
Skewness coefficient	$C_{s_X} = -1.1396$
Kurtosis coefficient	$C_{k_X} = 5.4$
Mode	$x_p = \zeta$
Median	$x_{0.5} = \zeta + \lambda \ln(-\ln 0.5) = \zeta - 0.3665\lambda$
Second L moment	$\lambda_X^{(2)} = \lambda \ln 2$
Third L moment	$\lambda_X^{(3)} = -(2 \ln 3 - 3 \ln 2)\, \lambda$
Fourth L moment	$\lambda_X^{(4)} = 2(8 \ln 2 - 5 \ln 3)\, \lambda$
L coefficient of variation	$\tau_X^{(2)} = \frac{\ln 2\, \lambda}{\zeta - \gamma_E\, \lambda}$
L skewness	$\tau_X^{(3)} = -2\frac{\ln 3}{\ln 2} + 3 \approx -0.1699$
L kurtosis	$\tau_X^{(4)} = 16 - 10\frac{\ln 3}{\ln 2} \approx 0.1504$

Parameter estimation

From the expressions of Table 6.14, the estimate of κ by the method of moments can be done from:

$$\frac{\Gamma(1 + 2\,\kappa)}{\Gamma^2(1 + \kappa)} = \hat{C}_{v_X}^{2} + 1 \tag{6.56}$$

This is implicit for κ and can be solved only numerically. An approximate solution with accuracy ±0.01 για $0 < \kappa < 3.2$ or $0 < C_{v_X} < 5$) is

$$\kappa = 2.56\ \left\{\exp\{0.41\ [\ln(C_v^2 + 1)]^{0.58}\} - 1\right\} \tag{6.57}$$

The L moment estimate is much simpler:

$$\kappa = \frac{-\ln(1 - \tau_X^{(2)})}{\ln 2} \tag{6.58}$$

Once κ has been estimated, the scale parameter for both the method of moments and L moments is

$$\alpha = \frac{\bar{x}}{\Gamma(1+\kappa)} \tag{6.59}$$

Table 6.16 Two-parameter Weibull distribution (type III of minima) conspectus.

Probability density function	$f(x) = \dfrac{1}{\kappa\,\alpha}\left(\dfrac{x}{\alpha}\right)^{1/\kappa - 1} \exp\left[-\left(\dfrac{x}{\alpha}\right)^{1/\kappa}\right]$
Distribution function	$F(x) = 1 - \exp\left[-\left(\dfrac{x}{\alpha}\right)^{1/\kappa}\right]$
Range	$0 < x < \infty$ (continuous)
Parameters	$\alpha > 0$: scale parameter $\kappa > 0$: shape parameter
Mean	$\mu_X = \alpha\Gamma(1+\kappa)$
Variance	$\sigma_X^2 = \alpha^2[\Gamma(1 + 2\kappa) - \Gamma^2(1 + \kappa)]$
Third central moment	$\mu_X^{(3)} = \alpha^3[\Gamma(1 + 3\kappa) - 3\Gamma(1 + 2\kappa)\,\Gamma(1 + \kappa) + 2\Gamma^3(1 + \kappa)]$
Coefficient of variation	$C_{\mathrm{v}X} = \dfrac{[\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)]^{1/2}}{\Gamma(1+\kappa)}$
Coefficient of skewness	$C_{\mathrm{s}X} = \dfrac{\Gamma(1+3\kappa) - 3\,\Gamma(1+2\kappa)\,\Gamma(1+\kappa) + 2\,\Gamma^3(1+\kappa)}{[\Gamma(1+2\kappa) - \Gamma^2(1+\kappa)]^{3/2}}$
Mode	$x_{\mathrm{p}} = \alpha(1-\kappa)^{\kappa}$ (for $\kappa > 1$)
Median	$x_{0.5} = \alpha(\ln 2)^{\kappa}$
Second L moment	$\lambda_X^{(2)} = \Gamma(1 + \kappa)\,(1 - 2^{-\kappa})\,\alpha$
Third L moment	$\lambda_X^{(3)} = \Gamma(1 + \kappa)\,[3(1 - 2^{-\kappa}) - 2(1 - 3^{-\kappa})]\,\alpha$
Fourth L moment	$\lambda_X^{(4)} = \Gamma(1 + \kappa)\,[5(1 - 4^{-\kappa}) - 10(1 - 3^{-\kappa}) + 6(1 - 2^{-\kappa})]\,\alpha$
L coefficient of variation	$\tau_X^{(2)} = 1 - 2^{-\kappa}$
L skewness	$\tau_X^{(3)} = 3 - 2\,\dfrac{1 - 3^{-\kappa}}{1 - 2^{-\kappa}}$
L kurtosis	$\tau_X^{(4)} = 6 + \dfrac{5(1 - 4^{-\kappa}) - 10(1 - 3^{-\kappa})}{1 - 2^{-\kappa}}$

We observe that the transformation $Z = \ln X$ results in

$$F_z(z) = 1 - \exp[-\mathrm{e}^{(z - \ln\alpha)/\kappa}] \tag{6.60}$$

which is a Gumbel distribution of minima with location parameter ln α and scale parameter κ. Thus, we can also use the parameter estimation methods of the Gumbel distribution applied on the logarithms of the observed sample values.

Weibull probability plot

A probability plot where the two-parameter Weibull distribution is depicted as a straight line is possible. The horizontal axis is $h = \ln[-\ln(1-F)]$ (similar to the plot of Gumbel of minima) and the vertical axis is $v = \ln x$ (logarithmic scale).

Numerical example

Table 6.17 lists a sample of annual minimum (average) daily discharge of the Evinos river upstream of the hydrometric station of Poros Reganiou. We wish to fit the Gumbel distribution of minima and the Weibull distribution and to determine the minimum 20-year discharge.

Table 6.17 Sample of annual minimum daily discharges (in m^3/s) of the river Evinos at the station of Poros Riganiou.

Hydrological year	Minimum. discharge	Hydrological year	Minimum discharge	Hydrological year	Minimum. discharge
1970-71	0.00	1977-78	2.14	1984-85	0.54
1971-72	2.19	1978-79	2.00	1985-86	0.54
1972-73	2.66	1979-80	1.93	1986-87	1.70
1973-74	2.13	1980-81	2.29	1987-88	1.70
1974-75	1.28	1981-82	2.66	1988-89	0.32
1975-76	0.56	1982-83	2.87	1989-90	1.37
1976-77	0.13	1983-84	1.88		

The sample mean is

$$\overline{x} = \sum x / n = 1.545 \text{ m}^3/\text{s}$$

The standard deviation is

$$s_X = \sqrt{\sum x^2 / n - \overline{x}^2} = 0.878 \text{ m}^3/\text{s}$$

and the coefficient of variation is

$$\hat{C}_{\mathrm{v}_X} = s_X / \overline{x} = 0.878/1.545 = 0.568$$

The skewness coefficient is

$$\hat{C}_{\mathrm{s}_X} = -0.40$$

The negative value of the skewness coefficient is expected for a sample of minimum discharges.

For the Gumbel distribution, the method of moments yields

$$\lambda = 0.78 \times 0.878 = 0.685 \text{ m}^3\text{/s}, \zeta = 1.545 + 0.577 \times 0.685 = 1.940 \text{ m}^3\text{/s}$$

The minimum discharge for $T = 20$ years, or equivalently for $u = 1/20 = 0.05$, is

$$x_{0.05} = 1.940 + 0.685 \times \ln[-\ln(1 - 0.05)] = -0.09 \text{ m}^3\text{/s}$$

Apparently, a negative value of discharge is meaningless; we can consider that the minimum 20-year discharge is zero.

For the two-parameter Weibull distribution, application of (6.57) for the method of moments gives

$$\kappa = 2.56 \left\{\exp\{0.41 [\ln(0.568^2 + 1)]^{0.58}\} -1\right\} = 0.55$$

Hence

$$\alpha = \frac{1.545}{\Gamma(1+0.55)} = 1.740 \text{ m}^3\text{/s}$$

and the 20-year mminimum daily discharge is estimated at

$$x_{0.05} = 1.740 \{[-\ln(1-0.05)]^{0.55}\} = 0.340 \text{ m}^3\text{/s}$$

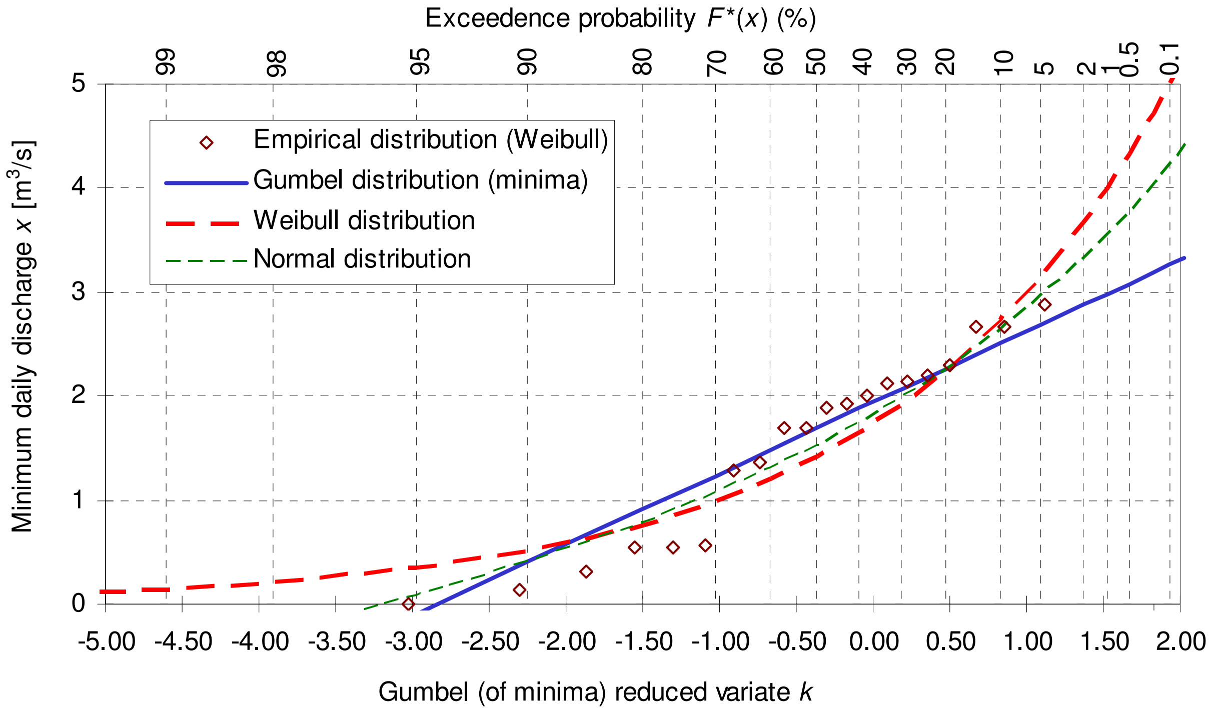

Fig. 6.4 Empirical and theoretical distribution function of the minimum daily discharge of the river Evinos at the station Poros Riganiou in Gumbel of minima probability paper.

Fig. 6.4 compares graphically the empirical distribution function with the two fitted theoretical distributions. For the empirical distribution we have used the Weibull plotting position. None the two theoretical distributions fits very well to the sample, but clearly the Gumbel distribution performs better, especially in the area of small exceedence probabilities.

The two-parameter Weibull distribution is defined for $x > 0$, which seems to be a theoretical advantage due to the consistency with nature. However, in practice it turns to be a disadvantage due to the departure of the empirical distribution for the lowest discharges. On the other hand, the Gumbel distribution of minima is theoretically inconsistent as it predicts negative values of discharge for high return periods. An *ad hoc* solution is to truncate the Gumbel distribution at zero, as we have done above. For comparison the normal distribution has been also plotted in Fig. 6.4 but we do not expect to be appropriate for this problem.

Acknowledgement I thank Simon Papalexiou for his help in translating some of my Greek texts into English and for his valuable suggestions.

References

Bobée, B., and F. Ashkar, *The Gamma Family and Derived Distributions Applied in Hydrology*, Water Resources Publications, Littleton, Colorado, 1991.

Gumbel, E. J., *Statistics of Extremes*, Columbia University Press, New York, 1958.

Kite, G. W., *Frequency and Risk Analyses in Hydrology*, Water Resources Publications, Littleton, Colorado, 1988.

Koutsoyiannis, D., *Statistical Hydrology*, Edition 4, 312 pages, National Technical University of Athens, Athens, 1997.

Koutsoyiannis, D., Statistics of extremes and estimation of extreme rainfall, 1, Theoretical investigation, *Hydrological Sciences Journal*, 49 (4), 575–590, 2004a.

Koutsoyiannis, D., Statistics of extremes and estimation of extreme rainfall, 2, Empirical investigation of long rainfall records, *Hydrological Sciences Journal*, 49 (4), 591–610, 2004b.

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

Koutsoyiannis, D., An entropic-stochastic representation of rainfall intermittency: The origin of clustering and persistence, *Water Resources Research*, 42 (1), W01401, 2006.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes*, Cambridge University Press, Cambridge, 1987.

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

Appendix

Table A1 Numerical values of the standard normal distribution.

z	$F(z)$	$F^*(z)$	z	$F(z)$	$F^*(z)$	z	$F(z)$	$F^*(z)$
0	0.5	0.5	1.35	0.91149	0.08851	2.70	0.99653	0.00347
0.05	0.51994	0.48006	1.40	0.91924	0.08076	2.75	0.99702	0.00298
0.10	0.53983	0.46017	1.45	0.92647	0.07353	2.80	0.99744	0.00256
0.15	0.55962	0.44038	1.50	0.93319	0.06681	2.85	0.99781	0.00219
0.20	0.57926	0.42074	1.55	0.93943	0.06057	2.8782	0.998	0.002
0.25	0.59871	0.40129	1.60	0.94520	0.05480	2.90	0.99813	0.00187
0.2533	0.6	0.4	1.6449	0.95	0.05	2.95	0.99841	0.00159
0.30	0.61791	0.38209	1.65	0.95053	0.04947	3.00	0.99865	0.00135
0.35	0.63683	0.36317	1.70	0.95543	0.04457	3.05	0.99841	0.00159
0.40	0.65542	0.34458	1.75	0.95994	0.04006	3.0902	0.999	0.001
0.45	0.67364	0.32636	1.80	0.96407	0.03593	3.10	0.99886	0.00114
0.50	0.69146	0.30854	1.85	0.96784	0.03216	3.15	0.99900	0.00100
0.5244	0.7	0.3	1.90	0.97128	0.02872	3.20	0.99903	0.00097
0.55	0.70884	0.29116	1.95	0.97441	0.02559	3.25	0.99918	0.00082
0.60	0.72575	0.27425	2.00	0.97725	0.02275	3.2905	0.9995	0.0005
0.65	0.74215	0.25785	2.05	0.97982	0.02018	3.30	0.99942	0.00058
0.70	0.75804	0.24196	2.0537	0.98	0.02	3.35	0.99950	0.00050
0.75	0.77337	0.22663	2.10	0.98214	0.01786	3.40	0.99952	0.00048
0.80	0.78814	0.21186	2.15	0.98422	0.01578	3.45	0.99960	0.00040
0.8416	0.8	0.2	2.20	0.98610	0.01390	3.50	0.99966	0.00034
0.85	0.80234	0.19766	2.25	0.98778	0.01222	3.5402	0.9998	0.0002
0.90	0.81594	0.18406	2.30	0.98928	0.01072	3.55	0.99977	0.00023
0.95	0.82894	0.17106	2.3263	0.99	0.01	3.60	0.99980	0.00020
1.00	0.84134	0.15866	2.35	0.99061	0.00939	3.65	0.99981	0.00019
1.05	0.85314	0.14686	2.40	0.99180	0.00820	3.70	0.99984	0.00016
1.10	0.86433	0.13567	2.45	0.99286	0.00714	3.7195	0.9999	10^{-4}
1.15	0.87493	0.12507	2.50	0.99379	0.00621	4.27	$1 - 10^{-5}$	10^{-5}
1.20	0.88493	0.11507	2.55	0.99461	0.00539	4.75	$1 - 10^{-6}$	10^{-6}
1.25	0.89435	0.10565	2.5758	0.995	0.005	5.20	$1 - 10^{-7}$	10^{-7}
1.2816	0.9	0.1	2.60	0.99534	0.00466	5.61	$1 - 10^{-8}$	10^{-8}
1.30	0.90320	0.09680	2.65	0.99598	0.00402	6.00	$1 - 10^{-9}$	10^{-9}
$-z$	$F^*(-z)$	$F(-z)$	$-z$	$F^*(-z)$	$F(-z)$	$-z$	$F^*(-z)$	$F(-z)$

Examples: $F(0.80) = 0.78814$ $F(-3.30) = 0.00058$

$z_{0.8} = 0.8416$ $z_{0.01} = -2.3263$

Table A2 Quantiles $\chi^2_u(n)$ of the χ^2 distribution for characteristic values of u and for n degrees of freedom.

u =	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
n=1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
35	17.19	18.51	20.57	22.47	24.80	46.06	49.80	53.20	57.34	60.27
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
45	24.31	25.90	28.37	30.61	33.35	57.51	61.66	65.41	69.96	73.17
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49

Examples: $\chi^2_{0.05}(5) = 1.15$ $\chi^2_{0.99}(10) = 23.21$

For $n \geq 50$: $\chi^2_u(n) = \frac{1}{2}\left(z_u + \sqrt{2n-1}\right)^2$

where z_u is the u-quantile of the standard normal distribution.

Table A3 Quantiles $t_u(n)$ of the t distribution for characteristic values of u and for n degrees of freedom.

u =	0.9	0.95	0.975	0.99	0.995
n =1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
22	1.32	1.72	2.07	2.51	2.82
24	1.32	1.71	2.06	2.49	2.80
26	1.31	1.71	2.06	2.48	2.78
28	1.31	1.70	2.05	2.47	2.76
30	1.31	1.70	2.04	2.46	2.75
35	1.31	1.69	2.03	2.44	2.72
40	1.30	1.68	2.02	2.42	2.70
45	1.30	1.68	2.01	2.41	2.69
50	1.30	1.68	2.01	2.40	2.68
∞	1.28	1.64	1.96	2.33	2.58

Example: $t_{0.95}(5) = 2.02$

For $n \geq 50$: $t_u(n) \approx z_u \sqrt{\dfrac{n}{n-2}}$

where z_u is the u-quantile of the standard normal distribution.

Table A4a Quantiles (k_u) of the standardized gamma distribution for characteristic values of the coefficient of skewness C_s (≤ 2) or the shape parameter κ (≥ 1).

$u=F$	$1-u=F^*$	C_s=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
		$\kappa=\infty$	400	100	44.44	25.00	16.00	11.11	8.163	6.250	4.938	4.000	3.306	2.778	2.367	2.041	1.778	1.563	1.384	1.235	1.108	1.000
0.0001	0.9999	-3.72	-3.51	-3.30	-3.10	-2.90	-2.71	-2.53	-2.35	-2.18	-2.03	-1.88	-1.75	-1.63	-1.52	-1.42	-1.33	-1.25	-1.18	-1.11	-1.05	-1.00
0.0002	0.9998	-3.54	-3.35	-3.16	-2.98	-2.80	-2.63	-2.46	-2.30	-2.14	-2.00	-1.86	-1.73	-1.62	-1.51	-1.41	-1.33	-1.25	-1.17	-1.11	-1.05	-1.00
0.0005	0.9995	-3.29	-3.13	-2.97	-2.81	-2.65	-2.50	-2.36	-2.21	-2.08	-1.95	-1.82	-1.71	-1.60	-1.50	-1.40	-1.32	-1.24	-1.17	-1.11	-1.05	-1.00
0.001	0.999	-3.09	-2.95	-2.81	-2.67	-2.53	-2.40	-2.27	-2.14	-2.02	-1.90	-1.79	-1.68	-1.58	-1.48	-1.39	-1.31	-1.24	-1.17	-1.11	-1.05	-1.00
0.002	0.998	-2.88	-2.76	-2.64	-2.52	-2.40	-2.28	-2.17	-2.06	-1.95	-1.84	-1.74	-1.64	-1.55	-1.46	-1.38	-1.30	-1.23	-1.17	-1.10	-1.05	-1.00
0.005	0.995	-2.58	-2.48	-2.39	-2.29	-2.20	-2.11	-2.02	-1.93	-1.84	-1.75	-1.66	-1.58	-1.50	-1.42	-1.35	-1.28	-1.22	-1.15	-1.10	-1.04	-0.99
0.01	0.99	-2.33	-2.25	-2.18	-2.10	-2.03	-1.95	-1.88	-1.81	-1.73	-1.66	-1.59	-1.52	-1.45	-1.38	-1.32	-1.26	-1.20	-1.14	-1.09	-1.04	-0.99
0.02	0.98	-2.05	-2.00	-1.94	-1.89	-1.83	-1.78	-1.72	-1.66	-1.61	-1.55	-1.49	-1.44	-1.38	-1.32	-1.27	-1.22	-1.17	-1.12	-1.07	-1.02	-0.98
0.05	0.95	-1.64	-1.62	-1.59	-1.56	-1.52	-1.49	-1.46	-1.42	-1.39	-1.35	-1.32	-1.28	-1.24	-1.21	-1.17	-1.13	-1.09	-1.06	-1.02	-0.98	-0.95
0.1	0.9	-1.28	-1.27	-1.26	-1.25	-1.23	-1.22	-1.20	-1.18	-1.17	-1.15	-1.13	-1.11	-1.09	-1.06	-1.04	-1.02	-0.99	-0.97	-0.94	-0.92	-0.89
0.2	0.8	-0.84	-0.85	-0.85	-0.85	-0.86	-0.86	-0.86	-0.86	-0.86	-0.85	-0.85	-0.85	-0.84	-0.84	-0.83	-0.83	-0.82	-0.81	-0.80	-0.79	-0.78
0.3	0.7	-0.52	-0.54	-0.55	-0.56	-0.57	-0.58	-0.59	-0.60	-0.60	-0.61	-0.62	-0.62	-0.63	-0.63	-0.64	-0.64	-0.64	-0.64	-0.64	-0.64	-0.64
0.4	0.6	-0.25	-0.27	-0.28	-0.30	-0.31	-0.33	-0.34	-0.36	-0.37	-0.38	-0.39	-0.41	-0.42	-0.43	-0.44	-0.45	-0.46	-0.47	-0.48	-0.48	-0.49
0.5	0.5	0.00	-0.02	-0.03	-0.05	-0.07	-0.08	-0.10	-0.12	-0.13	-0.15	-0.16	-0.18	-0.20	-0.21	-0.23	-0.24	-0.25	-0.27	-0.28	-0.29	-0.31
0.6	0.4	0.25	0.24	0.22	0.21	0.19	0.17	0.16	0.14	0.12	0.10	0.09	0.07	0.05	0.04	0.02	0.00	-0.02	-0.03	-0.05	-0.07	-0.08
0.7	0.3	0.52	0.51	0.50	0.49	0.47	0.46	0.44	0.43	0.41	0.40	0.38	0.36	0.35	0.33	0.31	0.30	0.28	0.26	0.24	0.22	0.20
0.8	0.2	0.84	0.84	0.83	0.82	0.82	0.81	0.80	0.79	0.78	0.77	0.76	0.75	0.73	0.72	0.71	0.69	0.68	0.66	0.64	0.63	0.61
0.9	0.1	1.28	1.29	1.30	1.31	1.32	1.32	1.33	1.33	1.34	1.34	1.34	1.34	1.34	1.34	1.34	1.33	1.33	1.32	1.32	1.31	1.30
0.95	0.05	1.64	1.67	1.70	1.73	1.75	1.77	1.80	1.82	1.84	1.86	1.88	1.89	1.91	1.92	1.94	1.95	1.96	1.97	1.98	1.99	2.00
0.98	0.02	2.05	2.11	2.16	2.21	2.26	2.31	2.36	2.41	2.45	2.50	2.54	2.58	2.63	2.67	2.71	2.74	2.78	2.81	2.85	2.88	2.91
0.99	0.01	2.33	2.40	2.47	2.54	2.62	2.69	2.76	2.82	2.89	2.96	3.02	3.09	3.15	3.21	3.27	3.33	3.39	3.44	3.50	3.55	3.61
0.995	0.005	2.58	2.67	2.76	2.86	2.95	3.04	3.13	3.22	3.31	3.40	3.49	3.58	3.66	3.74	3.83	3.91	3.99	4.07	4.15	4.22	4.30
0.998	0.002	2.88	3.00	3.12	3.24	3.37	3.49	3.61	3.73	3.85	3.97	4.09	4.21	4.32	4.44	4.55	4.67	4.78	4.89	5.00	5.11	5.21
0.999	0.001	3.09	3.23	3.38	3.52	3.67	3.81	3.96	4.10	4.24	4.39	4.53	4.67	4.81	4.96	5.10	5.23	5.37	5.51	5.64	5.78	5.91
0.9995	0.0005	3.29	3.46	3.62	3.79	3.96	4.12	4.29	4.46	4.63	4.80	4.97	5.13	5.30	5.47	5.63	5.80	5.96	6.12	6.28	6.44	6.60
0.9998	0.0002	3.54	3.73	3.93	4.13	4.33	4.53	4.73	4.93	5.13	5.33	5.53	5.74	5.94	6.14	6.34	6.54	6.73	6.93	7.13	7.32	7.52
0.9999	0.0001	3.72	3.93	4.15	4.37	4.60	4.82	5.05	5.27	5.50	5.73	5.96	6.18	6.41	6.64	6.87	7.09	7.32	7.54	7.77	7.99	8.21

Note: $k_u = (x - \mu_X) / \sigma_X \leftrightarrow x = \mu_X + \sigma_X k_u$ Example: For $C_s = 0.5$ ($\kappa = 16$): $k_{0.98} = 2.31$

Table A4b Quantiles (k_u) of the standardized gamma distribution for characteristic values of the coefficient of skewness C_s (≥ 2) or the shape parameter κ (≤ 1).

$u=F$	$1-u=F^*$	C_s=2	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
		$\kappa=1$	0.907	0.826	0.76	0.69	0.64	0.59	0.549	0.510	0.476	0.444	0.416	0.391	0.367	0.346	0.327	0.309	0.292	0.277	0.263	0.250
0.0001	0.9999	-1.00	-0.95	-0.91	-0.87	-0.83	-0.80	-0.77	-0.74	-0.71	-0.69	-0.67	-0.65	-0.62	-0.61	-0.59	-0.57	-0.56	-0.54	-0.53	-0.51	-0.50
0.0002	0.9998	-1.00	-0.95	-0.91	-0.87	-0.83	-0.80	-0.77	-0.74	-0.71	-0.69	-0.67	-0.65	-0.62	-0.61	-0.59	-0.57	-0.56	-0.54	-0.53	-0.51	-0.50
0.0005	0.9995	-1.00	-0.95	-0.91	-0.87	-0.83	-0.80	-0.77	-0.74	-0.71	-0.69	-0.67	-0.65	-0.62	-0.61	-0.59	-0.57	-0.56	-0.54	-0.53	-0.51	-0.50
0.001	0.999	-1.00	-0.95	-0.91	-0.87	-0.83	-0.80	-0.77	-0.74	-0.71	-0.69	-0.67	-0.65	-0.62	-0.61	-0.59	-0.57	-0.56	-0.54	-0.53	-0.51	-0.50
0.002	0.998	-1.00	-0.95	-0.91	-0.87	-0.83	-0.80	-0.77	-0.74	-0.71	-0.69	-0.67	-0.65	-0.62	-0.61	-0.59	-0.57	-0.56	-0.54	-0.53	-0.51	-0.50
0.005	0.995	-0.99	-0.95	-0.91	-0.87	-0.83	-0.80	-0.77	-0.74	-0.71	-0.69	-0.67	-0.65	-0.62	-0.61	-0.59	-0.57	-0.56	-0.54	-0.53	-0.51	-0.50
0.01	0.99	-0.99	-0.95	-0.91	-0.87	-0.83	-0.80	-0.77	-0.74	-0.71	-0.69	-0.67	-0.65	-0.62	-0.61	-0.59	-0.57	-0.56	-0.54	-0.53	-0.51	-0.50
0.02	0.98	-0.98	-0.94	-0.90	-0.86	-0.83	-0.80	-0.77	-0.74	-0.71	-0.69	-0.67	-0.65	-0.62	-0.61	-0.59	-0.57	-0.56	-0.54	-0.53	-0.51	-0.50
0.05	0.95	-0.95	-0.91	-0.88	-0.85	-0.82	-0.79	-0.76	-0.74	-0.71	-0.69	-0.67	-0.64	-0.62	-0.61	-0.59	-0.57	-0.56	-0.54	-0.53	-0.51	-0.50
0.1	0.9	-0.89	-0.87	-0.84	-0.82	-0.79	-0.77	-0.75	-0.72	-0.70	-0.68	-0.66	-0.64	-0.62	-0.60	-0.59	-0.57	-0.55	-0.54	-0.53	-0.51	-0.50
0.2	0.8	-0.78	-0.76	-0.75	-0.74	-0.72	-0.71	-0.70	-0.68	-0.67	-0.65	-0.64	-0.62	-0.61	-0.59	-0.58	-0.56	-0.55	-0.54	-0.52	-0.51	-0.50
0.3	0.7	-0.64	-0.64	-0.64	-0.63	-0.63	-0.62	-0.62	-0.61	-0.60	-0.60	-0.59	-0.58	-0.57	-0.56	-0.55	-0.54	-0.53	-0.52	-0.51	-0.50	-0.49
0.4	0.6	-0.49	-0.49	-0.50	-0.50	-0.51	-0.51	-0.51	-0.51	-0.51	-0.51	-0.51	-0.51	-0.51	-0.50	-0.50	-0.49	-0.49	-0.48	-0.48	-0.47	-0.46
0.5	0.5	-0.31	-0.32	-0.33	-0.34	-0.35	-0.36	-0.37	-0.38	-0.38	-0.39	-0.40	-0.40	-0.40	-0.41	-0.41	-0.41	-0.41	-0.41	-0.41	-0.41	-0.41
0.6	0.4	-0.08	-0.10	-0.12	-0.13	-0.15	-0.16	-0.18	-0.19	-0.20	-0.22	-0.23	-0.24	-0.25	-0.26	-0.27	-0.28	-0.29	-0.29	-0.30	-0.31	-0.31
0.7	0.3	0.20	0.19	0.17	0.15	0.13	0.11	0.09	0.08	0.06	0.04	0.02	0.01	-0.01	-0.03	-0.04	-0.06	-0.07	-0.09	-0.10	-0.11	-0.13
0.8	0.2	0.61	0.59	0.57	0.56	0.54	0.52	0.50	0.48	0.46	0.44	0.42	0.40	0.38	0.36	0.34	0.32	0.30	0.28	0.26	0.24	0.23
0.9	0.1	1.30	1.29	1.28	1.27	1.26	1.25	1.24	1.22	1.21	1.20	1.18	1.16	1.15	1.13	1.11	1.10	1.08	1.06	1.04	1.02	1.00
0.95	0.05	2.00	2.00	2.01	2.01	2.01	2.01	2.01	2.01	2.01	2.01	2.00	2.00	1.99	1.99	1.98	1.97	1.96	1.95	1.94	1.93	1.92
0.98	0.02	2.91	2.94	2.97	3.00	3.02	3.05	3.07	3.09	3.11	3.13	3.15	3.17	3.19	3.20	3.21	3.23	3.24	3.25	3.26	3.27	3.27
0.99	0.01	3.61	3.66	3.71	3.75	3.80	3.85	3.89	3.93	3.97	4.01	4.05	4.09	4.12	4.16	4.19	4.22	4.26	4.29	4.31	4.34	4.37
0.995	0.005	4.30	4.37	4.44	4.51	4.58	4.65	4.72	4.78	4.85	4.91	4.97	5.03	5.09	5.14	5.20	5.25	5.31	5.36	5.41	5.46	5.50
0.998	0.002	5.21	5.32	5.42	5.53	5.63	5.73	5.83	5.92	6.02	6.11	6.21	6.30	6.39	6.47	6.56	6.65	6.73	6.81	6.89	6.97	7.05
0.999	0.001	5.91	6.04	6.17	6.30	6.42	6.55	6.67	6.79	6.92	7.03	7.15	7.27	7.38	7.50	7.61	7.72	7.83	7.94	8.04	8.15	8.25
0.9995	0.0005	6.60	6.76	6.91	7.07	7.22	7.37	7.52	7.67	7.82	7.96	8.11	8.25	8.39	8.53	8.67	8.81	8.94	9.08	9.21	9.34	9.47
0.9998	0.0002	7.52	7.71	7.90	8.09	8.28	8.47	8.65	8.84	9.02	9.20	9.38	9.56	9.74	9.92	10.09	10.26	10.43	10.60	10.77	10.94	11.11
0.9999	0.0001	8.21	8.43	8.65	8.87	9.08	9.30	9.51	9.73	9.94	10.15	10.35	10.56	10.77	10.97	11.17	11.37	11.57	11.77	11.97	12.16	12.36

Note: $k_u = (x - \mu_X) / \sigma_X \leftrightarrow x = \mu_X + \sigma_X k_u$ Example: For $C_s = 2.5$ ($\kappa = 0.64$): $k_{0.98} = 3.05$