
Abstract

Trends are one of the deterministic parts of a given time series apart from the natural or artificial seasonality and uncertain components. Trend analysis is a search for deterministic trend in an uncertain environment, therefore, the basic concepts of uncertainty are explained as stochastic and completely random variables and their importance in trend identification studies. Since, probability and statistics are main subjects for such a search various probabilistic and statistical concepts are presented in an effective manner so that prior to a proper trend analysis the reader can appreciate the fundamental elementary concepts, which are in later chapters are employed for the main goal. In classical trend analyses, the most restrictive assumption requirement is the serial independence of given time series, various correlation measurement suggestions are reflected from the literature. In the meantime for classical trend analysis, the characteristics of a time series are explained for proper application of the methodologies.

Keywords

Correlation • Frequency • Histogram • Homogeneity • Seasonality • Stationarity • Uncertainty

2.1 General

Uncertainty has many connotations to common people and experts grasp it in rather different ways; some considers it as entirely unknown and unpredictable information and to some others, it is partial information and knowledge. The uncertainty is everywhere and one cannot get rid of it completely. Initial knowledge and information are concepts that depend on personal observations and experience.

Uncertainty can be avoided by a set of simplifying assumptions about the phenomenon concerned. For instance, Newtonian classical physics is entirely deterministic. Today almost all branches of science (environmental, atmospheric, earth, engineering, economics, health and social) are confronted with uncertainty ingredients and many scientific deterministic foundations take uncertainty forms in terms of random, probability, statistics, chaos, fractal, stochastic, quantum, and fuzzy implications. In many scientific and technological institutions determinism dominates the education systems. Famous philosophers and scientists spell out the uncertainty and fuzzy ingredients that are essential bases of scientific progress. For instance, Russell (1948) stated that

All traditional logic habitually assumes that precise symbols are being employed. It is, therefore, not applicable to this terrestrial life but only to an imagined celestial existence.

On the other hand, as for the verbal and linguistic fuzzy conceptions Zadeh (1965) said that

As the complexity of a system increases, our ability to make precise and yet significant statements about its behavior diminishes until a threshold is reached beyond which precision and significance (or relevance) become almost mutually exclusive characteristics.

During human thinking evolution, the premises include uncertainty elements such as vagueness, ambiguousness, possibility, probability, random, and fuzziness. Implication of mathematical structure from the mental thinking process might seem exact, but even today it is understood as a result of scientific development that at every stages of modeling, physical, or mechanical, there are uncertainty pieces, if not in the macroscale, at least at the microscale. It is clear today that mathematical conceptualization and idealization leading to satisfactory mathematical structure of any physical actuality is often an approximation, because as Popper (1954) states that scientific facts are falsifiable.

The word uncertainty reminds also the probability of occurrences with attachment of a certain percentage. It is not uncommon that everyone is confronted with probability statements, especially in natural, social, and financial conservations. For instance, what is the probability of weather status for tomorrow? what will be the income depending on the number of clients in the next month? what is the probability of investing on stock? These queries involve uncertainty and their daily answers are quantitatively through subjective percentage numbers, which are, the probability statements. Probability in the statistics context helps to investigate past records of uncertain events so as to make future predictions and dependable decisions.

Prior to the explanations of systematic components such as trends in any time series, it is preferable to equip the reader with uncertainty concepts. Uncertainty or randomness and stochasticity are counter concepts to determinism. For instance, astronomic events were thought as rather uncertain phenomena by early human beings, but today they are well known and even one is capable to calculate through the scientific methodologies the position of any planet at any future time.

Additionally, any gadget, instrument, automation or machine can be described by deterministic formulations, equations, and logical rules.

Many natural events in atmospheric, environmental, earth, oceanography, meteorology, hydrogeology, earthquake and social, economic, health, business, and similar events do not provide completely well-established behaviors, because they include some deterministic parts that can be identified by mathematical statistical methods, but the residuals from the deterministic components are random in character. Any natural phenomenon takes place under the combined effects of physical, chemical, mechanical, and thermodynamic conditions and evolves temporally and spatially according to a set of certain laws. These effects cannot be identified accurately whatever the monitoring instrumentation and scientific methodologies are. For instance, hydrological events such as rainfall, runoff, infiltration, and evaporation cannot be controlled precisely over large areas and future times. The uncertainties in these events affect the economic, environmental, social, and even physiological conditions of human societies. For instance, human civilization has long been deeply affected by impacts of droughts on economic, environmental and social sectors (Wilhite 1993).

On the other hand, the scientific models that are suitable for description of an event might not be reliable with high confidence. At this stage, it is very convenient to remember the statement by Einstein that

so far as the laws of mathematics refer to reality, they are not certain. And so far as they are certain, they do not refer to reality.

Natural event uncertainty is associated with not knowing if and/or when, say for instance, a rainfall event will cause to the exceedence of a given design discharge. Additionally, model uncertainty is the inaccuracy of the model used to estimate the design discharge. In addition to these two uncertainty types the third one is the measurement error.

The uncertainty in the earth and atmospheric systems arises from the conviction that generalizations are immensely complicated instantiations of abstract and often universal physical laws. Such generalizations always contain assumptions of boundary and initial conditions. The researchers cannot control these conditions with certainty.

Earth systems sciences deal with spatial and temporal structures (trends, periodicities, jumps) in natural phenomena at every scale for the purpose of predicting the future replicas of the similar phenomenon, which help to make significant decisions in planning, management, operation, and maintenance of natural events that are related to social, environmental, and engineering activities. These phenomena are sampled by measurements with uncertainty ingredient; their analysis, control, and prediction need to use uncertainty techniques for reliable predictions. Natural phenomena cannot be monitored at a set of desired instances and locations, and therefore, such restrictive time and location conditions bring additional irregularity into the measurements. For instance, floods, earthquakes, car accidents, illnesses, and rock fracture occurrences are among the irregularly distributed temporal and spatial events. Uncertainty and irregularity are the common properties of

natural phenomena measurements in many researches, but the analytical solutions through numerical approximations require mostly regularly available initial and boundary conditions that cannot be obtained by lying regular measurement sites or time intervals. In an uncertain environment any cause is associated with different effects each with different level of possibility. Herein, possibility means some preference index for the occurrence of each effect. The greater the possibility index, the more frequent the event's occurrence.

2.2 Random and Randomness

Random and randomness are the two terms that are used in statistical sense to describe any phenomenon, which is unpredictable with any degree of certainty. An illuminating definition of randomness is provided by famous statistician Parzen (1960) as,

A random (or chance) phenomenon is an empirical phenomenon characterized by the property that its observation under a given set of circumstances does not always lead to the same observed outcome (so that there is no deterministic regularity) but rather to different outcomes in such a way that there is a statistical regularity.

The statistical regularity implies group and subgroup behaviors of a large number of observations so that the predictions can be made for each group more accurately than individual ones. For instance, provided that a long sequence of temperature observations are available at a location, it is then possible to say quite confidently that the weather will be warm, cool, cold, or hot tomorrow than specifying exactly by degree of centigrade prediction. The statistical regularities are as a result of some astronomical, natural, environmental, and social effects.

Deterministic phenomena are those in which outcomes of individual events are predictable with complete certainty under a given set of circumstances, provided that the initial and boundary conditions are known. It is necessary to check the validity of the assumption sets and initial conditions. In a way, with idealization concepts, assumptions, and simplifications deterministic scientific researches yield conclusions in the forms of algorithms, procedures, or mathematical formulations, which should be used with caution. The very essence of determinism is the idealization and assumptions so that uncertain phenomenon becomes graspable and conceivable to work with the available physical concepts and mathematical procedures. In a way, idealization and assumption sets render random phenomenon into conceptually certain case by trashing out the uncertainty components. A significant question that may be asked at this point is that, is there not any benefit from the deterministic approaches in natural studies, where the events are uncertain? The answer to this question is affirmative, because in spite of the simplifying assumptions and idealizations, the skeleton of the uncertain phenomenon can be captured by deterministic methods. For instance, determination of a trend component in a time series is a good example.

Even after the separation of a time series from its systematic components such as trend, the residuals should be checked for various feature properties so as to be able to apply probabilistic, statistical, and stochastic methodologies, which have a set of assumptions such as stationarity or weakly stationarity, homogeneity, independence or dependence, persistence or fuzziness.

2.3 Empirical Frequency and Distribution Function

An empirical work is possible provided that there is a set of measurements about the event concerned. In the content of this book measurements are considered as a sequence of records at equal time intervals, which are referred to as the time series. In mathematical notation Y_1, Y_2, \dots, Y_n is a time series with n samples. It can be shown succinctly as Y_i ($i = 1, 2, \dots, n$). Such a time series is shown for annual precipitation records in Fig. 2.1

Visual inspection of this figure indicates that obviously there is not a systematic follow-up between the successive years, and therefore, it is a random time series. This visualization is obtained by looking at the figure directly. It is possible to search visually for any trend, seasonality, and shift components. However, in this figure, it is not possible to identify any one of these deterministic components.

It is possible to obtain the number of frequencies provided that the variation domain is divided into equal length classes, which is shown in Fig. 2.2 with five classes. The frequency means the number of data values that fall within the class considered. For instance, in Fig. 2.3 the numbers of frequencies from left to right classes are 3, 30, 43, 27, and 13, respectively. The summation of these frequencies is equal to 116, which is the number of data in Fig. 2.1.

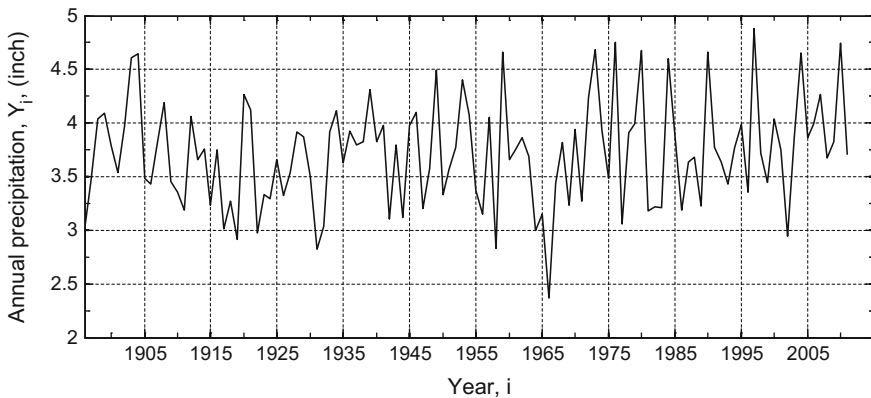


Fig. 2.1 Annual precipitation time series

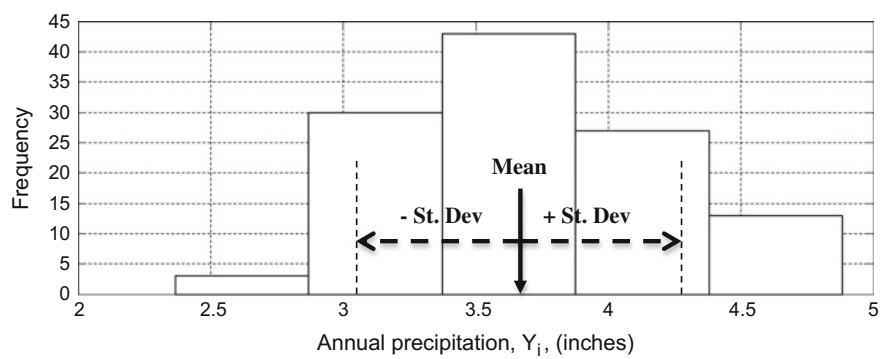


Fig. 2.2 Empirical frequency distributions

In order to appreciate the frequency concept, Fig. 2.2 is shown on the left-hand side in Fig. 2.3 vertically, and hence, one can understand, which data values fall into which class.

The frequency diagram in Fig. 2.2 seems almost symmetrical, which means that the number of data values more (less) than the arithmetic average is almost 50%. Such a symmetrical frequency distribution function has the arithmetic average (mean) value at or very close to the symmetry axis as shown in the same figure. As a statistical rule, in symmetrical distributions, the mean value is almost equal to the mode (the most frequently occurring value) and median (the value that divides the frequency distribution function into two halves) values.

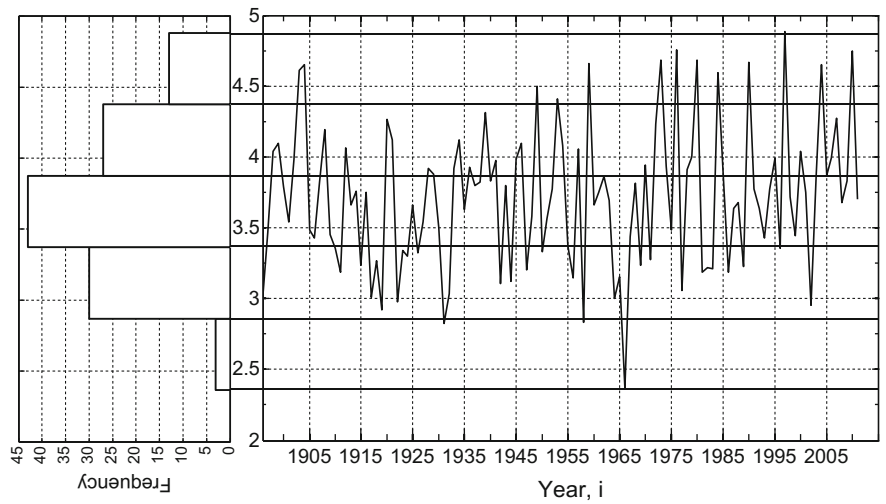


Fig. 2.3 Time-series and frequency combination

After all what have been explained above, it is possible to write down an equation among the class frequencies in a time series. If there are m classes each with frequencies f_i ($i = 1, 2, \dots, m$) then,

$$f_1 + f_2 + \dots + f_m = n, \quad (2.1)$$

where n is the number of data in the given time series.

It is also possible to appreciate the standard deviation value, which indicates arithmetic average deviations around the mean value. With this concept in mind, positive and negative standard deviation values are shown on the right and left of the arithmetic average value in Fig. 2.2. Apart from the symmetric frequency distribution function, various skewed (nonsymmetric) types are shown in Fig. 2.4.

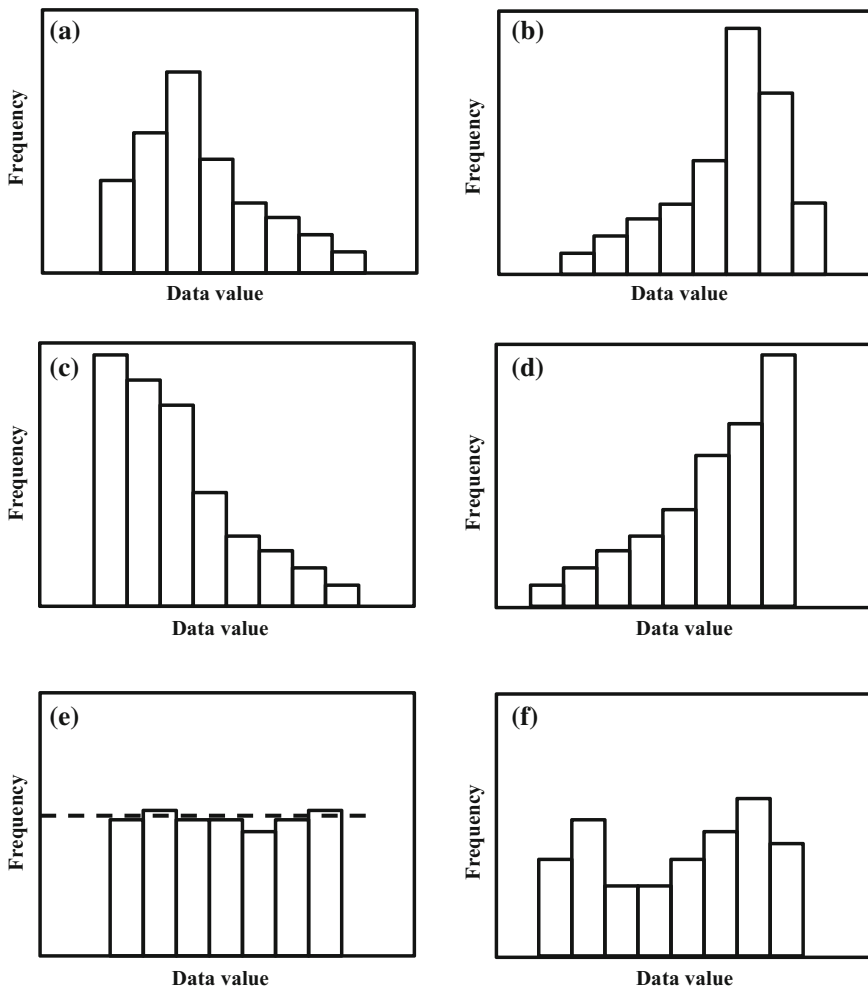


Fig. 2.4 a Negatively skewed, b positively skewed, c exponential, d J-shaped, e uniform, f bimodal empirical distributions

In practical studies, any one of these empirical frequency distribution functions emerges from a given time series data. It is possible to make various visual interpretations from each empirical frequency distribution function and this point is left to the reader so that s/he can increase personal expert views.

A very significant question at this stage; is it possible to identify any systematic component from the empirical frequency distribution functions? The answer is that it is not possible to deduce any trend component from the empirical frequency distributions.

2.3.1 Empirical Frequency and Trend

The unique way to be able to identify trend component through the employment of empirical frequency analysis is possible if the given time series is divided into two halves and for each half the empirical frequency distribution functions are obtained and compared with each other. For visual inspection, Fig. 2.5 presents a time series of 100 data values, where one can see visually that there is an increasing trend.

In order to see objectively whether there is a trend in the given time series or not, the time series is divided into two halves 50 and 50 data values and the resulting two empirical frequency distributions are given in Fig. 2.6.

Comparison of the two-half empirical frequency distributions indicates that there is significant shift toward the high values as obvious from Fig. 2.6b.

Figure 2.7 indicates decreasing trend component in the time series and it is identifiable even by naked eye visually without any quantitative methodology applications.

This time series is also divided into two halves and their empirical frequency distribution functions' comparison provides information about the decreasing trend as the frequency distributions in Fig. 2.8.

Comparison of the second half empirical frequency distribution with the first one indicates that there is a decreasing trend.

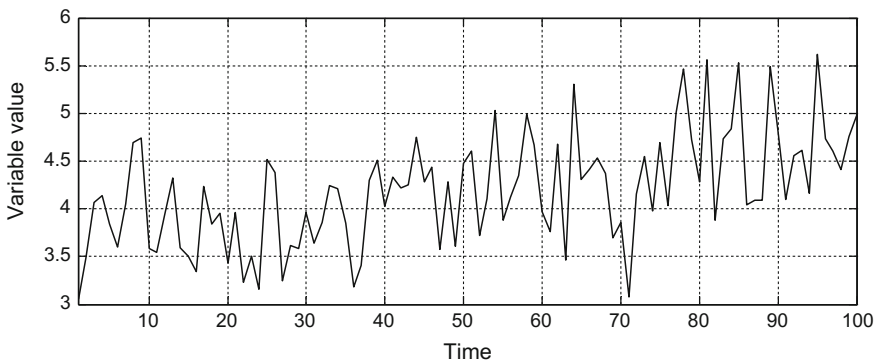


Fig. 2.5 Time series with visual increasing trend

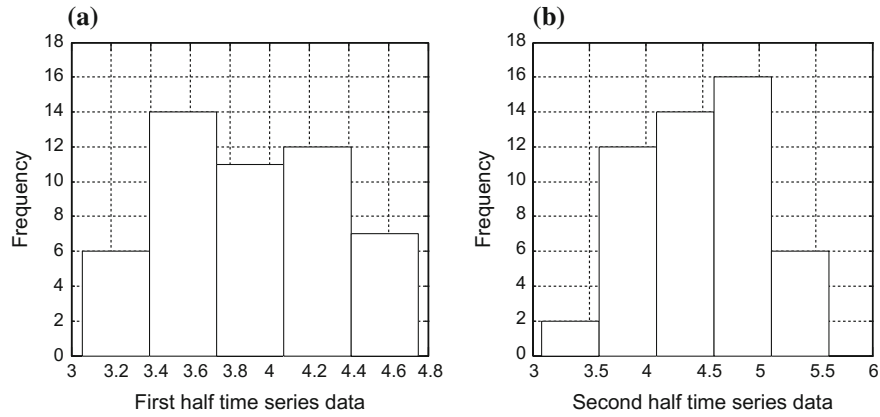


Fig. 2.6 Two halves empirical frequency distributions

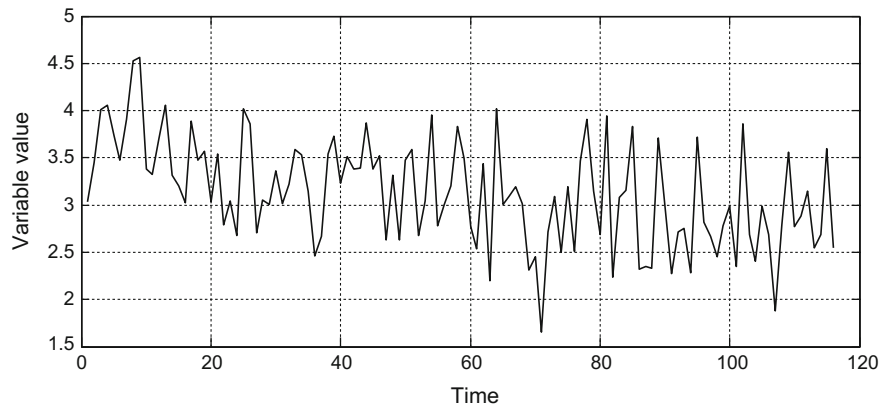


Fig. 2.7 Time series with visual decreasing trend

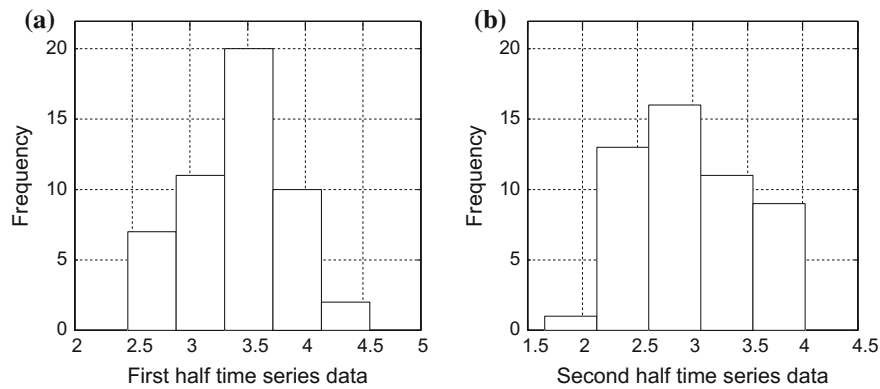


Fig. 2.8 Two halves empirical frequency distributions

Objective decision about the trend component existence within a time series by two-half empirical frequency distributions is possible through the chi square test.

Şen (2012, 2014) has proposed an innovative trend analysis methodology by using the concept of dividing a given time series into two equal halves. However, it is also possible to divide the time series into more than two equal size pieces and then in each piece, one can search for possible trend through the same innovative methodology. Various forms of the innovative trend methodology are explained in different chapters of this book (Chaps. 5–8).

2.4 Theoretical Probability Distribution Function (Pdf)

The probability has been expressed in daily life as percentages, but in the statistical context it may assume any value between 0 and 1, inclusive. When its value is equal to zero (one) then the event is absolutely impossible (possible). If the question is what is the probability in a given class then one can divide both sides of Eq. (2.1) by n , and hence, the probability of i th class is defined objectively as f_i/n . Finally, Eq. (2.1) can be written as,

$$\frac{f_1}{n} + \frac{f_2}{n} + \dots + \frac{f_n}{n} = 1 \quad (2.2)$$

or with probability, p_i ($i = 1, 2, \dots, m$), notations,

$$p_1 + p_2 + \dots + p_m = 1 \quad (2.3)$$

After these definitions, it becomes clear that in order to obtain the pdf one needs to divide each class frequency diagram in the empirical frequency distribution by the number of data and the resulting graph is referred to as histogram. Since by definition the area under any theoretical pdf is equal to 1, the histograms must be prepared preferably in such a way that the empirical area under it must also be equal to 1.

As mentioned in the previous subsection for the empirical frequency distribution, it is possible to search for possible trend component by comparison of two-half pdfs of a given time series. This is used descriptively for global warming discussions as in Fig. 2.9, where pdfs are shown theoretically as continuous curves.

In this figure, the pdf A can be regarded as the first half of a time series and B and C are the second halves. Comparison of the first half pdf (A) with the second (B) indicates that there is a shift toward the high values. If this shift is not sudden but gradual then it evolves with time along a smooth trend component. This last statement implies that there is an increasing trend in the given time series. If the amount of increase is asked then one can say that the arithmetic average (the peak) value, μ , of the first half has shifted toward the right by amount of $\Delta\mu$, and since this amount took place during the half time, $n/2$, evolution of a given time series of

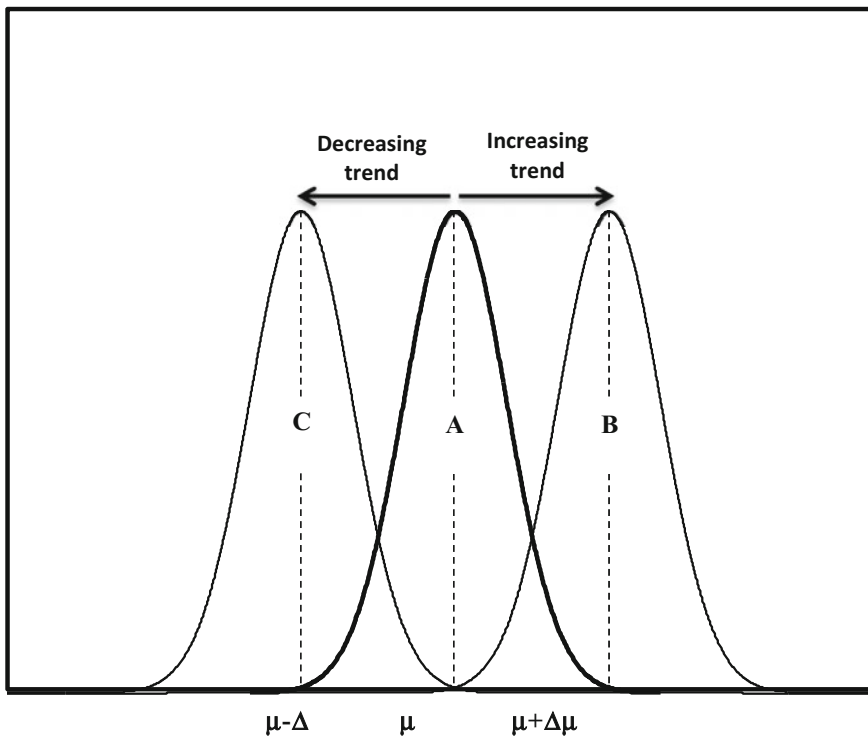


Fig. 2.9 Probability distribution function trend implications

duration (sample number), n , then the slope, S , of the trend component can be calculated as,

$$S = \frac{(\mu + \Delta\mu) - (\mu)}{(n/2)} = + \frac{\Delta\mu}{(n/2)} \quad (2.4)$$

This statement says that in order to find the slope of the trend, take the difference between the arithmetic averages of the two halves and divide it by the half number of data. This statement is one of the fundamental points about the innovative trend analysis in Chap. 5. In cases of nonsymmetrical pdfs, instead of arithmetic average, if possible, mode or preferably median value can be adapted.

On the other hand, if pdf C is considered as the second half, then there is a decreasing trend component in the time series and the slope should be calculated similar to Eq. (2.4) as follows.

$$S = \frac{\mu - (\mu + \Delta\mu)}{(n/2)} = - \frac{\Delta\mu}{(n/2)} \quad (2.5)$$

These interpretations indicate that visual inspections can lead to quantitative trend slope calculations in the simplest manner. These expressions will be used in the subsequent chapters, and especially during the explanation of the innovative trend analyses methodologies (Chap. 5).

2.5 Statistical Modeling

Complex interactions among the natural event characteristics give rise to spatial and temporal evolution of the phenomenon concerned, which must be controlled in a scientific manner so as to render its consequences to beneficial forms for human activities. For instance, prior to computer age the runoff event analysis dates back to the original work of Ripple (1881), who presented a deterministic graphical method for determination of the necessary reservoir capacity from an available sequence of recorded runoffs. This capacity is regarded in its simplest form as a prediction for future runoff regulations. However, such an approach has several drawbacks as follows.

- (1) The historical sequences will not reappear in the same order in future,
- (2) The statistical correlation structure will not have the same pattern,
- (3) The location of the extreme values along the time axis will not be in the same order as in the historic records.

The use of computers in natural event modeling led researchers to an explosion of simulation models for prediction purposes. Subsequently, a host of physical, conceptual, or black box type models are developed continuously and introduced into the literature. However, initially most of these models aimed at preserving some low order statistics, but later more specific real-time prediction processes are presented with rather simple recurrence model types, which extract necessary information from the available historical data, and later, their future predictions are achieved. Among the most important statistical parameters are the mean, standard deviation, coefficient of skewness, and the autocorrelation coefficients.

Especially, in long periods of time, the Hurst coefficient is also suggested for modeling purposes. Mandelbrot and Wallis (1969a, b, c) works led them to set horizons of the fractal geometry, which plays significant role in the investigation of chaotic behaviors of dynamic systems. In order to construct a dynamic model for the simulation of any natural phenomena, it is necessary to have a finite record of past observations. Given a historical record, the estimation process consists of computing an estimate of the variable concerned at time lead k , the position of which relative to observation period leads to three types of estimations problems.

- (1) The estimation of state at any time instant during observation period is referred to in statistics as “smoothing” operation or in mathematics as “interpolation”,

- (2) Estimation of the state at the final observation time instant is called as “filtering”,
- (3) State variable estimation at a time instant after the final observation, which is referred to in uncertainty domain as “prediction” or in certainty, i.e., in mathematics domain as “extrapolation”.

In addition to these stages, after the model adaptation and determination of its parameters there is “verification” stage where the suitability of chosen model to the historical observation sequence is sought. This stage includes search for suitable model theoretically to make parameter estimates for the model and to check the suitability of the model. However, in any study, the most important stages are the identification, verification, and subsequent prediction phases, and they follow each other.

The basic estimation work has been performed by Gauss in early 1800s who tried to fit the most suitable curve through the scatter of points by having the least squares technique as a criterion, which constitutes without exception the basis of any uncertainty event assessment in statistics and stochastic process modeling studies. The successful application of the least squares technique for almost two centuries is due to the following factors.

- (1) The minimization of sum of squared errors leads to a system of linear equations, which are easy to solve and do not require an extensive theory. This approach is used frequently in trend identification calculations,
- (2) The sum of the squares corresponds in many different contexts to various interpretations such as in physics, the energy is expressed as the sum of squares; in mechanics it represents moment of inertia, in statistics it provides the variance about the fitted curve, and consequently, it can be used as a measure of the goodness-of-fit test,
- (3) An assumption of a definite explicit analytical form to represent that the observed data constitutes the principal application of the classical least squares technique,
- (4) Without proposing an explicit analytical expression, it is possible to apply the least squares technique to filtering problems. For instance, a known differential equation may represent the phenomenon concerned. Likewise, the storage and continuity equations are explicit expressions for certain phenomenon in physics,
- (5) Wiener (1949) has founded a different application version of the least squares technique by assuming certain statistical properties for the useful signal and noise constituents of observation sequences. The significant difference of Wiener’s approach lies in the fact that the useful and noise parts are characterized not by analytical forms, but by their statistical properties, such as the mean values are supposed to be zero or rendered to zero and also serial and cross-autocorrelations,

- (6) After 1960 in order to reduce the computation burden, Kalman suggested an elegant procedure for the adaptive prediction in the form of recursive filtering. This technique is generally considered as igniting the widespread interest in the subject of estimation.

2.5.1 Deterministic-Uncertain Model

In various modeling studies in different disciplines, there are input and output random structured variables. In any system design, the input and output variable measurements show randomness in the sense that any data value cannot be predicted from the previous data values with certainty, therefore, they must be treated by probabilistic, statistical, and stochastic models. On the basis a convenient prediction model, future replica of the output variables can be obtained within certain limits of errors such as ± 5 or 10%. This was the main problem in front of system planners and managers and the question is which value to adopt in the model procedural design? In the beginning of the twentieth century, Hazen (1914) has suggested the use of the following procedure without the availability of any computer and even calculator. His method was random drawing of paper pieces that are mixed in a sack. Each paper piece had a certain number written on it and then folded and put into the sack. If the past observations of any phenomenon are denoted as a sequence, Y_1, Y_2, \dots, Y_n , it is possible to calculate its various statistical parameters (arithmetic average, standard deviation, serial correlation coefficient, etc.). This sequence is the naturally ordered record of measurements, and the statistical parameters are dependent on the whole data values and they are valid for the record duration only. These historical data series can be used for the simple prediction of the future values according to the following steps.

- (1) Historical data record: There are daily, monthly, or annual records of past variable measurements. Let the number of records be n ,
- (2) Each one of these measurements is written separately on equal size paper pieces. They are folded and then put into a sack,
- (3) The pieces are drawn one after the other, and hence, a new time series is constructed of the same duration or even longer with the same data values but at different times,
- (4) After each drawing, the paper pieces are either returned to the sack or not. In the former case, it is possible to generate sequences as long as desired. However, in the latter, the maximum length of the synthetic data can be equal to the length of the original data. After the generation is complete, whole pieces can be returned into the sack and again another random sequence (replica) can be generated. In this manner one is able to obtain an ensemble of synthetic sequences (time series).

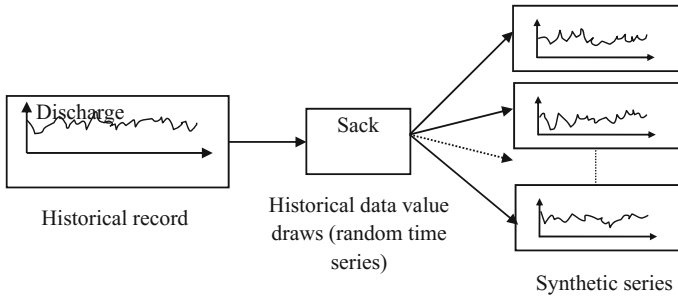


Fig. 2.10 Deterministic-uncertain method sequences

After the completion of such a generation procedure, the input sequences are treated for the assessment of decision or determination of design quantity. For instance, if the sequences are supply levels of some quantity then by knowing the demand level, it is possible to decide about the supply sufficiency time durations. In the case of insufficiency, additional supply is withdrawn from available storages.

The sequences obtained by this aforementioned deterministic-uncertain method can be referred to as synthetic sequences (replicas), which have the following points in common.

- (1) Each synthetic sequence has the same arithmetic average as the original sequence,
- (2) Each synthetic sequence has the same variance and the standard deviation as the original sequence,
- (3) Other statistical parameters (mode, median, skewness, kurtosis, etc.) are also the same in addition to the relative frequencies, hence also the relative frequency distribution,
- (4) The major assumption in such a draw system is that each one of the generated sequence is regarded as independent from others, but this is not valid practically. Each one of the generated sequence has its own serial correlation coefficient that may be significantly different from each other. The general procedural function of this deterministic-uncertain methodology is shown in Fig. 2.10.

2.5.2 Probabilistic-Statistical Model

This procedure is more developed than the previous one and instead of using the same data values in synthetic sequence generation, the relative frequency distribution of the measured data is adopted as the root of the generation procedure and it is fitted with the most convenient theoretical pdf through the chi square test. This time the data are not drawn from the sack with the repetition of the historical data,

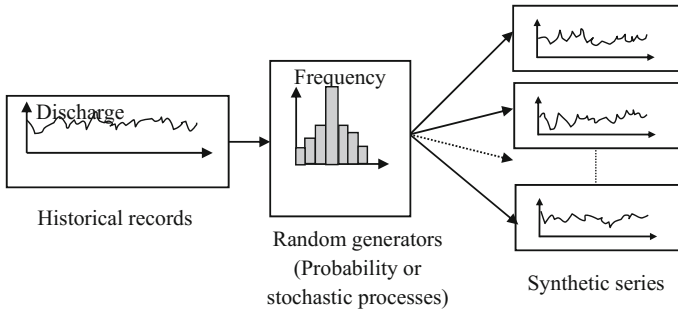


Fig. 2.11 Statistical procedure stages

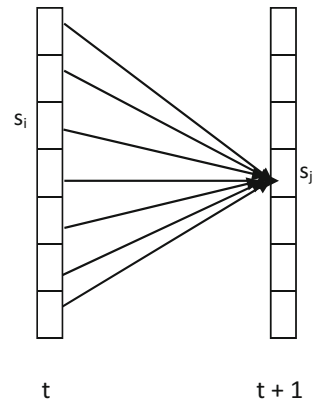
but the synthetic sequence values are drawn from the theoretical pdf automatically in computers from random number generators. This approach yields synthetic sequences that have in the long run almost the same statistical parameters, but it also provides extreme value contribution into the generation procedure, which is never possible with the deterministic-uncertainty method. Figure 2.11 shows the basic stages in the statistical procedure.

In this generation process, although the original time series statistical structure is preserved, but the serial correlation coefficient is not taken into consideration.

2.5.3 Transitional Probability Model

Although the probabilistic-statistical methods are based on the statistical parameter preservation by a theoretical pdf adaptation as explained above, they depend on the class interval relative frequencies obtained from a given measurement series. In the transitional probability approach, the sequence of class intervals are considered to remain the same at each time instant as shown in Fig. 2.1, but transition probabilities or transition frequencies are considered from one class interval at t instant to the next one at $t + 1$ instant as in Fig. 2.12.

Fig. 2.12 State and transition probabilities



If there are m class intervals, there will be m class interval relative frequencies, which are referred to herein as the state probabilities. Furthermore, there are $m \times m$ interclass interval transition probabilities, which are relative joint frequencies. Hence, instead of the statistical parameters, the state and transition probabilities are used in the modeling of a given time series. These are known in the literature as the Markov chain models (Feller 1968; Box and Jenkins 1970). Their application requires the following steps:

- (1) Construction of the histogram from a given time series,
- (2) Calculation of the class interval frequencies (state probabilities) from the histogram,
- (3) Calculation of transition relative frequencies (transition probabilities) between two successive time instances.

The transition probabilities can be considered in the form of a matrix, where rows are for time instant t , and columns for $t - 1$. Such a matrix is called as the transition probability matrix, P_T .

$$P_T = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & \cdot & \cdot & \cdot & \cdot & \cdot & p_{1n} \\ p_{21} & p_{22} & p_{23} & p_{24} & \cdot & \cdot & \cdot & \cdot & \cdot & p_{2n} \\ p_{31} & p_{32} & p_{33} & p_{34} & \cdot & \cdot & \cdot & \cdot & \cdot & p_{3n} \\ p_{41} & p_{42} & p_{43} & p_{44} & \cdot & \cdot & \cdot & \cdot & \cdot & p_{4n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ p_{n1} & p_{n2} & p_{n3} & p_{n4} & \cdot & \cdot & \cdot & \cdot & \cdot & p_{nn} \end{bmatrix} \quad (2.6)$$

Since the transition from state, say, s_i at time t , to state, s_j at time $t - s$ is the same as the transition from state s_j at time $t - 1$ to s_i at time t , the transition matrix will have a diagonal symmetric form, i.e., $p_{ij} = p_{ji}$. Hence, $m(m - 1)/2$ transition probabilities are necessary for the definition of the transition probability matrix. The transition probabilities along the major diagonal are all equal to 1 ($p_{ii} = 1$), because they represent the transition from a state to itself. The state and the transition matrix provide the basis of future phenomenon prediction.

2.6 Stochastic Models

These are the most advanced alternatives for synthetic time series generation with inclusion of all the properties in the previous models and additionally they have sound probabilistic, statistical, and transitional foundations collectively. Any time series Y_i ($i = 1, 2, \dots, n$), has in general four distinctive components as the periodic

fluctuations, P_i , trend, T_i , sudden jump (step) J_i and stochastic, S_i components. Hence, it is possible to write a given time series mathematically as,

$$Y_i = P_i + T_i + J_i + S_i \quad (2.7)$$

After the identification of sudden jump and its separation, this expression becomes with remaining components of jump (sudden change) free Y_i time series as,

$$X_i = P_i + T_i + S_i \quad (2.8)$$

After the trend separation trendless time series, Z_i , expression takes the following form.

$$Z_i = X_i - T_i = P_i + S_i \quad (2.9)$$

It is now time to try and separate the periodic component, which can be done through the harmonic analysis (Sect. 2.6.3).

Following the separation of the periodic component, the remaining stochastic part, S_i , has inherit random variability that can be treated by probabilistic, statistical, and stochastic evaluations methodologies (Box and Jenkins 1970). Most often, these methodologies are applied automatically to available records, and consequently, there are numerous papers published in different disciplinary journals that do not provide any new approach, but the application of well-known methods to specific data sets leads to desired information. It is recommended that original methodology development foundations are first based on the qualitative information deductions, which help to establish theoretical backgrounds with a set of fundamental assumptions among which the homogeneity and stationarity are the most important ones.

2.6.1 Homogeneity (Consistency)

This assumption is valid only in the case when the record series originate from the same population. This implies that the record series has a constant time invariant arithmetic average, which also means that the record temporal variation is free of trend or jump (shift) components. Otherwise, the records have heterogeneous structure, which need comparatively rather complex mathematical, probabilistic, statistical, and stochastic treatments.

Since, in homogeneous series the arithmetic average is time variant, the simplest method to check for homogeneity is to compare the arithmetic averages after the division of the original time series into two or more portions of the same length. Buishand (1982, 1984), Jayawardena and Lau (1990) have summarized the application of three statistical homogeneity tests.

Under a null hypothesis, H_o , a given time series, Y_i ($i = 1, 2, \dots, n$) has the same mean value throughout the effective time period. The alternative hypothesis, H_a , is generally vague, since often no reliable prior information is available about possible changes in the mean. Of course, Y_i 's have some empirical pdf and in the application of the homogeneity test, a theoretical joint pdf is assumed for the Y_i 's.

In general, tests require serially independent structure for time series. If the tests are performed on seasonal or annual time series then this point cannot be a significant restriction. The test statistic pdfs are derived for stochastically independent and identically distributed time series. If there are slight departures from the normality, the test can still be applied confidently. In practical homogeneity tests, generally the pdf of test statistics is overlooked. The properties of test statistics are illustrated for the case that the Y_i 's are normally distributed with mean (Buishand 1982).

$$E(Y_i) = \begin{cases} \mu & i = 1, 2, \dots, m \\ \mu + \Delta & i = m + 1, m + 2, \dots, n \end{cases} \quad (2.10)$$

and the variance of the time series is simply,

$$\text{Var}(Y_i) = \sigma_Y^2 \quad (2.11)$$

According to this model, there is a shift (sudden jump) in the time series of magnitude Δ after m observations, and therefore, it is not a homogeneous time series. Homogeneity implies that the data in the series belong to one population, and hence, have a time invariant mean. Heterogeneity may arise due to changes in the method of data collection and/or the environment in which it is done (Fernando and Jayawardena 1994).

2.6.2 Stationarity

Different samples from the same population have practically the same statistical parameters within the range of sampling error (variability). Any time series with all the statistical parameters without significant change is referred to as the strictly stationary process. This is an impossible property in natural records. However, in practical applications, weakly (second order) stationary records are suitable for the application of the classical statistical methodologies including the stochastic processes. This type of stationarity implies that the time series has the first-order (arithmetic average) and second-order (variance) moments depending on the time differences (Box and Jenkins 1970). Independence of the variance from time is referred to as the homoscedascity in the statistics literature.

In order to check the stationarity property, at least two non-overlapping parts are considered from the original time series. If these two subseries look similar then visually one can say that the original series is stationary. This implies that stationary time series cannot include trends, jumps, or periodicities. Stationarity can be

checked either by parametric or nonparametric tests. Parametric tests are employed usually in the analysis of economic time series based on a certain number of data (Aigner et al. 1977; Bauer 1990).

The researchers who care for the frequency properties of time series prefer to work with nonparametric stationarity tests. Among these researchers are electronic engineers and a certain branch of statisticians and stochastic process experts. They consider the whole system as a “black box”, where only input and output signals are important and the system identification may be achieved through some simple procedures such as the regression technique and spectral analyses. Depending on the work type, researcher uses parametric and nonparametric approaches. The significance of nonparametric approaches is that they are not based on the assumption of normal pdf. This point makes the nonparametric approaches to be used more frequently in practical applications even though they are less powerful than the parametric alternatives. As suggested by Bethea and Rhinehart (1991) in order to reach almost to the same conclusions, the nonparametric tests need 5–35% more data than parametric tests.

2.6.3 Periodicity (Seasonality)

It is well known that the periodic fluctuations are embedded into a natural time series as a result of mainly astronomic events such as Earth’s rotation around the sun annually with implication of seasonality; diurnal variations due to day–night variations. In the social and economic time series, seasonality is the main factor for the periodic component existence. In general, such variations in any time series records become graspable and quantifiable at time scales less than a year (daily, weekly, monthly, three-monthly, and six-monthly). Figure 2.13 presents different periodicities in the given time series.

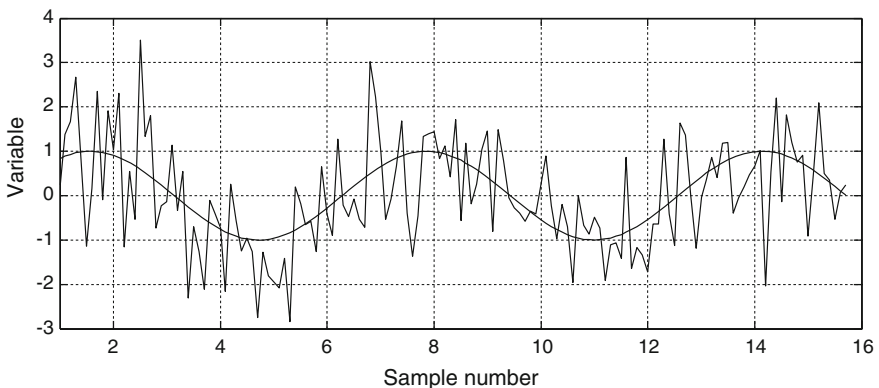


Fig. 2.13 Periodicity (seasonality) components

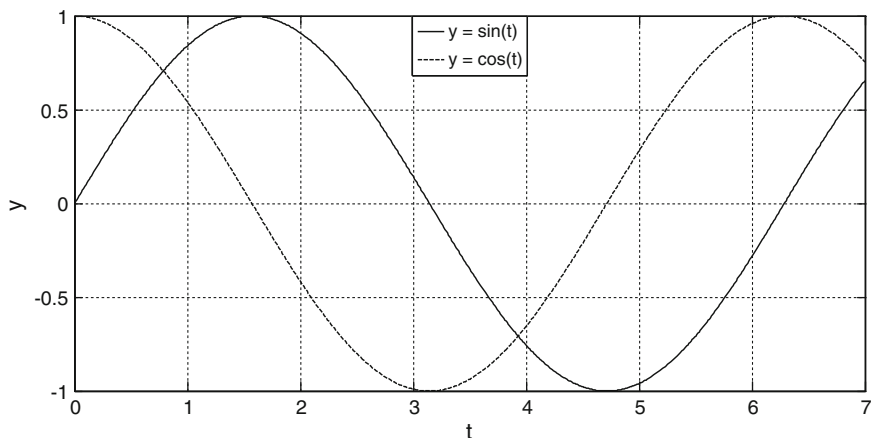


Fig. 2.14 Sine and cosine waves

In order to quantify and detect the periodicity, the commonly used methodology is the Fourier series (Maidment and Parzen 1984; Kite 1989; Jayawardena and Lai 1989; Pugacheva et al. 2003). Some researchers like Jayawardena and Lai (1989) have used the autocorrelation technique for testing the periodicity in time series.

Periodic component is the part of time series which reflects the seasonal effects. The astronomical effects in any time series can be observed provided that record durations are less than one year such as day, week, month, and season. Periodic fluctuations can be expressed as regular sine and cosine waves as in Fig. 2.14.

These waves have their amplitudes, a , basic wave period, T , and phase angle, Θ . These three quantities define a sine wave as,

$$Y_t = a \sin\left(2\pi \frac{t}{T} + \Theta\right) \quad (2.12)$$

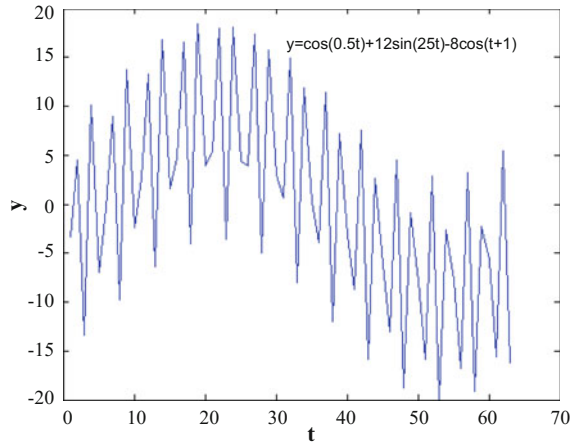
Cosine waves are also defined in a similar way. In order to identify the periodic component in a time series, a series of waves first a sine (and cosine) wave is considered with basic wave length equal to the whole record length, then equal to the half of the total length, then one-third, etc. The summation of regular waves leads to a rather random (irregular) looking wave as in Fig. 2.15.

In this manner, it is possible to approach an irregular wave, like a given time series, by the summation of various regular waves. A regular sine wave can be expressed as,

$$Y_{i1} = a_1 \sin\left(2\pi \frac{1}{n} i + \Theta_1\right) \quad (2.13)$$

Each wave is called as a harmonic. In general, the frequency of j th harmonic has a frequency as j/n and its regular wave expression is,

Fig. 2.15 Regular wave summations



$$Y_{ij} = a_j \sin\left(2\pi \frac{j}{n} i + \Theta_j\right) \quad (2.14)$$

The expansion of this sine wave gives,

$$Y_{ij} = a_j \sin\left(2\pi \frac{j}{n} i\right) \cos \Theta_j + a_j \cos\left(2\pi \frac{j}{n} i\right) \sin \Theta_j \quad (2.15)$$

Since $a_j \cos \Theta_j$ and $a_j \sin \Theta_j$ are constant, they are represented by A_j and B_j , and hence, the previous expression takes the following form,

$$Y_{ij} = A_j \sin\left(2\pi \frac{j}{n} i\right) + B_j \cos\left(2\pi \frac{j}{n} i\right) \quad (2.16)$$

This is the contribution of j th harmonic to i th data value. If m harmonics are considered then the time series will have the approximation as,

$$Y_i = \sum_{j=1}^m \left[A_j \sin\left(2\pi \frac{j}{n} i\right) + B_j \cos\left(2\pi \frac{j}{n} i\right) \right] \quad (2.17)$$

This has a zero arithmetic average value, and hence, it alone cannot represent the arithmetic average of the time series. Therefore, it is necessary to add the average term, \bar{Y} , which leads to,

$$Y_i = \bar{Y} + \sum_{j=1}^m \left[A_j \sin\left(2\pi \frac{j}{n} i\right) + B_j \cos\left(2\pi \frac{j}{n} i\right) \right] \quad (2.18)$$

In practice, since the number of harmonics, namely m , is a finite value not more than 7, this last expression approaches a given time series with error term, h_i , and finally, the periodicity equation takes the following form.

$$Y_i = \bar{Y} + \sum_{j=1}^m \left[A_j \sin\left(2\pi \frac{j}{n} i\right) + B_j \cos\left(2\pi \frac{j}{n} i\right) \right] + h_i \quad (2.19)$$

In this expression, A_j 's and B_j ($j = 1, 2, \dots, m$), there are $2m$ unknowns. They can be obtained from a given time series value by the minimization of sum of error squares, $\min(\sum h_i^2)$ condition, leading to the following expressions.

$$A_j = \frac{2}{n} \sum_{i=1}^{n-1} Y_i \cos\left(2\pi \frac{j}{n} i\right) \quad (2.20)$$

and

$$B_j = \frac{2}{n} \sum_{i=1}^{n-1} Y_i \sin\left(2\pi \frac{j}{n} i\right) \quad (2.21)$$

The summation of the squares of these terms is equivalent to the variance of the given time series as,

$$\sigma_j^2 = A_j^2 + B_j^2 \quad (2.22)$$

and the phase angle is defined as,

$$\Theta_j = \tan^{-1}\left(\frac{A_j}{B_j}\right) \quad (2.23)$$

The major defect of this approach is that the frequencies must be whole number divisions as $1/n, 2/n, \dots, m/n$.

2.6.3.1 Known Period Case

If the basic period in a time series is known, then the periodicity component can be eliminated by using simple statistical parameters without any consideration of trigonometric functions. For instance, if hourly data are available, then it is known that the periodicities are confined within the 24-h period, and therefore, a table similar to Table 2.1 can be presented for the exposition of available data and there are $N = 24n$ hourly values, where n is the number of days. If such a time series is shown as $Y_0, Y_1, Y_2, \dots, Y_{N-1}$, their exposition is given in Table 2.1.

In the last two rows, the arithmetic averages, $(\bar{Y}_i, i = 0, 1, 2, \dots, 23)$, and the standard deviations, $(\sigma_i, i = 0, 1, 2, \dots, 23)$, of hourly data are calculated. If the arithmetic average of each hour is subtracted from the corresponding hourly data,

Table 2.1 Hourly data

Y_0	y_1	y_2	.	.	.	y_{24}
Y_{24}	y_{25}	y_{26}	.	.	.	y_{48}
Y_{48}	y_{49}	y_{50}	.	.	.	y_{60}
.
.
.
$Y_{24(n-1)}$	$y_{24(n-1)+1}$	$y_{24(n-1)+2}$.	.	.	y_{24n-1}
\bar{Y}_0	\bar{Y}_1	\bar{Y}_2	.	.	.	\bar{Y}_{23}
σ_0	σ_1	σ_2	.	.	.	σ_{23}

Table 2.2 Periodicity free hourly data

$Y_0 - \bar{Y}_0$	$Y_1 - \bar{Y}_1$.	.	.	$Y_{23} - \bar{Y}_{23}$
$Y_{24} - \bar{Y}_0$	$Y_{25} - \bar{Y}_1$.	.	.	$Y_{47} - \bar{Y}_{23}$
$Y_{48} - \bar{Y}_0$	$Y_{49} - \bar{Y}_1$.	.	.	$Y_{60} - \bar{Y}_{23}$
.
.
.
$Y_{24(n-1)} - \bar{Y}_0$	$Y_{25(n-1)+1} - \bar{Y}_1$.	.	.	$Y_{24n-1} - \bar{Y}_{23}$
0	0	.	.	.	0
σ_0	σ_1	σ_2	.	.	σ_{23}

then the remaining term does not have any more periodic fluctuation on the arithmetic average level. This subtraction procedure is shown in Table 2.2.

One can notice that in this table, the arithmetic average of each column is zero, but the standard deviations remain without any change. In order to eliminate the periodicity effect in the standard deviation, each column in the previous table must be divided by the standard deviation of the column leading to Table 2.3. The time series in this table is a standardized data, because it has zero arithmetic average and unit variance.

Table 2.3 Standardized data

$(Y_0 - \bar{Y}_0)/S_0$	$(Y_1 - \bar{Y}_1)/S_1$.	.	.	$(Y_{23} - \bar{Y}_{23})/S_{23}$
$(Y_{24} - \bar{Y}_0)/S_0$	$(Y_{25} - \bar{Y}_1)/S_1$.	.	.	$(Y_{47} - \bar{Y}_{23})/S_{23}$
$(Y_{48} - \bar{Y}_0)/S_0$	$(Y_{49} - \bar{Y}_1)/S_1$.	.	.	$(Y_{60} - \bar{Y}_{23})/S_{23}$
.
.
.
$(Y_{24(n-1)} - \bar{Y}_0)/S_0$	$(Y_{25(n-1)+1} - \bar{Y}_1)/S_1$.	.	.	$(Y_{24n-1} - \bar{Y}_{23})/S_{23}$
0	0	.	.	.	0
1	1	.	.	.	1

2.7 Time Series Truncation

It is possible to explore the internal structure of any time series by truncating it at a certain truncation level, Y_0 (Şen 2015). Such a truncation gives rise to two-valued verbal variables such as deficit/surplus, dry/wet, cloudy/non-cloudy, flood/drought, hot/cold, rainy/non-rainy, gain/loss, etc. These two-valued variables help decision maker to base his/her final plans toward a certain goal. In some system design studies, the variables must be categorized into two classes on the basis of a certain truncation level. Let us consider that for practical applications, the time series given in Fig. 2.16 is truncated at Y_0 level.

After the truncation, the time series is converted into two mutually exclusive events along the time axis as surplus, S_i , and deficit, D_i . In mathematical sense, surpluses have positive and deficits have negative values. In general, when a time series, Y_i ($i = 1, 2, \dots, n$) is truncated at Y_0 constant level then at the i th location, there is either $S_i = Y_i - Y_0 > 0$ or deficit $D_i = Y_0 - Y_i < 0$. The following properties are observable from such a truncation.

- (1) Along the time series, there are appearances of S_i and D_i in a randomly alternate manner. The first important point is that at two successive time instances there are four possible events as deficit–surplus (DS), deficit–deficit (DD), surplus–deficit (SD), or surplus–surplus (SS),
- (2) If there are n elements in a time series with n_d deficits then the number of surpluses is,

$$n_s = n - n_d$$

or

$$n_d + n_s = n \quad (2.24)$$

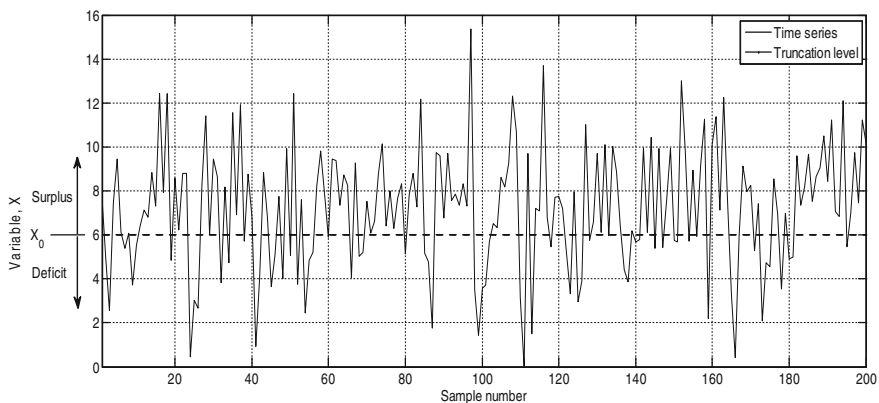


Fig. 2.16 Time series truncation

Dividing both sides by the total number of elements, n , yields,

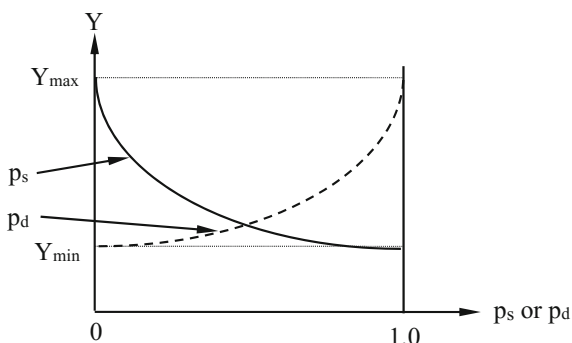
$$p_s + p_d = 1 \quad (2.25)$$

Herein, $p_d = n_d/n$ is the probability (percentage) of deficits and likewise $p_s = n_s/n$ is for surplus probability. Depending on the level of truncation, Fig. 2.17 indicates the relationship between the truncation level and these probabilities,

It is noted that the truncation level changes between the maximum, Y_{\max} and minimum, Y_{\min} data values. In the case of symmetric relative frequency distribution like the normal (Gaussian) pdf, the truncation level that is equal to the arithmetic average is also equal to the median and model levels, which implies that $p_s = p_d = 0.5$. In this case approximately, $Y_0 = (Y_{\max} + Y_{\min})/2$.

- (3) In case of uninterrupted sequence of two or more deficit (surplus) events, a deficit (surplus) period is valid. These periods follow each other along the time axis alternatively. In a given time series, the difference between the number of surplus and deficit periods is either 0 or 1.
- (4) The maximum deficit duration corresponds to the critical deficit period within the given time series,
- (5) The transition from a deficit period to surplus has DS bivariate event, whereas SD bivariate event is valid in the case of surplus followed by deficit period. The first bivariate event is referred to as the upcrossing and the second one as downcrossing event. The more these bivariate variables are in a time series the less is the dependence.
- (6) The summation of deficits (surpluses) along a deficit (surplus) duration is referred to as the deficit (surplus) magnitude.
- (7) The division of magnitude to duration is the deficit (surplus) intensity.

Fig. 2.17 Surplus and deficit percentages



2.7.1 Statistical Truncations

In the statistical studies, the deviations from the arithmetic average for a given time series data $(Y_i - \bar{Y})$ are very significant. Such deviations constitute the basic definitions of variance, covariance, correlation coefficients, and the coefficient of determination in regression analysis. It is possible to categorize the overall time series on the basis of standard deviation distances above and below of the arithmetic average as in Fig. 2.18.

In this figure, σ indicates the standard deviation of the whole time series. With 1, 2, and 3 standard deviation limits, a given time series may be viewed in seven categories as in Table 2.4 with different specifications.

In practice, most of the time series values fall within the normal limits with extreme values outside above and below normal extreme limits. In order to standardize all the time series to a common dimensionless base, the standard values, y_i , can be obtained according to the following formulation.

$$y_i = \frac{Y_i - \bar{Y}}{\sigma_Y} \quad (i = 1, 2, \dots, n) \quad (2.26)$$

In the case of normal (Gaussian) pdf consideration of 1, 2, 3, and 4 standard deviation values around the arithmetic mean leads to the following numerical percentages.

In interval,

- $-1 < y_i < +1$ 68.269%, i.e., with probability 0.68269
- $-2 < y_i < +2$ 95.450%, i.e., with probability 0.95450
- $-3 < y_i < +3$ 99.730%, i.e., with probability 0.99730
- $-4 < y_i < +4$ 99.994%, i.e., with probability 0.99994

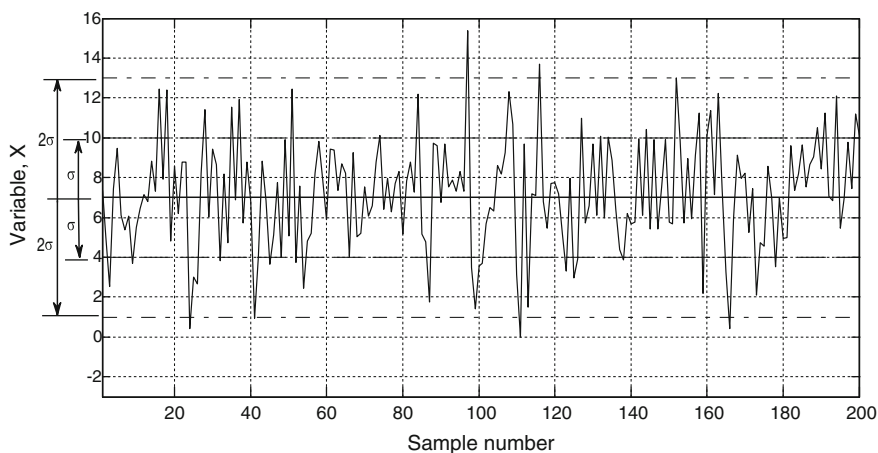


Fig. 2.18 Standard deviation truncation

Table 2.4 Truncation levels and specifications

Truncation	Specification
$\bar{Y} + 3\sigma < Y_i$	Above normal extreme
$\bar{Y} + 3\sigma < Y_i < \bar{Y} + 2\sigma$	Rather super-normal extreme
$\bar{Y} + 2\sigma < Y_i < \bar{Y} + 1\sigma$	Above normal
$\bar{Y} + 1\sigma < Y_i < \bar{Y} - 1\sigma$	Normal
$\bar{Y} - 1\sigma < Y_i < \bar{Y} - 2\sigma$	Below normal
$\bar{Y} - 2\sigma < Y_i < \bar{Y} - 3\sigma$	Rather subnormal extreme
$Y_i < \bar{Y} - 3\sigma$	Below normal extreme

In Fig. 2.19, a standard normal (Gaussian) pdf is shown with arithmetic mean $\bar{y} = 0$ and variance 1 in addition to the categorical division according to the standard deviation at three levels.

Different from the statistical truncation, there are others that are useful for various human activities. In such truncations, the comfort and benefit of humans are taken into consideration. These are referred to herein as the engineering truncations. For instance, for the comfort of humans, the temperature must not be under 15 °C, and for the plant life below 7 °C. The daily water demand of Istanbul City, Turkey, is $1.5 \times 10^6 \text{ m}^3$, which can be considered as the truncation level for water supply to the city.

In the previous explanations, the values below and above of any truncation level are given in terms of numbers, percentages, or probabilities. However, as shown in Table 2.5, it is also possible to specify different phenomena with different words verbally.

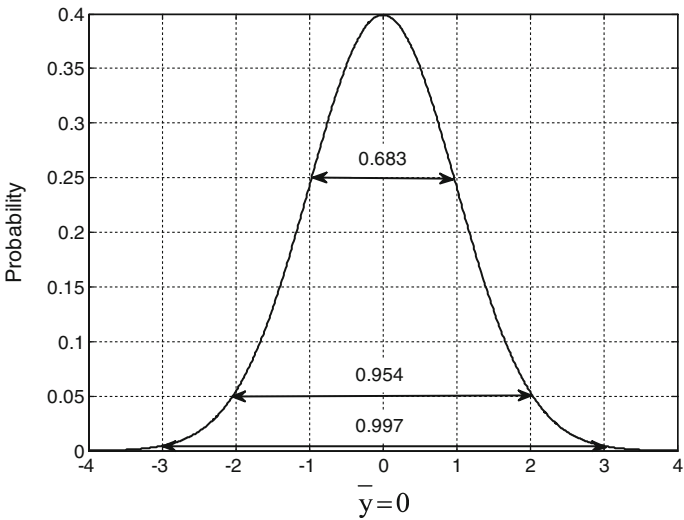


Fig. 2.19 Standart normal distributions

Table 2.5 Time series truncation and specifications

	Temperature	Rainfall	Runoff	Humidity	Cloud	General
If $Y_i < Y_0$	Cold	Rainy	Dry	Humid	Open	Deficit
If $Y_i > Y_0$	Hot	Non-rainy	Wet	Non-humid	Close	Surplus

These bivariate specifications play important role in many diverse human social, environmental, economic, health, and engineering activities.

2.8 Data Smoothing

In the records of time series, there are local haphazard and very random fluctuations that may sometimes hide the general variation trend. In order to get rid of these disturbances, it is necessary to smooth the time series through some procedures. In general, the equation for a time series can be written as composed of deterministic, D_i , and stochastic, S_i , parts similar to Eq. (2.7). The time series components are already explained in Sect. 2.6. The summation of the random component, and hence, its arithmetic average is equal to zero, and therefore, the arithmetic average of the process is equal to the arithmetic average of the deterministic part, i.e., $\bar{Y} = \bar{D}$. This shows that random component can be eliminated through some average procedure. The remaining deterministic part is the smoothened part of the time series which is shown in Fig. 2.20.

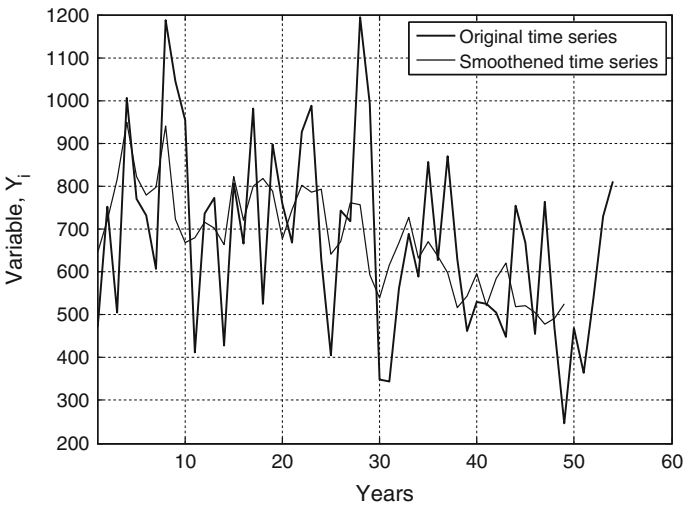


Fig. 2.20 Smoothened time series

2.8.1 Moving Averages

Informal regression methods based on moving averages at certain window widths are used as the smoothing techniques to disclose possible hidden trend components. Moving average methodology helps to identify and highlight possible long-term nonlinear trends with smoothing of short-term fluctuations. Moving average procedure is commonly used in many economy sectors.

The most frequently used procedure for smoothing is the moving average approach, where a certain length of series is replaced in an overlapping manner by the arithmetic average. For instance, in Fig. 2.20, 5-year window width is used for successive arithmetic moving average smoothening. In most applications, first, third-order moving average procedure is recommended and subsequently the order can be increased up to seventh order, if necessary. If a time series is, Y_i ($i = 1, 2, \dots, n$), its third-order moving average smoothing X_i ($i = 1, 2, \dots, n - 2$) can be achieved as follows.

$$X_1 = \frac{Y_1 + Y_2 + Y_3}{3}, X_2 = \frac{Y_2 + Y_3 + Y_4}{3}, \dots, X_{n-2} = \frac{Y_{n-2} + Y_{n-1} + Y_n}{3}$$

In a third-order moving average, there are $n - 2$ terms from a time series of length n . In the case of m -order moving average procedure, there are $n - m + 1$ terms, X_i ($i = 1, 2, \dots, n - m + 1$).

The above-mentioned moving average gives equal weights to each part of smoothened time series. However, in some cases, it is necessary to give different weights to each smoothening term. In practice, the most frequently used versions are as follows.

$$X_i = \frac{Y_i + 2Y_{i+1} + Y_{i+2}}{4} \quad (2.27)$$

or

$$X_i = \frac{Y_i + 4Y_{i+1} + 6Y_{i+2} + 4Y_{i+3} + Y_{i+4}}{16} \quad (2.28)$$

2.8.2 Difference Smoothing

A very simple procedure is the successive difference method, which for a given series, Y_i , with lag-one difference operation, the terms in a new time series, X_i , become with $n - 1$ terms as,

$$X_1^{(1)} = Y_2 - Y_1, X_2^{(1)} = Y_3 - Y_2, \dots, X_{n-1}^{(1)} = Y_n - Y_{n-1}$$

If the difference is taken at lag- m apart from each other, then the new series X_i has $n - m$ terms as,

$$X_1^{(m)} = Y_m - Y_1, X_2^{(m)} = Y_{m+1} - Y_2, \dots, X_m^{(m)} = Y_n - Y_m$$

Example 2.1 In Table 2.6, 23 terms are given as a time series in the first column and it is smoothened according to difference procedure at lags 1, 2, 3, and 10 in the same table. It is obvious that lag-10 differences have more fluctuations, because the successive terms become more independent from each other. In general, the further away the two values are from each other the less is the dependence between them.

Table 2.6 Application of difference procedure

Data (X_i)	Differences			
	1	2	3	10
9.05				
	-0.96			
8.08		0.79		
	-0.17		0.96	
7.97		1.75		
	1.58		-1.00	
9.50		0.75		
	2.33		-4.67	
11.83		-3.92		
	-1.59		7.05	
10.84		3.13		242.2
	1.54		-1.08	
11.78		2.06		67.2
	3.59		-4.95	
15.37		-2.90		-366.1
	0.69		3.38	
16.06		0.48		544.5
	1.17		-4.83	
17.23		-4.35		-622.2
	-3.18		6.67	
14.05		2.32		661.8
	-0.86		-3.13	
13.19		-0.81		-714.2
	-1.67		5.79	
11.52		4.98		759.1
	3.31		-8.88	
14.83		3.90		-708.8
	-0.59		5.02	
14.23		1.12		533.7

(continued)

Table 2.6 (continued)

Data (X_t)	Differences			
	0.53		-3.13	
14.77		-2.01		-317.6
			1.03	
13.29		-0.98		162.2
	-2.46		0.78	
10.83		-0.20		-90.9
	-2.66		2.15	
8.17		1.95		
	-0.71		0.04	
7.46		1.99		
	1.28		-1.22	
8.74		0.77		
	2.05		-3.75	
10.79		-2.98		
9.86				
<i>Sample number, n</i>				
23	22	21	20	13
<i>Average</i>				
11.68	0.0372	0.001	-0.189	11.60

2.9 Jump (Shift)

This implies a sudden change (downward or upward) in a time series, which may also have a linear downward or upward trend. Such changes are common in financial time series and also in a surface flow discharge record series after the construction of a dam or a diversion channel (Fig. 2.21).

For a jump component, there is almost a sudden change in the effective environmental conditions.

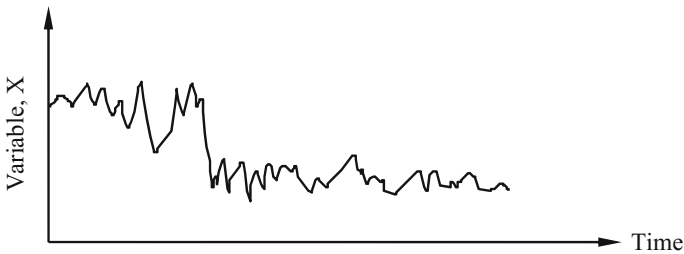


Fig. 2.21 Sudden jump components in a time series

- (1) If the location of the measurement station is changed, then the change in the environmental conditions may lead to sudden effects in the time series measurements. Such a jump may be very distinctive as shown in Fig. 2.21,
- (2) After natural hazards, there may appear sudden jumps in the measurements. For instance, eruptions may load the lower atmosphere with dust, and accordingly, this may cause sudden jumps in some meteorological records,
- (3) Due to miss-maintenance, local defects in the instruments may lead to sudden jumps. For instance, if there appears a small hole in the rain gauge then the measurements will be lower than before.

2.10 Correlation Coefficients

In any time series, apart from the visible components in a graph, there are also non-visible features that need identification and quantitative evaluation. These features are concerned with the internal structure of a given time series. The most significant one is the serial dependence, which is concerned with the question, whether there is an effect of any event occurrence in a time to the next time step occurrence? Such a feature, which is referred to as the serial autocorrelation is an indispensable component relevant to any time series whether natural or artificial. It is also known as a memory effect in two types as the short-memory and long-memory effects, where the latter type is the persistence. Scientific terminology for the short memory effect is the autocorrelation coefficient. Linguistically and qualitatively, the short-term memory effect can be expressed as “time series low values follow low values and high values follow high values”. This expression provides an ability to visualize a time series and then to deduce whether there is a short-memory effect or not.

In general, processes can be viewed under two very broad categories as dependent and independent according to time scale considerations. Usually, the smaller the time interval between two successive events the greater is the dependence, and hence, there is persistence, but large-time apart natural events imply independence. This classification is also in accord with the rarity or frequency of the event. For instance, flood and earthquake occurrences are among rare natural events that occur along time axis, and therefore, they are considered as independent from each other. Similar arguments are valid also for low natural events such as droughts. In such problems, the serial (internal) correlation coefficient is ignored and the probabilistic treatment of the successive event occurrences becomes very easy according to the probabilistic modeling.

Correlation coefficients are useful in determination of the relationship strength between two variables. Two different time series can be related to each other proportionally, inversely or there may not be any correlation between them such a relationship is calculated by the cross-correlation coefficient. For the quantification

of correlation, there are different procedures as parametric and nonparametric alternatives.

On the other hand, within the same time series the successive data values might affect each other, which is expressed by the serial correlation coefficient. For instance, rainfall of today might be affected partially from yesterday's rainfall occurrence. In general, rainy periods follow rainy periods and dry periods follow dry periods. Furthermore, high rainfall amounts follow high amounts and low values follow low values. These two statements indicate that so far as the rainfall occurrences and their amounts are concerned, there are serial (internal) relationships to a certain extent.

In mathematics, when two variables are related to each other their variation or plots on a Cartesian coordinate system does not appear as a horizontal or vertical line (see Fig. 1.2e, f), but rather a straight line with a slope or a curve with many tangential slopes (see Fig. 1.2a–d). The simplest form of dependence has linearity, which is always used in the statistics or stochastic modeling works. If there are two different time series, they can be plotted as one versus the other. By visual inspection of the scatter points, one can appreciate whether the dependence is high or low and directly or reversely proportional. In the case of scatters around a straight line (trend line) there is dependence.

For serial correlation structure, time series Y_1, Y_2, \dots, Y_n , is shifted by a certain lag (for instance lag-one) so as to obtain another parallel time series as $Y_2, Y_3, Y_4, \dots, Y_{n-1}$. They have $n - 1$ common point. The scatter diagram of these two time series gives rise to $n - 1$ scatter points on the Cartesian coordinate system (Fig. 2.22).

If straight-line trend appears through the scatter points then it is possible to conclude that there is dependence between the two variables, otherwise they are independent. The most suitable straight line through these scatter points gives the dependence measurement as its slope. The more the deviation of the slope from $\pm 45^\circ$ (1:1 and -1:-1) line is, the smaller is the dependence. In Fig. 2.23 an independent scatter diagram is shown.

2.10.1 Pearson Correlation Coefficient

There are two types such as serial correlation and cross-correlation. The serial correlation coefficient, ρ_k , is expressed for a given time series, Y_i ($i = 1, 2, \dots, n$) and lag- k as follows.

$$\rho_{sk} = \frac{\sum_{i=1}^{n-k} (Y_i - \bar{Y})(Y_{i-k} - \bar{Y})}{\sqrt{\sum_{i=1}^{n-k} (Y_i - \bar{Y})^2} \sqrt{\sum_{i=n-k}^n (Y_i - \bar{Y})^2}} \quad (2.29)$$

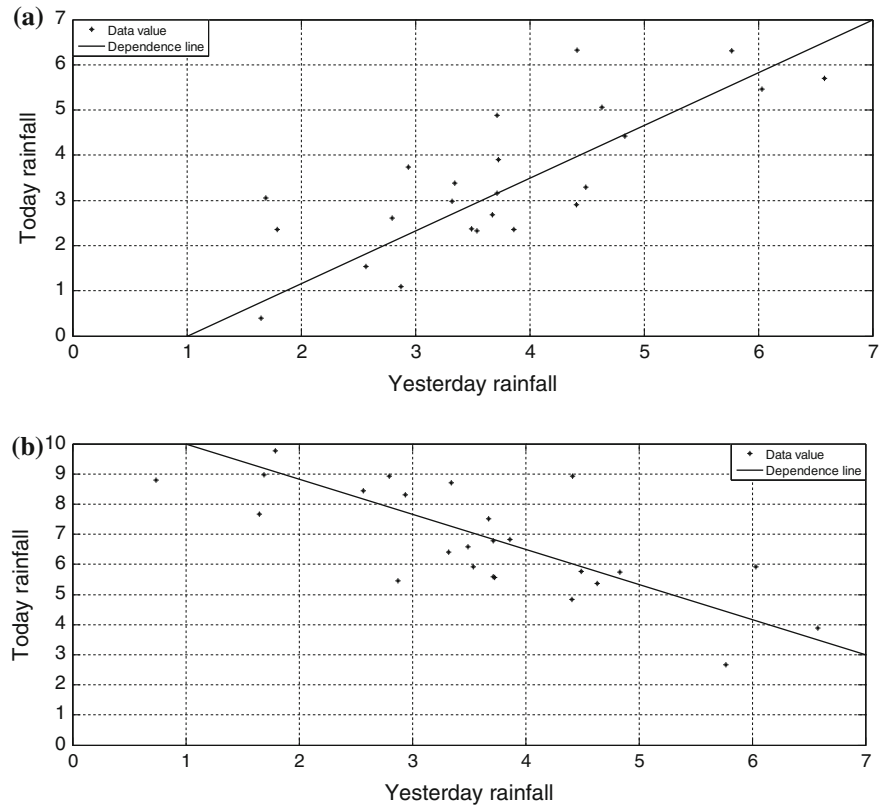


Fig. 2.22 Dependent scatter diagrams, **a** positive dependence, **b** negative dependence

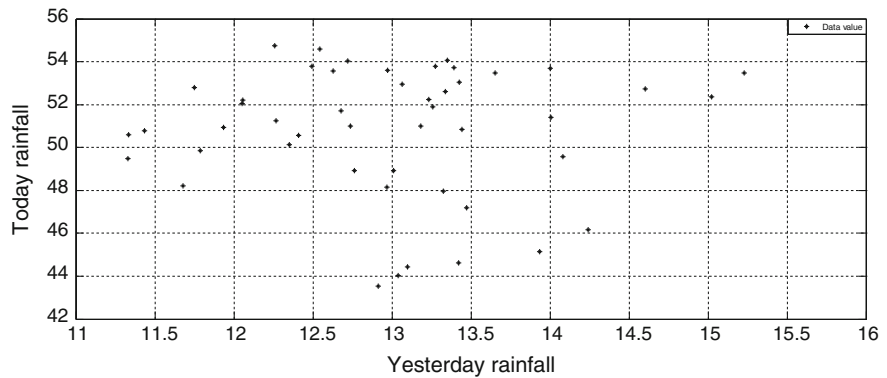


Fig. 2.23 Independent scatter diagrams

On the other hand, similarly the Pearson cross-correlation, ρ_c , between two time series Y_i and X_i is defined as,

$$\rho_c = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}, \quad (2.30)$$

where \bar{X} and \bar{Y} are the mean of respective time series. It is also possible to search for lag-one or more cross-correlations.

The correlation coefficient takes values between -1 and $+1$. The closer the cross-correlation coefficient values are to zero the more random, i.e., independent are the two time series, otherwise, values close to $+1$ (-1) imply positively (negatively) strong cross- or serial correlations. Positive correlation means direct proportionality (see Fig. 2.22a) and negative value shows inverse proportionality (see Fig. 2.22b). In the case of positive dependence, high (low) values follow high (low) values, whereas in the case of negative dependence high (low) values follow low (high) values. The dependence that is calculated through Eq. (2.29) is the serial correlation or autocorrelation coefficient. Similarly to lag-one, lag-two or more lag correlation coefficients can be calculated simply from a given time series. In general, there is an upper practical limit for lag- k as $k \leq n/3$. The theoretical distribution parameters as the average and variance for lag-one in Eq. (2.29) are given as (Anderson 1942),

$$\overline{\rho_{sk}} = -\frac{1}{(n-1)} \quad (2.31)$$

and,

$$\sigma_{sk}^2 = \frac{1}{(n-1)}, \quad (2.32)$$

respectively. The pdf of ρ_c was shown to be asymptotically normal with the mean $E(\rho_P) = 0$ and variance as in Eq. (2.32). In the test of serial correlation, a single-tail normal pdf is used.

As mentioned earlier, the Pearson correlation coefficients assume any value between -1 and $+1$, inclusive with specifications in Table 2.7. One should not memorize this table, because they are more or less the subjective opinion of the author. Other authors may deviate slightly from these specifications owing to their experiences, but such deviations are not significant in practical works. It must be kept in mind that correlation coefficients are the measure of linear dependence between two variables or within the same time series. If the correlation is not linear, then these definitions are invalid.

Table 2.7 Correlation coefficient classes

Numerical value intervals	Linguistic interpretations
$\rho_P = -1.0$	Completely negative dependence
$-1.0 < \rho_P < -0.9$	Strong negative dependent
$-0.9 < \rho_P < -0.7$	Quite negative dependence
$-0.7 < \rho_P < -0.5$	Weak negative dependence
$-0.5 < \rho_P < -0.3$	Very weak negative dependence
$-0.3 < \rho_P < -0.1$	Insignificant negative dependence
$\rho_P = 0.0$	Complete independence
$0.1 < \rho_P < 0.3$	Insignificant positive dependence
$0.3 < \rho_P < 0.5$	Very weak positive dependence
$0.5 < \rho_P < 0.7$	Weak positive dependence
$0.7 < \rho_P < 0.9$	Quite positive dependence
$0.9 < \rho_P < 1.0$	Strong positive dependence
$\rho_P = 1.0$	Complete positive dependence

The following points are the deficiencies of the Pearson correlation coefficient concept in practical applications.

- (1) Even if the correlation is not linear, the correlation value will appear between -1 and $+1$. This may not have logical and physical meaning, because the Pearson correlation coefficient definition is valid for linear relationships,
- (2) If there are one or more extreme values in a time series, then these values affect Eq. (2.30) in such a manner that the correlation coefficient appears biased and/or unrepresentative,
- (3) The data must abide with a normal pdf, otherwise, the correlation coefficient is not meaningful,
- (4) For meaningful and reliable correlation coefficient calculations, the standard deviation of the data must be constant, i.e., homoscedasticity property must be valid,
- (5) Correlation coefficient definition in Eq. (2.30) cannot be used for verbal and linguistic data,
- (6) If data is transformed by any means to have normal pdf then the correlation coefficient of the transformed data is not the same with the original data (Şen 1977). Even the reverse transformation does not guarantee that the correlation coefficient is equal to the observed correlation value.

After what has been said above, it is obvious that the domain of the Pearson correlation coefficient is rather restrictive, and prior to its use all necessary assumptions must be cared for their validity.

2.10.2 Kendall Correlation Coefficient

In order to alleviate the defects in the Pearson correlation coefficient, other procedures are suggested for the same purpose. One of these techniques is the consideration of data ranks instead of data values in natural sequence. This is the requirement of the Kendall correlation coefficient, ρ_K , which gets rid of the extreme value effects. This coefficient can be used even though the data may have a skewed pdf. It is applicable even in the cases of some missing data or incomplete measurements. In general, for the same data series Kendall correlation coefficient is smaller than the Pearson coefficient. For this reason, although strong correlation is observed through the use of the Pearson correlation coefficient as 0.9 or more, the same is valid as 0.7 in the case of Kendall correlation coefficient. Kendall coefficient can be calculated without any calculator even by hand. It is also capable to measure nonlinear correlations. The superiority of this coefficient over the Pearson coefficient is due to the following points.

- (1) It measures even the nonlinear relationships,
- (2) It is not affected by extreme values,
- (3) Even after the transformations Kendall coefficient remains the same.

For instance, if $\log(Y_i)$ and $\log(X_i)$ are used instead of the original data (Y_i and X_i), the Kendall correlation coefficient will remain the same. In the calculation of this correlation coefficient the following steps are necessary,

- (1) Rank one of the time series into ascending order and replace the data in the next time series with the ranks of this time series. Hence, one of the time series is ordered and the other had replacement of values according to the ranks of the first one. In the case of correlation, there will appear simultaneous increase in both series. If there is increase in the ranked time series, and decrease in the other time series implies then a negative correlation is valid. Otherwise, there is no correlation between them,
- (2) Any data value in the ranked series, say Y_i , is compared with all the data after its location, Y_j ($j = i + 1, i + 2, \dots, n$), and if $Y_i < Y_j$ then +, otherwise for $Y_i > Y_j$ a - sign is attached. For the data value at place i , there are $(n - i)$ number of alternative + and - signs. If the same procedure is repeated for all the data values without any equal data value to each other, then there are $n(n - 1)/2$ signs. If half of these have + sign then the data sequence is considered as independent. If + (-) signs are more than - (+) signs then there is positive (negative) dependence,
- (3) If the total numbers of + and - signs are denoted by P and N then the Kendal correlation coefficient is defined as,

$$\rho_K = \frac{2(P - N)}{n(n - 1)} \quad (2.33)$$

By definition this has values between -1 and +1,

- (4) For the test of independence the necessary statistical quantity, K , is defined as,

$$K = P - N \quad (2.34)$$

which is the difference between the numbers of + and - signs. The K value may be positive or negative with zero expectation. Theoretical studies indicate that its standard deviation can be expressed as,

$$\sigma_K = \sqrt{\frac{n(n-1)(2n+5)}{18}} \quad (2.35)$$

On the other hand, for more than 10 data values, the distribution of the Kendall correlation coefficient approaches the Gaussian pdf. Whether ρ_K is different from zero can be tested by the standard normal pdf. If the necessary standard value is S , then the test statistics can be calculated as,

$$z_s = \begin{cases} \frac{S-1}{\sigma_s} & \text{if } S \geq 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sigma_s} & \text{if } S \leq 0 \end{cases} \quad (2.36)$$

If this standard value is less than the critical value found on the basis of a certain significance level then the data is considered as independent.

2.10.3 Spearman Correlation Coefficient

In the nonparametric statistics domain, the analogous to the Pearson correlation is named as the Spearman's rank correlation coefficient. Pearson correlation coefficient requires that both variables should comply by the normal pdf, which is not the case in many disciplines. For calculating this nonparametric correlation coefficient, both data sets are ordered separately from each other. Hence, there are two sequences of ranks, one for Y time series, $R(Y_i)$, and other for X time series, $R(X_i)$. If for each i the ranks are of Y_i equal to ranks of X_i , then the Spearman's rank correlation is regarded as perfect. The rank correlation is defined as the sum of the difference between the corresponding ranks of Y_i and X_i . Analogous to the parametric version of the coefficients the correlation values are scaled between -1 (perfect negative) and $+1$ (perfect positive) correlation. In between the value is equal to zero indicating no correlation. Spearman's rank correlation calculation steps are as follows.

- (1) As the null hypothesis, H_0 , the correlation between Y_i and X_i is assumed as equal to zero. This is referred to as the hypothetical correlation value, $\rho_s = 0$,
- (2) Alternative hypothesis, H_a is that this correlation coefficient is different than zero,

- (3) The test statistic, ρ_S , which is referred to as the Spearman's rank correlation coefficient is defined in terms of each data set ranks and the number, n , of data in each set as,

$$\rho_S = 1 - \frac{6 \sum_{i=1}^n [R(Y_i) - R(X_i)]}{n(n^2 - 1)} \quad (2.37)$$

As with the other nonparametric methods, values of X_i and Y_i can vary extensively without affecting the final result. It is necessary to keep in mind that ρ_S does not imply good linear relationship, rather than linearity. It is quite possible to obtain low Spearman's rank correlation coefficient for high Pearson's parametric correlation coefficient. However, in many applications, it is unusual for Pearson's coefficient to provide a statistical test result markedly superior to Spearman's rank correlation approach even with normally distributed data. Another version of the previously defined Spearman correlation coefficient can be found as follows,

$$\rho_S = \frac{\sum_{i=1}^n R(Y_i)R(X_i) - n\left(\frac{n+1}{2}\right)^2}{\frac{n(n^2-1)}{12}} \quad (2.38)$$

In the case of positive correlation, high values of Y_i ranks follow high X_i ranks; otherwise there is a negative correlation. Theoretical studies indicate that in the case of trend nonexistence for big data values, this coefficient appears according to a Gaussian pdf with the following average and variance expressions as,

$$\bar{\rho}_S = 0 \quad (2.39)$$

and

$$V_{\rho_S} = \frac{1}{(n-1)}, \quad (2.40)$$

respectively. The test must be carried out with two-tailed pdf by the assumption as the null hypothesis that there is no trend component in the time series. The test value at any significance level results as $\rho_{\text{sig}} > \rho_S$, otherwise the time series is not homogeneous. In the case of significant level $\rho_S > 0$ implies the existence of an increasing trend.

2.11 Persistence/Nonrandomness

Persistence is one of the most important properties in many system designs concerning the storage capacity of reservoirs, average return periods, failure risks, hidden periodicities, trends, and drought properties. Its consideration in analytical derivations of design criteria presents difficulties and for this reason most often the analytical expressions are obtained on the basis of nonpersistent (independent or short-memory) processes. Although the conventional autocorrelation coefficients and functions are used in many design problems, but the very definition of the autocorrelation function requires that the underlying process generating mechanism abide with normal (Gaussian) pdf. It is therefore, necessary to convert non-Gaussian pdf into normal pdf in order to make benefit of the available analytical expressions. During the transformation process, the very persistence genuine property of the basic variables is not preserved although the statistical parameters such as the average, standard deviation, skewness coefficient, and kurtosis are maintained in the transformed normal pdf.

Persistence and randomness are two distinctive properties of a time series. Randomness is another term for nonpersistence and it is defined as the independence among time series values, whereas persistence (correlation) occurs provided that the successive time series data affect each other. Persistence (correlation) is a tendency of the successive time series values to “remember” their antecedent values’ influence.

2.11.1 Short-Memory (Correlation) Components

Simple successive dependence models are representations of a linear line on a Cartesian coordinate system between the value from the time series and the following value. Hence, given a time series of Y_1, Y_2, \dots, Y_n with n observations for the simplest successive dependence at lag-one apart, this sequence yields $n - 1$ points on a scatter diagram as shown in Fig. 2.22a, b. It is the trend slope of the straight line that is a representative of the simple dependence, i.e., short-term correlation. All the serial correlations are obtained in this manner. In Fig. 2.22a, b, the horizontal axis represents the previous value, say, Y_{i-1} , whereas the vertical axis is for current value, Y_i . It is possible to infer the simplest model (lag-one Markovian) mathematically as the straight line with deviations, u_i , from this line as,

$$Y_i = a + bY_{i-1} + u_i, \quad (2.41)$$

where, a and b are the model parameters. Such a simple model does not have any assumption concerning the pdf of the time series, but the model is based on the linearity assumption. This is one of the most significant conclusions about the serial dependence that the classical correlation coefficient measures the linear dependence, and therefore, prior to its application, it is necessary to look at the scatter diagram of successive values so as to infer whether this assumption is valid.

Otherwise, in the case of nonlinearity one can still obtain classical correlation coefficient but without knowing the underlying facts of nonlinearity. Unfortunately, most often in practical applications, this point is overlooked and moreover the correlation coefficient is calculated and applied rather blindly. The model parameters, a and b , can be obtained in any manner without formal procedures. For instance, Eq. (2.41) can be considered without error and in this case any two data values give rise to two equations, which can be solved for a and b parameters. If this procedure is applied to all possible pairs from the sequence, then a set of a and b parameters can be calculated and their averages are adopted as a and b . However, this is very naive way of parameters estimation (Chap. 4).

If the necessary tests are not performed and the data are not checked for the basic assumptions then all what have been explained above leave suspicions in the coefficient estimations. In practical studies, researchers most often do not care or even think about these restrictive assumptions, and consequently, the coefficient estimations might remain biased. Even the amount of the global bias is not known, and therefore, bias correction procedures cannot be defined and applied (Şen 1974). Hence, the parameter estimates of Eq. (2.41) remain under suspicion. In order to avoid all these restrictive assumptions rather than the application of procedural regression analysis to data with a set of restrictive assumptions, it may be preferable to try and preserve only the arithmetic averages and variances of the sequence. After all the arithmetic averages and variances are the most significant statistical parameters in any design work.

Equation (2.41) can also be interpreted as a first-order Markov process. In such a case, since always a physical value is assumed to exist, i.e., there is no zero value, it is possible to consider that $a = 0$ in this equation. On the other hand, with theoretical restrictive assumptions similar to the regression approach especially the normal (Gaussian) pdf of the physical variable, it can be shown that, Eq. (2.41) can be brought into a stochastic process form as follows.

$$(X_i - \mu) = \rho(X_{i-1} - \mu) + \sigma(1 - \rho^2)^{1/2}u_i, \quad (2.42)$$

where μ , σ , ρ , and u_i are the arithmetic average, standard deviation, first-order correlation coefficient, and uncertain residual error term, respectively. In such a model, u_i is normally distributed random variable with zero mean and unit standard deviation. The stochastic model in Eq. (2.42) generates normally (Gaussian) distributed variables. It is important to notice at this stage that the correlation coefficient is defined for linear dependence and for normal pdf stochastic variates only.

2.11.2 Long-Memory (Persistence) Component

Persistence is commonly referred to as long-term dependence between successive observation values in a time series. Natural variables (landslide, earthquake, flood, and tsunami) are uncertain in character but there are embedded features that give

rise to better quantitative description of these series. Among these features are long-term averages and standard deviations and better frequency distribution behaviors according to a theoretical pdf such as normal, logarithmic normal, Gamma, Gumbel, Pearson, etc. However, none of these features are capable to give the measure of successive dependence, except the correlation coefficient or persistence measures such as rescaled ranges (Hurst 1951; Şen 1974). Dependence measures including classical correlation and persistence strength decreases as the basic time interval of the time series increases. For instance, daily records have more dependence characteristic than annual series. Even in high (floods) or low (droughts) natural events there are persistence, but unfortunately for the sake of brevity and simplicity these are ignored in many practical applications. Assumption of serial independence makes calculations simple within the probability theory only but the conclusions always appear as over-estimations. For instance, the ideal expected size of a storage reservoir, $E(R)$, is given simply for a first-order Markovian process as (Şen 1974),

$$E(R) = \sqrt{\frac{2(1 - \rho)}{\pi(1 + \rho)}} \sigma, \quad (2.43)$$

where ρ is the lag-one serial correlation coefficient and σ is the standard deviation of time series. The ratio, $(1 - \rho)/(1 + \rho)$ is smaller than 1, and consequently, the expected size for dependent process is smaller than the independent case, which appears for $\rho = 0$ as,

$$E(R) = \sqrt{\frac{2}{\pi}} \sigma \quad (2.44)$$

It is, therefore, very significant to consider the short- or long-term dependences within any natural phenomena, if the design is expected to perform in the best possible manner economically. On the other hand, Douglas et al. (2000) have shown that even in the low values of seven-day basic time interval, there are significant serial correlations reaching to 25% at the runoff stations throughout the USA.

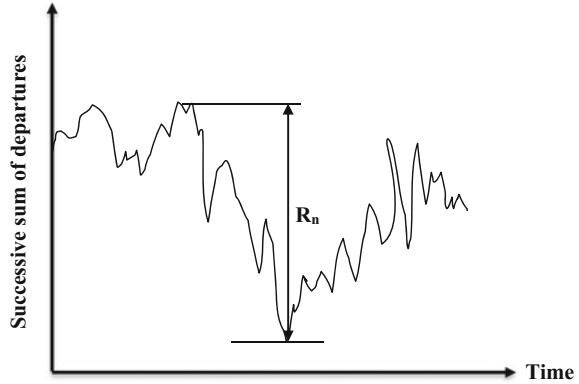
2.11.2.1 Rescaled Range and Hurst Phenomenon

Hurst (1951, 1956), Şen (1974) studied long-term fluctuations within a large number of geophysical records and found that

$$\frac{R_n}{S_n} \approx n^h, \quad (2.45)$$

where R_n is the range of cumulative departures from the sample mean and S_n is the standard deviation estimation. The range is based on the cumulative sums of departures from the time series arithmetic average as in Fig. 2.24.

Fig. 2.24 Range of a time series



In this empirical formulation, h is referred to as the Hurst coefficient, which assumes values theoretically between 0 and 1 but for independent processes its value is equal to 0.5. In many geophysical phenomena h does not appear as 0.5, and hence, its deviation from 0.5 is called as the “Hurst phenomenon” implying long-term dependence, i.e., persistence. Such a discrepancy has been accounted on the basis of three factors.

- (1) The non-normality of the pdf of the underlying variables,
- (2) Effect of small samples, i.e., bias effect in the statistical sense,
- (3) The autocorrelation structure.

Especially, the last factor has caused introduction of different theoretical stochastic models among which the “fractional Brownian processes” (Mandelbrot and Wallis 1968) are the major ones in addition to the white Markov (Şen 1974) or AutoRegressive Integrated Moving Average (ARIMA) processes (Box and Jenkins 1974). Division of the range, R , which is the storage volume, by the standard deviation, S , theoretically leads for serially independent processes to the expectation value as (Feller 1968),

$$E\left(\frac{R_n}{S_n}\right) = 2\sqrt{\frac{2}{\pi}}n\sigma, \quad (2.46)$$

where n is the sample length. However, for serially dependent processes, the expectation of the rescaled range, R/S , is derived by Şen (1974) as follows.

$$E\left(\frac{R}{S}\right) = \frac{2\sqrt{n}}{\sqrt{\pi(n-1)}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \sum_{k=1}^n k^{-\frac{1}{2}} \left[\frac{1+\rho}{1-\rho} - \frac{2\rho(1-\rho^k)}{k(1-\rho)^2} \right]^{\frac{1}{2}}. \quad (2.51)$$

References

- Aigner, D., Knox Lovell, C. A., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6, 21–37.
- Anderson, T. W. (1942). Distribution of the serial correlation coefficients. *The Annals of Mathematical Statistics*, 13(1), 1–13.
- Bauer, P. W. (1990). Recent developments in the econometric estimation of frontiers. *Journal of Econometrics*, 46, 39–56.
- Bethea, R., & Rhinehart, R. (1991). *Applied engineering statistics* (p. 312). Marcel Dekker Inc.
- Box, G. E. P., and Jenkins, G. (1974). *Time series analysis, forecasting and control*, Holden-Day, San Francisco.
- Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.
- Buishand, T. A. (1984). Tests for detecting a shift in the mean of hydrological time series. *Journal of Hydrology*, 73, 51–69.
- Buishand, T. A. (1982). Some methods for testing the homogeneity of rainfall records. *Journal of Hydrology*, 58, 11–27.
- Douglas, E. M., Vogel, R. M., & Kroll, C. N. (2000). Trends in flood and low flows in the United States: Impact of spatial correlation. *Journal of Hydrology*, 1–2, 90–105.
- Fernando, D. A. K., & Jayawardena, A. W. (1994). Generation of forecasting of monsoon rainfall data. In: *Proceedings of the 20th WEDC Conference on Affordable Water Supply and Sanitation*, Colombo, Sri Lanka (pp. 310–313).
- Feller, W. (1968). *An introduction to probability theory and its applications* (Vol. I, 3rd ed.). John Wiley and Sons. Co.
- Hazen, A. (1914). Storage to be provided in impounding reservoirs for municipal water supply. *Trans. ASCE*, 77, 1308.
- Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116, 70–808.
- Hurst, H.E. (1956). Methods of using long-term storage in reservoirs. In: *Proceedings of the Institution of Civil Engineers, Part I*, (pp. 519–542).
- Jayawardena, A. W., & Lai, F. (1989). Time series analysis of water quality data in Pearl river, China. *Journal of Environmental Engineering*, 115(3), 590–607. ASCE.
- Jayawardena, A. W. & Lau, W. H. (1990). Homogeneity tests for rainfall data. *Journal of the Hong Kong Institution of Engineers*, 22–25.
- Kite, G. (1989). Use of time series analyses to detect climatic change. *Journal of Hydrology*, 111, 259–279.
- Maidment, D. R., & Parzen, E. (1984). Time patterns of water use in six Texas cities. *Journal of Water Resources Planning and Management*, 110(1), 90–106. ASCE.
- Mandelbrot, B. B., and Wallis, J. R. (1968). Noah, Joseph and operational hydrology. *Water Resources Research*, 4(5), 909–918.
- Mandelbrot, B. B., & Wallis, J. R. (1969a). Computer experiments with fractional Gaussian Noises. Part I—Averages and variances. *Water Resources Research*, 5(1), 228–241.
- Mandelbrot, B. B., & Wallis, J. R. (1969b). Some long run properties of geophysical records. *Water Resources Research*, 5(2), 321–340.
- Mandelbrot, B. B., & Wallis, J. R. (1969c). Robustness of the rescaled range R/S in the measurement of non-cyclic long-run statistical dependence. *Water Resources Research*.
- Parzen, E. (1960). *Modern Probability and its applications*. New York: Wiley and Sons.
- Popper, K. (1954). *The logic of scientific discovery*. Routledge is an imprint of the Taylor & Francis group. 494 p. ISBN 0-203-99462-0
- Pugacheva, G., Gusev, A., Martin, I., Schuch, N., & Pankov, V. (2003). 22-year periodicity in rainfalls in littoral Brazil. Geophysical Research Abstracts, EGS - AGU -EUG Joint Assembly, Abstracts from the meeting held in Nice, France, April 6–11, 2003, 6797.
- Ripple (1881) Diagram for storage capacity determination.

- Russel, B. (1948). *Human knowledge: Its scope and limits*. London: George Allen and Unwin.
- Şen Z. (1974). Small sample properties of stationary stochastic processes and hurst phenomenon in hydrology. Unpublished Ph. D. Thesis, University of London, 256 pp.
- Şen, Z. (1977). Run-sums of annual flow series. *Journal of Hydrology*, 35, 311–324.
- Şen, Z. (2012). Innovative trend analysis methodology. *Journal of Hydrologic Engineering*, 17(9), 1042–1046.
- Şen, Z. (2014). Trend identification simulation and application. *Journal of Hydrologic Engineering*, 19(3), 635–642.
- Şen, Z. (2015). *Drought modeling, prediction and mitigation* (p. 396). : Elsevier.
- Weiner, N. (1949). *Extrapolation, interpolation and smoothing of stationary time series*. New York: Wiley.
- Wilhite, D. A. (1993). The enigma of drought. In D. A. Wilhite (Ed.), *Drought assessment, management and planning: Theory and case studies*. Boston: Kluwer Academic Publishers.
- Zadeh, L. A. (1965). Fuzzy Sets. *Information Control*, 8, 338–353.

Innovative Trend Methodologies in Science and
Engineering

Şen, Z.

2017, XIII, 349 p. 163 illus., 51 illus. in color., Hardcover

ISBN: 978-3-319-52337-8