Springer Transactions in Civil and Environmental Engineering

Rajib Maity

Statistical Methods in Hydrology and Hydroclimatology

EXTRAS ONLINE

Springer

Springer Transactions in Civil and Environmental Engineering

More information about this series at http://www.springer.com/series/13593

Rajib Maity

Statistical Methods in Hydrology and Hydroclimatology

Springer

Rajib Maity
Department of Civil Engineering
Indian Institute of Technology Kharagpur
Kharagpur
India

Additional material to this book can be downloaded from http://extras.springer.com.

ISSN 2363-7633 ISSN 2363-7641 (electronic)
Springer Transactions in Civil and Environmental Engineering
ISBN 978-981-10-8778-3 ISBN 978-981-10-8779-0 (eBook)
https://doi.org/10.1007/978-981-10-8779-0

Library of Congress Control Number: 2018934932

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
part of Springer Nature
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Dedicated to my
Parents
and my wife
Mitali

Preface

In the areas of hydrology and hydroclimatology, usage of different statistical methods is inevitable due to inherent uncertainty. Hydrology and climatology are two areas of science that involve studies related to hydrologic and climatic systems/subsystems, respectively. In connection with the climate change and its impacts on water resource engineering, hydrologic and hydroclimatic problems are now being addressed hand in hand. Random variability of hydrologic variables has a long history since its recognition, and several statistical techniques are currently in use. Further, the correspondence between climatic variability and hydrologic variability has produced a relatively new interdisciplinary field, known as hydroclimatology. It provides a platform to analyze the relationship between climatic factors and hydrologic variables over space and time. Spatio-temporal evolution of such relationship is essential in the context of climate change. Several statistical methodologies are currently being developed and introduced in this subject area to tackle new emerging challenges.

This book focuses on a wide range of statistical methods ranging from fundamental concepts to advanced theories that are found to be potential and essential to deal with the real-life problems in the fields of hydrology and hydroclimatology. Besides other advanced theories, the book also introduces the theory of copulas and its applications in a chapter with many illustrative examples and MATLAB-based small codes to deal with the problems and solutions in hydrology and hydroclimatology.

Part of the book is intended to serve as a textbook for graduate courses on stochastic methods in hydrology and related disciplines. The book may also be a valuable resource for researchers, professionals, and doctorate students in the areas of hydrology, hydroclimatology, and related fields. This book is broadly organized as follows: Chapter 1 provides a basic introduction on the subject area and role of statistical methods in it. Chapters 2 and 3 are introductory in nature and present a thorough discussion on the basic concepts of random experiment, random variables, and some basic exploratory statistical properties. Chapter 4 provides mathematical and conceptual foundations of commonly used probability distributions in the domains of hydrology and hydroclimatology. Chapter 5 deals with frequency analysis, risk, and uncertainty in hydroclimatic analysis. Hypothesis testing and nonparametric tests are discussed in Chap. 6. Regression analysis and multivariate analysis including ANOVA and wavelet analysis are covered in Chaps. 7 and 8, respectively. Chapter 9 presents the concepts of hydroclimatic time series analysis and forecasting including stationarity, homogeneity, periodicity. Chapter 10 portrays the potential of copula theory in hydrology and hydroclimatology. Copulas help to develop the joint distribution between multiple associated hydroclimatic variables. Its potential in frequency analysis, multivariate modeling, simulation, and prediction is discussed for hydroclimatic problems.

Kharagpur, India Rajib Maity

Acknowledgements

There are many individuals who directly or indirectly contributed to this book. It starts with many professors and academicians in India and abroad with whom I interacted through various collaborations.

I would also like to acknowledge the support offered by many research students. Specifically, help from Ph.D. students—Mayank Suman and Subharthi Sarkar—and MS student Riya Dutta is highly appreciated and acknowledged. Technical help from Mayank is also acknowledged. Support from Subbarao Pichuka, A Naren and indirect support from Manali Pal and Subhasmita Dash are also acknowledged.

Finally, it is my wife Mitali, a friend, philosopher, and guide, who is always with me in all my ups and downs. To state the least, I must mention her dedication, sacrifice, and love without which everything is meaningless.

Contents

About the Author

Dr. Rajib Maity is Associate Professor in the Department of Civil Engineering, Indian Institute of Technology Kharagpur, India. His research areas include hydroclimatology, stochastic hydrology, climate impacts on water resources, hydrologic time series analyses and forecasting. He has published a book on ‘Hydroclimatic Teleconnection: Indian Perspective,’ several chapters, and over 80 research articles in various peer-reviewed journals and conferences. His research work has been funded by various agencies such as the Department of Science and Technology (DST), Indian Space Research Organisation (ISRO), Ministry of Earth Sciences (MoES), Ministry of Human Resource Development (MHRD), Australia-India Strategic Research Fund (AISRF), and IBM. Some of his professional awards/honors include Humboldt Fellowship (experienced category) from Alexander von Humboldt Foundation (Germany), James Rennell MoES Young Fellowship (MoES), the Prof. R. J. Garde Research Award, ASCE 2011 Outstanding Reviewer (USA), Emerging Leaders Fellowship (Australia), BOYSCAST Fellowship (India/USA), IEI Young Engineers Award, DAAD Fellowship for IIT faculty (Germany), International ICE WaRM Fellowship (Australia), and Prof. N. S. Govinda Rao Memorial Gold Medal, IISc. He is also currently serving as an Associate Editor of the Journal of Earth System Science (JESS), Springer, and ISH Journal of Hydraulic Engineering, Taylor and Francis.

Chapter 1
Introduction

It is oblivious to state the need of statistical methods in any field of engineering and science. In the area of hydrology and hydroclimatology, use of different statistical methods is inevitable due to inherent uncertainty. This chapter starts with some basic definitions and scope in hydrology, climatology, and hydroclimatology. Role of statistical methods in the context of inherent variability and uncertainty is discussed afterward. Organization of the book is also presented at the end of this chapter.

1.1 Definitions and Scope

Hydrology is the science that involves studies related to occurrence and movement of water (in any phase of solid, liquid, or vapor) in the combined system of surface, subsurface, and atmosphere. Hydrologic cycle, also known as water cycle, is the basis of the hydrologic science. It offers a platform to manage the available water in the context of water use, water control, and water pollution.

Climatology is the field of study related to exchange of mass, momentum, and energy between land/ocean surface and atmosphere. Vertical and horizontal fluxes of these quantities drive the interaction between earth surface (both land and ocean) and atmosphere. These fluxes also control the atmospheric circulation at different scales. Atmospheric component of hydrologic cycle is coupled with climatic phenomena, and thus, any change or variability may affect each other through different feedback systems.

Hydroclimatology is an interdisciplinary area of study that deals with the interaction between hydrology and climatology to identify the influence of the climatic system on different hydrologic processes, which are the parts of hydrologic cycle. For example, hydrologic variables, such as rainfall, soil moisture, streamflow, etc., are significantly influenced by various global or local scale atmospheric circulations. In the context of climate change, role of hydroclimatic studies has become crucial

R. Maity, *Statistical Methods in Hydrology and Hydroclimatology*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-10-8779-0_1

in many applications. In general, hydroclimatology provides a platform to analyze the relationship between climatic factors and hydrologic variables over space and time. Such relationship and their possible changes vary over time and space and are essential in the context of climate change.

1.2 Role of Statistical Methods

1.2.1 Hydrologic and Hydroclimatic Variability

Hydrologic and climatic systems, and their combination, i.e., hydroclimatic systems, consist of several interrelated processes. Such processes are not amenable to deterministic analysis. In most of the cases, if not all, hydrologic and hydroclimatic variables are associated with randomness/uncertainty and should be treated as random variables. Examples include peak discharge, streamflow, annual maximum rainfall, number of rainy days, etc. It is rather hard to identify any hydroclimatic variable that is free from any randomness.

1.2.2 Need of Statistical Methods

Statistical methods deal with the uncertainty and provide the ways to take practical decisions or choosing mitigation strategies. Role of statistical methods in the context of uncertainty includes evaluation and quantification of uncertainty, making inferences based on the available data, frequency analysis, forecasting, and so on.

Need of statistical methods in hydrology was felt long back. Recently, in the context of climate change and its possible impact on hydrology and water resources engineering, statistical methods are inevitable. In general, numerous hydrologic and hydroclimatic variables are associated with each other. Several considerations come into play for the development of statistical models. These include the nature of the associated variable(s) (precipitation, temperature, streamflows, storage levels, etc.), data availability, scale of analysis.

The utility of statistical methods in analysis of hydroclimatic systems is beneficial for understanding the interrelated processes involved and to perform risk and vulnerability analysis. Though there is a plethora of deterministic models available, yet the presence of several source and types of uncertainties associated with spatial and temporal variability in hydroclimatic variables demands statistical methods. However, most of the statistical methods, if not all, depend on some parametric assumptions of data sets and the predefined nature of correspondence. Thus, it is essential to extract the characteristics of the data using different statistical tools. Any statistical modeling approach involves exploring the mutual relationship between the input and target hydroclimatic variables.

Prediction of hydroclimatic variables is another important aspect to be accomplished through statistical modeling. Reliable prediction is always helpful in resource management and impact assessment studies in the context of climate change. In general, some of the variables are considered as inputs (also known as predictors or independent variables) from which information is extracted and rest are considered as the response variables (also known as predictands or dependent variables). Sometimes information of the same variable from the previous time steps (lagged values) is also considered in the set of inputs. The role of inputs may vary in both space and time. Traditionally, the selection of predictors has been accomplished by some statistical methods, such as regression or cross-correlation analysis. For instance, monthly streamflow prediction at a basin scale is a challenging problem because of the complex roles of multiple interacting hydroclimatic variables, such as precipitation, evaporation, soil moisture, temperature, pressure, wind speed that directly or indirectly contribute to flow generation. While several target variables, such as rainfall, streamflows are known to depend on various hydroclimatic variables, dependence patterns may not be known with certainty and vary from one basin to another. Statistical methods are required for the competent predictor selection, and it is an important part of the development of effective prediction or simulation models. Apart from selecting variables based on our understanding of the physical system, temporal relations between the predictor set and predictand need to be accounted for using techniques such as time series autocorrelation and partial autocorrelation and/or cross-correlation analysis.

Another issue concerns about numerous hydroclimatic variables that may have possible influence on the target variable at multiple lags, which may yield a prohibitively large number of variables in the predictor set. This leads to *curse of dimensionality* and may pose serious challenges in parameter estimation and lead to a highly complex prediction model. Sometimes it may also be burdened with redundancy in information from multiple inputs. In such situation, some techniques related to multivariate analysis are helpful in prioritizing the relevant features in the set of potential predictor variables. It has several advantages including better understanding of the data and dimensionality reduction of multivariate data to avoid the *curse of dimensionality*. Examples include principal component analysis (PCA), supervised principal component analysis (SPCA), canonical correlation analysis (CCA), empirical orthogonal function (EOF) analysis, analysis of variance (ANOVA).

A substantial impact on the available water resources due to climate change is realized almost everywhere across the world. Such impacts may vary spatio-temporally that influence the characteristics of the extreme events, such as droughts and floods, including number, magnitude, severity, duration. Spatio-temporal variation in any hydroclimatic variables may cause spatio-temporal variation in other associated hydroclimatic variables also. The characteristics of hydrologic extreme events are influenced by triggers that may be manifested in specific patterns of hydroclimatic variables. Identification of these triggers also requires statistical methods for devising effective mitigation plans against extreme phenomena.

The development of joint probability distribution among the associated hydroclimatic variables is needed in many modeling schemes. It may be noted that multivariate Gaussian distribution ensures that the marginal distribution of each of the associated variables is normally distributed. However, reverse is not true; i.e., when the distributions of all the associated variables are normal, joint distribution is not necessarily multivariate Gaussian. In general, even though the marginal distributions of each of the associated variables are known, their joint distributions may not be easy to derive from these marginal distributions. However, copula can be used to obtain their joint distribution, using scale-free measures of dependence between the variables. *Kendall's tau* and *Spearman's rho* are the most commonly used scale-free measure of association, and these are nonparametric, i.e., free from any specific parametric assumption. In most of the hydroclimatic analysis, some interrelationship among the associated variables may exhibit more prominence as compared to others even though other factors may influence the target variable. For example, rainfall and runoff may exhibit more prominent association but other hydroclimatic variables, such as spatial variation of soil moisture, may also influence the runoff generation. In such cases, multivariate copulas are helpful. In some cases, combination of several statistical methods is also found beneficial, for instance, extraction of principal components from the set of input variables and then application of copulas using the principal components as inputs.

In brief, probabilistic assessment in the field of hydrology and hydroclimatology is unavoidable. This requires a thorough knowledge on wide range of statistical tools from basics to advanced theories and their applications.

1.3 Organization of the Book

Keeping all the aspects in consideration as discussed in the last section, the book is organized in such a way the readers will build up their knowledge from basic concept to advanced theories and apply to the real-life hydrologic and hydroclimatic problems and interpret the results. It starts with some basic concepts of probability and statistics (Chaps. 2 and 3). All the statistical methods discussed in the subsequent chapters require in-depth knowledge of probability theory. Chapter 2 presents a thorough discussion on the basic concepts of random experiment, random variables, events, and assignment of probability to events with relevant examples. Chapter 3 starts with some basic exploratory statistical properties, which is the first step of any statistical method to be attempted. Concept of moment and expectation, moment generating, and characteristic functions is considered afterward. Different methods for parameter estimation build the foundation for many statistical inferences in the field of hydrology and hydroclimatology.

As mentioned before, presence of uncertainty is unavoidable in any hydrologic and hydroclimatic variable. First step to deal with it is to probabilistically represent the data using different probability distributions. In Chap. 4, commonly used distributions with their parameters, properties of the distribution supported by graphical representation, and their plausible applications in hydrology and hydroclimatology are explained. Discussion on each distribution is presented in the order of their basics, interpretation of the random variable, parameters, probability mass/density function, description, potential applications, and illustrative examples. This order is expected to help the readers to understand the distribution and to develop the knowledge base for its further applications.

Frequency of extreme events like severe storms, floods, droughts is an essential component of hydrology and hydroclimatology. In the context of climate change, such events are found to occur more frequently. It is oblivious to state that more extreme events have catastrophic impact on the entire agro-socioeconomic sector of the society. Chapter 5 deals with frequency analysis, risk, and uncertainty in hydroclimatic analysis.

Hypothesis testing and nonparametric tests are discussed in Chap. 6. Available data is generally limited in the domain of hydrology and hydroclimatology. Hypothesis testing is useful to assess the changes that might have occurred owing to climate change. It helps to make statistical inferences about some parameter of the population based on the available data. Nonparametric tests also help to assess the change in the data over time or space using the concept of hypothesis testing. Such tests are useful in absence of long data and/or if the available data does not fit any known and commonly used distribution.

Rest of the book covers the modeling of relationship/association/dependence between the associated variables. Many applications in hydrology and hydroclimatology, such as simulation, prediction, depend on the relationship between the associated variables. In Chap. 7, the procedure of developing such relationship between dependent and independent variables through regression analysis and curve fitting is discussed. Multivariate analysis techniques are taken up next in Chap. 8, since it is often noticed that many hydroclimatic variables are associated with each other. Generally such associations are complex and are required to be analyzed simultaneously using multivariate hydroclimatic analysis.

Hydroclimatic time series vary with space and time due to continuously evolving nature of hydroclimatic variables. The objective of Chap. 9 is to introduce different types of time series analysis techniques. This requires an understanding of time series analysis techniques and time series properties like stationarity, homogeneity, periodicity, which is the subject matter of this chapter.

Chapter 10 portrays the potential of copula theory in hydrology and hydroclimatology. This chapter starts with an introduction to basic concept and the theoretical background. Copulas help to develop the joint distribution between multiple variables that are associated with each other. Basic mathematical formulations for most commonly used copulas are discussed and illustrative examples are provided. Its potential in frequency analysis, multivariate modeling, simulation, and prediction is discussed for hydroclimatic problems.

Throughout the book, the illustrative examples are of three types – (i) with very small data showing the calculations very clearly so that readers can get an idea on the computing procedure, (ii) with sufficiently large data so that the results can be interpreted and the theory can be applied to other similar problems, and

(iii) with real data and computer code (MATLAB platform). The illustrative examples with very few data points help to show the calculation steps explicitly. Please note that any statistical analysis should be done with sufficiently long data. Once the readers understand the steps, computer codes can be written easily for large data sets. Examples of MATLAB codes are also provided at the end of each chapter.

Chapter 2
Basic Concepts of Probability and Statistics

Probability is the measure of chance of occurrence of a particular event. The basic concept of probability is widely used in the field of hydrology and hydroclimatology due to its stochastic nature. The inferences like the expected frequency of events, prediction of hydrologic phenomena based on the dependent variables, risk assessment and modeling require in-depth knowledge of probability theory. This chapter starts with the basic concepts of probability that is required for a clear understanding of random experiment, random variables, events, and assignment of probability to events. The axioms of probability and the fundamental rules are explained with the help of Venn diagrams. Later, the concepts of univariate and bivariate random variables along with their respective forms of probability distribution function, cumulative distribution function, and joint probability distribution are discussed. Application of the probability theories in the field of hydrology and hydroclimatology is illustrated with different examples.

2.1 Concepts of Random Experiments and Random Variables

2.1.1 Random Experiments, Sample Space, and Events

An experiment is a set of conditions under which behavior of some variables is observed. Random experiment is an experiment, conducted under certain conditions, in which the outcome cannot be predicted with certainty. Each run of a random experiment is generally referred as a *trial*. Possible outcome(s) of each trial varies (vary); reason to call it *random*. In the domain of hydrology and hydroclimatology, counting the number of rainy days in a particular month (say June), measuring the rainfall depth, soil moisture content, wind speed, etc., are the few examples of random experiments.

R. Maity, *Statistical Methods in Hydrology and Hydroclimatology*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-10-8779-0_2

All possible outcomes of a random experiment constitute *sample space*, and each outcome is called a *sample point*. For example, for the random experiment, ‘*counting the number of rainy days in June*’, the sample space consists of only integers from 0 to 30. Outcome of ‘*measuring the rainfall depth*’, ‘*soil moisture content*’, or ‘*wind speed*’ at a location may take any nonnegative values. Thus, the sample space of these random experiments consists of any real number in the range of 0 to ∞.

The *sample space* can be classified either as *discrete* or *continuous* sample space. A *sample space* is discrete if it has finite or countably infinite elements. For example, the sample space of the random experiment, ‘*counting the number of rainy days in June*’, consists of discrete numbers only (0–30). This is an example of discrete sample space that contains finite number of elements. Another example of ‘*inter-arrival time (in days) between two rainfall events*’ is also a discrete sample space. However, this sample space contains countably infinite elements. On the other hand, the sample space that consists of a continuum, i.e., all possible values within a range of real numbers, is known a continuous sample space. The sample spaces of ‘*measuring the rainfall depth*’, ‘*soil moisture content*’, or ‘*wind speed*’ at a location are the examples of continuous sample space that consist of any real number in the range of 0 to ∞.

An *event* can be defined as a subset of a sample space. Event may consist of a single/multiple sample points (discrete sample space) or a range from the sample space (continuous sample space). Number of rainy days in June equal to 10 is an example of *event* from the sample space of ‘*counting the number of rainy days in June*’. Similarly, wind speed greater than 100 km/h is an event from the continuous sample space of ‘*wind speed*’ at a location.

2.1.2 Concept of Random Variables and Events

According to classical concept, a random variable (RV) is a function that maps each outcomes of an experiment over a sample space to a numerical value on the real line (Fig. 2.1). Thus a RV is not a variable, rather a function. A random variable is generally denoted by an uppercase letter, say X, and the corresponding lowercase letter, i.e., x is used to represent a specific value of that random variable. The convention of course varies; however, it will be uniformly followed in this book. Thus, X denotes a random variable and x denotes a specific value of the random variable X. Random variable may be discrete or continuous depending on the nature of the associated sample space. The random variable associated with a discrete (continuous) sample space is a discrete (continuous) random variable. Thus, if the set of values a random variable can assume is finite or countably infinite, the random variable is said to be *discrete random variable*. If the set of values a random variable can assume is a continuum, i.e., all possible values within a range of real numbers, then the random variable is known as *continuous random variable*. An example of a discrete random variable would be the ‘*number of rainy days in June*’ at a particular location, whereas ‘*the rainfall depth*’ at a location is a continuous random variable. Any function of a

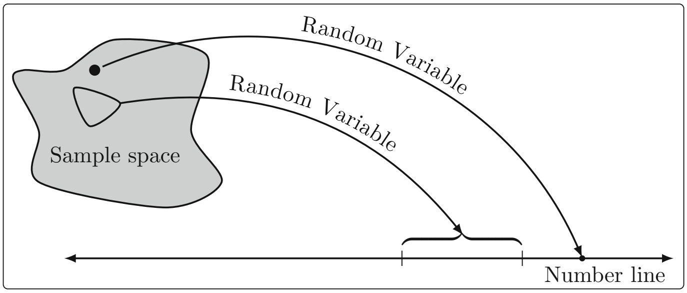


Fig. 2.1 Representation of random variable

random variable is also a random variable. If X is a random variable, then $Z = g\,(X)$ is also a random variable.

Since the subset of a sample space forms an event, a specific value or a range of values of a random variable is also an event. For example, $X = 3$, $X \geq 5$, $0 \leq X \leq 50$ are the examples of events of the random variable X. Probability is assigned to the events, and this assignment of probability to events is the key for any probabilistic assessment. It requires the concept of set theory that includes the inter-relationships between events, such as *union* (symbolized as $A \cup B$), *intersection* (symbolized as $A \cap B$ or AB), and *complement* (symbolized as A^c). It is expected that the readers are well aware of these concepts. Graphical representation of sample space, events, and their inter-relationships is generally depicted by *Venn diagram*. A typical Venn diagram showing *sample space* (S), *events* (E_1, E_2, … etc.), and their *inter-relationships* is shown in Fig. 2.2. Further details can be referred to any basic book on probability and statistics.

Mutually Exclusive Events

Two events E_1 and E_2 are called mutually exclusive when none of the outcomes in E_1 belongs to E_2 or vice versa. This is denoted as: $E_1 \cap E_2 = \phi$, where ϕ indicates a null set. In Fig. 2.2a, mutually exclusive events are shown by no overlap between them.

Collectively Exhaustive Events

When union of all events (E_1, $E_2, \ldots E_n$) comprise the whole sample space, 'S', then $E_1, E_2, \ldots, E_n$ are called collectively exhaustive events. This is denoted as: $E_1 \cup E_2 \cup \cdots \cup E_n = S$. However, the *intersection* of any two events need not be null set.

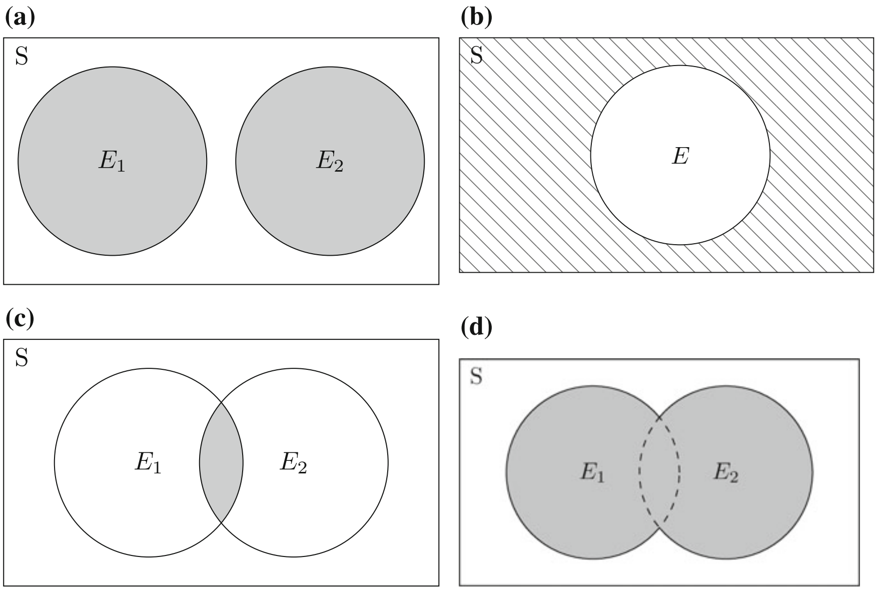
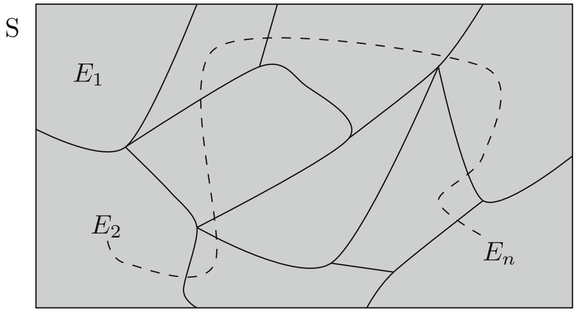


Fig. 2.2 Venn diagrams showing *sample space*, *events* (E, E_1, E_2), and their *inter-relationships*: **a** The events E_1 and E_2 *are mutually exclusive*; **b** the hatched area is *complement* of event E; **c** the shaded area is *intersection* of events E_1 and E_2; and **d** the shaded area is *union* of events E_1 and E_2

Mutually Exclusive and Collectively Exhaustive Events

When the entire sample space is partitioned by n different events in such a way that intersection between any two of them is a null set and the union of all the events forms the entire sample space, the events are known as mutually exclusive and collectively exhaustive events. It is denoted as: $E_1 \cup E_2 \cup \cdots \cup E_n = S$ where $E_i \cap E_j = \phi$ for $\forall\, i \neq j$. The Venn diagram is shown in Fig. 2.3.

Fig. 2.3 Venn diagram showing mutually exclusive and collectively exhaustive events



In hydrology and hydroclimatology, the categorization of any variables into different groups is the example of mutually exclusive and collectively exhaustive events. For example, daily rainfall depth (X in mm) can be grouped as $X = 0\,\text{mm}$, $0\,\text{mm} < X \le 5\,\text{mm}$, $5\,\text{mm} \le X \le 10\,\text{mm}$, and $X \ge 10\,\text{mm}$. These events are mutually exclusive and collectively exhaustive events.

2.2 Basic Concepts of Probability

In any random experiment, there is always uncertainty as to whether a specific event will occur or not. The probability concept was proposed originally to explain the uncertainty involved in the outcome of a random experiment. Probability is assigned to the events, and this assignment of probability to events is the key for any probabilistic assessment.

As a measurement of the chance or probability, with which an event can be expected to occur, it is convenient to assign a number between 0 and 1. According to the classical definition, the probability of an event A, denoted as $P\,(A)$, is determined a priori without actual experimentation. It is given by the ratio

$$P(A) = \frac{N_A}{N} \tag{2.1}$$

where N is the number of possible outcomes and N_A is the number of outcomes that are favorable to the event A.

The definition of probability in hydrology and hydroclimatology is more effectively expressed in terms of relative frequencies. If a random event occurs a large number of times N and the event A occurs in n of these occurrences, then the probability of the occurrence of the event A is:

$$P(A) = \lim_{N\to\infty} \frac{n}{N} \tag{2.2}$$

2.2.1 *The Axioms of Probability*

Probability of any event A in a sample space S, denoted as $P(A)$, is assigned in such a way that it satisfies certain conditions. These conditions for assigning probability are known as *axioms of probability*. There are three such axioms defined as follows.

Axiom 1: $0 \le P(A) \le 1$ for each event A in S.
This states that the probabilities are real numbers in the interval from 0 to 1, including the boundary, i.e., 0 and 1.

Axiom 2: $P(S) = 1$
This states that the sample space as a whole is assigned a probability of 1. Since S contains all possible outcomes, one of these must always occur.
Axiom 3: If A and B are mutually exclusive events in S, then $P(A \cup B) = P(A) + P(B)$
This states that the probability functions must be additive, i.e., the probability of union is the sum of the two probabilities when the two events have no outcome in common. All conclusions drawn on probability theory are either directly or indirectly related to these three axioms.

2.2.2 *Some Elementary Properties on Probability*

From the axioms of probability, some elementary property can be proved that are important in further work.

Property 1: If $E_1, E_2, \ldots, E_n$ are mutually exclusive events, then probability of union of all these events is equal to summation of probability of individual events. This is mathematically denoted as,

$$P(E_1 \cup E_2 \cup \ldots \cup E_n) = P(E_1) + P(E_2) + \ldots + P(E_n) \qquad (2.3)$$

This is basically the extension of Axiom 3, considering any number of mutually exclusive events. This is known as property of finite additivity.
Property 2: If an event E_2 belongs to another event E_1, then probability of E_2, $P(E_2)$ will be less than or equal to probability of E_1, $P(E_1)$. And the probability of difference between these two events, $P(E_1 - E_2)$, will be equal to the difference between probability of E_1 and E_2, i.e., $P(E_1)$ and $P(E_2)$. In other words, if $E_2 \in E_1$, then $P(E_2) \leq P(E_1)$ and $P(E_1 - E_2) = P(E_1) - P(E_2)$. The visualization is given in Venn diagram (Fig. 2.4).

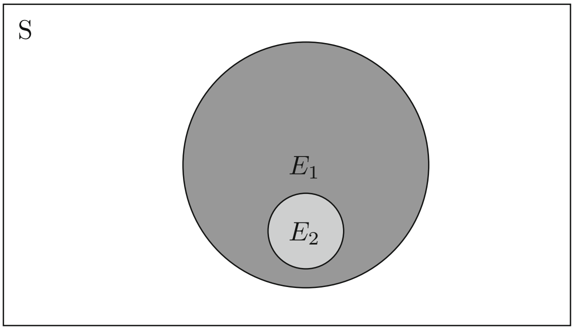


Fig. 2.4 Venn diagram related to elementary *Property 2*

Property 3: If any event E_1 is complementary to another event E_2, then probability of E_1 can be determined using probability of E_2 from *Axiom 1*. This is mathematically denoted as, if $E_1 = E_2^c$ then $P(E_1) = 1 - P(E_2)$.

Property 4: If an event, E is the union of events $E_1, E_2, \ldots, E_n$, where E_1, $E_2, \ldots, E_n$ are mutually exclusive, then probability of E is the summation of probability of each of these events. This is mathematically denoted as,

$$P(E) = P(E_1) + P(E_2) + \ldots + P(E_n) \tag{2.4}$$

The visualization is given in Fig. 2.3.

Property 5: For any two events, E_1 and E_2 that belong to sample space S, probability of E_1, $(P(E_1))$ can be determined as summation of probability of E_1 intersection E_2 and the probability of E_1 intersection the complement of E_2. It is mathematically denoted as:

$$P(E_1) = P(E_1 \cap E_2) + P\left(E_1 \cap E_2^c\right) \tag{2.5}$$

Property 6: If E_1 and E_2 are any two events in sample space, S, then probability of union of E_1 and E_2 can be determined by deducting the probability of intersection of E_1 and E_2 from the summation of their individual probabilities. It is mathematically denoted as:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) \tag{2.6}$$

The visualization is shown in Fig. 2.2d. This property can be proved using axioms and other properties. From Fig. 2.2d, considering the different parts of the shaded areas,

$$\begin{aligned} P(E_1 \cup E_2) &= P(E_1 \cap E_2) + P\left(E_1 \cap E_2^c\right) + P\left(E_1^c \cap E_2\right) \\ &= \left[P(E_1 \cap E_2) + P\left(E_1 \cap E_2^c\right)\right] + \left[P(E_1 \cap E_2) + P\left(E_1^c \cap E_2\right)\right] - P(E_1 \cap E_2) \\ &= P(E_1) + P(E_2) - P(E_1 \cap E_2) \end{aligned} \tag{2.7}$$

Extending this property, if E_1, E_2, and E_3 are any three events,

$$\begin{aligned} P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_2 \cap E_3) \\ - P(E_3 \cap E_1) + P(E_1 \cap E_2 \cap E_3) \end{aligned} \tag{2.8}$$

This can be visualized graphically in Fig. 2.5.

Property 7: For mutually exclusive and collectively exhaustive events, E_1, $E_2, \ldots, E_n$ in the sample space S, the probability of another event E is equal to the sum of probabilities of intersections between E and each of the event $E_1, E_2, \ldots, E_n$. It is mathematically expressed as,

$$P(E) = P(E \cap E_1) + P(E \cap E_2) + \ldots + P(E \cap E_n) \tag{2.9}$$

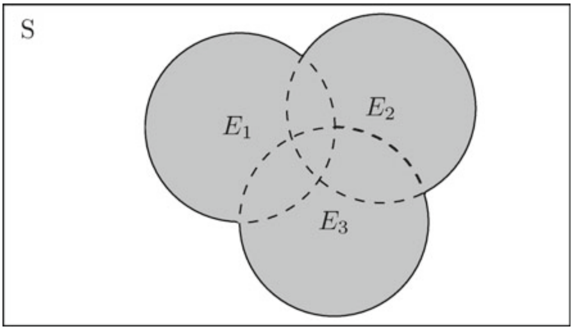


Fig. 2.5 Venn diagram related to *Property 6* for three events

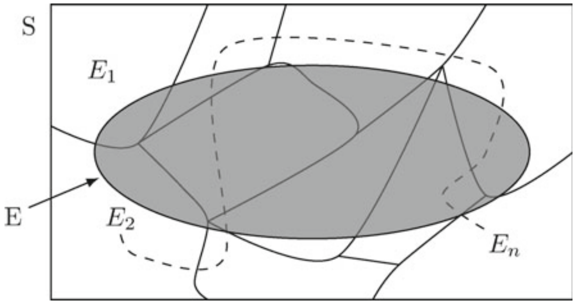


Fig. 2.6 Venn diagram related to elementary *Property 7*

The visualization is presented in Fig. 2.6.

Example 2.2.1
A field is irrigated using the supply from either of canal water or groundwater or rainfall. At any given time, the probability of failure due to inadequate supply of water from at least one of these sources is 0.4. Assuming that the probability of failure of canal supply, groundwater, and rainfall individually are 0.2, 0.05, and 0.25, respectively, information on their simultaneous failures is as follows:

(a) Probability of simultaneous failure of canal supply and groundwater is 0.1.
(b) Probability of simultaneous failure of groundwater and rainfall is 0.01.
(c) Probability of simultaneous failure of canal supply and rainfall is 0.3.

What is the probability of simultaneous failure of all the sources?

Solution Let us denote,

E_1 = Failure of canal supply;
E_2 = Failure of groundwater source;
E_3 = Failure of rainfall source (no rainfall occurs).

Thus,

$$
\begin{aligned}
&P\left(E_1\right)=0.2\\
&P\left(E_2\right)=0.05\\
&P\left(E_3\right)=0.25\\
&P\left(E_1\cap E_2\right)=0.1\\
&P\left(E_2\cap E_3\right)=0.01\\
&P\left(E_1\cap E_3\right)=0.3\\
&P\left(E_1\cup E_2\cup E_3\right)=0.4
\end{aligned}
$$

Thus, from Property 6,

$$
\begin{aligned}
P\left(E_1\cup E_2\cup E_3\right)=&P\left(E_1\right)+P\left(E_2\right)+P\left(E_3\right)-P\left(E_1\cap E_2\right)-P\left(E_2\cap E_3\right)\\
&-P\left(E_3\cap E_1\right)+P\left(E_1\cap E_2\cap E_3\right)\\
\Rightarrow 0.4=&0.2+0.05+0.25-0.1-0.01-0.3+P\left(E_1\cap E_2\cap E_3\right)
\end{aligned}
$$

or, $P\left(E_1\cap E_2\cap E_3\right)=0.31$

Thus, the probability of failure of all the sources is 0.31.

2.3 Conditional Probability Theorem

If A and B are two events in a sample space S, and $P(B)\neq 0$, the conditional probability of B given that A has already occurred is obtained as the ratio of probability of intersection of A and B, and probability of A. It is mathematically expressed as,

$$P\left(B/A\right)=\frac{P\left(A\cap B\right)}{P\left(A\right)} \tag{2.10}$$

For any three events E_1, E_2, and E_3, the probability that all of them occur is the same as the probability of E_1 times probability of E_2 given that E_1 has occurred times the probability of E_3 given that both E_1 and E_2 have occurred. It is mathematically expressed as,

$$P\left(E_1\cap E_2\cap E_3\right)=P\left(E_1\right)P\left(E_2/E_1\right)P\left(E_3/E_1\cap E_2\right) \tag{2.11}$$

This theorem can be generalized for any n number of events $E_1, E_2, \ldots, E_n$.

Example 2.3.1
Daily rainfall records are obtained from two rain gauge stations A and B, located 150 km apart. The probability of occurrence of wet day (rainfall > 2.5 mm/day) at each station is 0.1. However, the probability of occurrence of wet day at one station, given that the other station experience wet day is 0.80. What is the probability of occurrence of wet day either at station A or B?

Solution Let the event A denote the wet day at station A and the event B denote wet day at station B. The probability of occurrence of wet day at either station A or B is the union of the events A and B. Using property 6,

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A)P\left(B/A\right) \\ &= 0.1 + 0.1 - 0.1 \times 0.8 \\ &= 0.12 \end{aligned}$$

Example 2.3.2
The probabilities that the rain gauge instruments at stations A and B will function uninterruptedly for 20 months are 0.8 and 0.9, respectively. Proper functioning of the rain gauge instruments is independent. Find the probability that in 20 months (a) both, (b) neither (c) at least one, will be in function.

Solution Considering A and B are the events that the rain gauges instruments function uninterruptedly for 20 months at stations A and B, respectively.

Thus, $P(A) = 0.8$ and $P(B) = 0.9$. Since the events A and B are independent,

(a) P(both will be in function) = $P(A \cap B) = P(A)P(B) = 0.8 \times 0.9 = 0.72$.
(b) P(neither will be in function) = $P(A^c \cap B^c) = P(A^c)P(B^c) = (1 - 0.8) \times (1 - 0.9) = 0.02$.
(c) P(at least one will be in function) = $1 - P$(neither will be in function) = 0.98.

Example 2.3.3
The probability of occurrence of rainfall on a particular day in monsoon is 0.4. The probability of occurrence of rainfall on two consecutive days is 0.1. What is the probability of occurrence of rainfall on 26th July given that the rainfall occurred on 25th July?

Solution Let X and Y be the event of occurrence of rainfall on 25th and 26th July, respectively.

$$P\left(Y/X\right) = \frac{P\left(X \cap Y\right)}{P\left(X\right)} = \frac{0.1}{0.4} = 0.25$$

Hence, probability of rainfall on 26th July given that rainfall occurs on 25th July is 0.25.

Example 2.3.4
An embankment may fail either due to releasing the excess water from the upstream reservoir or due to the heavy rainfall or due to their simultaneous occurrences. The probability of failure due to excess water release from upstream reservoir is 0.01, and the same due to heavy rainfall is 0.08. However, probability of failure of embankment due to excess release during heavy rainfall is quite high and estimated as 0.5. Determine

(a) The probability of failure of the embankment.
(b) The probability that the failure due to heavy rainfall only (no excess release from upstream reservoir).

Solution Let E and R represent the events of failure due to excess water release from upstream reservoir and due to heavy rainfall, respectively.

$$P\left(E\right) = 0.01,\ P\left(R\right) = 0.08 \text{ and } P\left(E/R\right) = 0.5$$

(a) Probability of failure of the embankment is given as,

$$\begin{aligned} P(F) = P(E \cup R) &= P(E) + P(R) - P(E \cap R) \\ &= P(E) + P(R) - P\left(E/R\right) P(R) \\ &= 0.01 + 0.08 - 0.5 \times 0.08 \\ &= 0.05 \end{aligned}$$

(b) The probability that the failure due to heavy rainfall only (no excess release from upstream reservoir) is given as,

$$\begin{aligned} P(R \cap E^c) &= P(E^c/R)P(R) \\ &= [1 - P(E/R)]P(R) \\ &= (1 - 0.5) \times (0.08) \\ &= 0.04 \end{aligned}$$

Example 2.3.5
There are several industries located on the bank of a river. It is observed that the wastes from those industries are mixing in the river without proper treatment. The water samples are collected every day from two different sections 1 and 2 on the river to check the pollution level. Let X denotes the event that pollution is detected at section 1 and Y denotes the same for section 2. Following information is obtained from laboratory test: $P(X) = 0.158$, $P(Y) = 0.25$ and the probability that at least one section is polluted on any given day is 0.27. Determine the probability that

(a) Section 1 is polluted given that section 2 is already found polluted.
(b) Section 2 is polluted given that section 1 is already found polluted.

Solution First, the probability of both the reaches are polluted is to be computed

$$\begin{aligned} P(X \cup Y) &= P(X) + P(Y) - P(X \cap Y) \\ \Rightarrow P(X \cap Y) &= P(X) + P(Y) - P(X \cup Y) \\ &= 0.158 + 0.25 - 0.27 \\ &= 0.138 \end{aligned}$$

(a) Probability of reach 1 is polluted given that reach 2 is already found polluted is

$$P(X/Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{0.138}{0.25} = 0.552$$

(b) Probability of reach 2 is polluted given that reach 1 is already found polluted is

$$P(Y/X) = \frac{P(X \cap Y)}{P(X)} = \frac{0.138}{0.158} = 0.873$$

2.4 Total Probability Theorem and Bayes' Rule

Let $E_1, E_2, \ldots, E_n$ represent a set of mutually exclusive and collectively exhaustive events as shown in Fig. 2.3. Also, consider another event A that belongs to the same sample space. The probability of occurrence of the event A depends on the events (E_i) that have already occurred. Probability of the event A can be evaluated using the Property 7 as follows:

$$P(A) = P(A \cap E_1) + P(A \cap E_2) + \ldots + P(A \cap E_n) \tag{2.12}$$

Next, using the conditional probability theorem (Eq. 2.10):

$$P(A/E_i) = \frac{P(A \cap E_i)}{P(E_i)} \tag{2.13}$$

$$\Rightarrow P(A \cap E_i) = P(E_i)\, P(A/E_i) \tag{2.14}$$

Now, $P(A)$ can be evaluated as (from Eqs. 2.12 and 2.14)

$$P(A) = P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + \cdots + P(E_n)P(A/E_n)$$

$$\Rightarrow P(A) = \sum_{i=1}^{n} P(E_i)\, P(A/E_i) \tag{2.15}$$

This is known as the *Theorem of Total Probability*.

Bayes' Rule: Next, if we are interested to know the probability of occurrence of any particular event E_i, given that event A has occurred, conditional probability theorem (Eq. 2.10) can be used to evaluate the same as follows,

$$P\left(A \cap E_i\right) = P\left(E_i \cap A\right) \Rightarrow P\left(E_i\right) P\left(A/E_i\right) = P\left(A\right) P\left(E_i/A\right)$$

Therefore, the desired probability is,

$$P\left(E_i/A\right) = \frac{P\left(E_i\right) P\left(A/E_i\right)}{P\left(A\right)} \tag{2.16}$$

Utilizing the total probability theorem from the expression $P\left(A\right) = \sum_{i=1}^{n} P\left(E_i\right) P\left(A/E_i\right)$, it can be written as,

$$P\left(E_i/A\right) = \frac{P\left(E_i\right) P\left(A/E_i\right)}{\sum_{i=1}^{n} P\left(E_i\right) P\left(A/E_i\right)} \tag{2.17}$$

This is known as the *Bayes' rule*.

The denominator on the *r.h.s.*, i.e., $\sum_{i=1}^{n} P\left(E_i\right) P\left(A/E_i\right)$, is a constant term. Thus, using proportionality,

$$P\left(E_i/A\right) \propto P\left(E_i\right) P\left(A/E_i\right)$$

In this expression, the term $P\left(E_i\right)$ is the probability of occurrence of E_i, without knowing any other information. This term is referred as *prior*. Next, knowing that event A has occurred, probability of occurrence of E_i, i.e., $P\left(E_i/A\right)$, is updated. Thus, this term is referred as *posterior*. The probability of occurrence of the event A, given that E_i has occurred, is generally evaluated/estimated from historical records/experience. This term, $P\left(A/E_i\right)$ is, referred as *likelihood*. Using these terms, the Bayes' rule is often expressed as,

$$Posterior \propto Prior \times Likelihood$$

Example 2.4.1

Municipality of a city uses 70% of its required water from a nearby river and remaining from the groundwater. There could be various reasons of not getting the required supply from either sources including pump failure, non-availability of sufficient water. If probability of shortage of water from river is 0.3 and that from groundwater is 0.15, what is the probability of insufficient supply of water to the city?

Solution Let us first denote the events mentioned in the example,

Event A: insufficient supply of water to the city;
Event R: sufficient supply of water from the river;
Event G: sufficient supply of water from groundwater.

Thus, we get $P(R) = 0.7$, $P(G) = 0.3$, $P(A/R) = 0.3$, $P(A/G) = 0.15$
Using Theorem of Total Probability,

$$P\left(A\right) = P\left(R\right) \times P\left(A/R\right) + P\left(G\right) \times P\left(A/G\right) = 0.7 \times 0.3 + 0.3 \times 0.15 = 0.255$$

Example 2.4.2
A series of rainfall record is assimilated from the measurements obtained from three different instruments. 30% measurements are taken by instrument A that yields one missing data out of 200 on an average, 45% measurements are taken by instrument B that yields one missing data out of 150 on an average and rest by instrument C that yields one missing data out of 100 on an average. One measurement is found to yield a missing data, what is the probability that the measurement was taken by instrument A.

Solution The probability that the measurement was made by instrument A on condition that the measurement is wrong can be calculated using Bayes' theorem. Let X_1, X_2, and X_3 represent the events that the measurement was made by instrument A, B, and C, respectively. Let Y represents the event that the measurement was missing.

$$P(X_1/Y) = \frac{P(Y/X_1)P(X_1)}{\sum_{i=1}^{3} P(Y/X_i)P(X_i)}$$

$$P(X_1/Y) = \frac{(1/200) \times 0.3}{(1/200) \times 0.3 + (1/150) \times 0.45 + (1/100) \times 0.25} = 0.214$$

The probability that the measurement was made by instrument A given that the measurement is wrong is 0.214.

Example 2.4.3
A series of soil moisture data is prepared by collecting samples from two different sources. Though the sources are random for any month, a total of 600 samples are obtained from source-A that contains 3% erroneous data and a total of 400 samples are obtained from source-B that contains 1% erroneous data.

(a) What is the probability that the data for a month selected at random is obtained from source-A?
(b) What is the overall percentage of erroneous data?
(c) An erroneous data is selected at random, what is the probability that it is from source-A?

Solution Let us denote the following events:

A: data obtained from source-A;
B: data obtained from source-B;
E: selected data is erroneous.

(a) Thus, the probability of data obtained from source-A, i.e., $P(A)$ is given by,

$$P(A) = \frac{600}{(600 + 400)} = 0.6$$

(b) The erroneous data may come from either source-A or source-B. Therefore, we need to apply the total probability theorem to calculate the probability of event E; i.e., the selected value is erroneous:

$$\begin{aligned} P(E) &= P(E/A)P(A) + P(E/B)P(B) \\ &= 0.03 \times 0.6 + 0.01 \times 0.4 \\ &= 0.022 \end{aligned}$$

(c) If the sample data selected at random is erroneous, probability that it comes from source-A is not 0.6 as in case of solution (a), it is because of the change of sample space. Instead of entire data, the new sample space consists of only erroneous data. Thus, using Bayes' rule,

$$P(A/E) = \frac{P\left(E/A\right)P\left(A\right)}{P\left(E/A\right)P\left(A\right) + P\left(E/B\right)P\left(B\right)} = \frac{0.03 \times 0.6}{0.03 \times 0.6 + 0.01 \times 0.4} = 0.818$$

Example 2.4.4
The flood damages at a location are caused mainly due to poor management of different measures. These measures can be classified into two major groups—structural and non-structural measures. The Flood Management Authority (FMA) analyzes various issues involved and found that the possibility of improving the structural and non-structural measures to prevent flood are 70 and 55%, respectively, considering various socioeconomic factors. If only one of these two measures is successfully implemented, the probability of preventing the flood damages is 80%. Assuming that flood damages caused by poor management of structural and non-structural measures are independent,

(a) What is the probability of preventing the flood damages?
(b) If the flood damages are not prevented, what is the probability that it is entirely caused by the failure due to poor management of non-structural measures?
(c) If the flood damages are not prevented, what is probability that it is caused by the failure due to poor management of non-structural measures?

Solution Let us define the events as follows:

A: prevention of flood damages due to improvement of structural measures;
B: prevention of flood damages due to improvement of non-structural measures;
E: prevention of the flood damages.

Since the events A and B are independent, we have,

$$\begin{aligned} P(AB) &= 0.70 \times 0.55 = 0.385 \\ P(A^c B) &= 0.30 \times 0.55 = 0.165 \\ P(AB^c) &= 0.70 \times 0.45 = 0.315 \\ P(A^c B^c) &= 0.30 \times 0.45 = 0.135 \end{aligned}$$

It is also known that if only one of the two measures is successfully implemented, the probability of reducing the flood damages is 80%. Thus,

$$P\left(E/A^c B\right) = 0.8 \text{ and } P\left(E/AB^c\right) = 0.8$$

It is also implied that prevention of flood damages due to improvement of both structural and non-structural measures is certain and that due to no improvement either structural or non-structural measures is 0, i.e., $P\left(E/AB\right) = 1$ and $P\left(E/A^c B^c\right) = 0$.

(a) Thus, using the total probability theorem, the probability of prevention of the flood damages is

$$\begin{aligned} P(E) = {} & P\left(E/AB\right) P(AB) + P\left(E/A^c B\right) P(A^c B) + P\left(E/AB^c\right) P(AB^c) \\ & + P\left(E/A^c B^c\right) P(A^c B^c) \\ = {} & 1 \times 0.385 + 0.8 \times 0.165 + 0.8 \times 0.315 + 0 \times 0.135 \\ = {} & 0.769 \end{aligned}$$

(b) Next, if the flood damages are not prevented, i.e., E^c, the probability that it is entirely caused by the failure due to poor management of non-structural measures, i.e., AB^c

$$P\left(AB^c/E^c\right) = \frac{P\left(E^c/AB^c\right) P(AB^c)}{P(E^c)} = \frac{(1 - 0.8) \times 0.315}{(1 - 0.769)} = 0.273$$

(c) In this question the word ‘entirely’ is not used. Thus, we need to calculate the probability of $P(B^c/E^c)$.

$$
\begin{aligned}
P\left(B^c/E^c\right) &= P\left(AB^c \cup A^cB^c/E^c\right)\\
&= P\left(AB^c/E^c\right) + P\left(A^cB^c/E^c\right)\\
&= \frac{P\left(E^c/AB^c\right)P(AB^c)}{P(E^c)} + \frac{P\left(E^c/A^cB^c\right)P(A^cB^c)}{P(E^c)}\\
&= \frac{0.2 \times 0.315}{(1-0.769)} + \frac{1 \times 0.135}{(1-0.769)} \quad = \quad 0.857
\end{aligned}
$$

2.5 Univariate and Bivariate Probability Distribution of Random Variables

As mentioned before, the random variable is a function on the sample space that maps the outcomes of a random experiment to a real number. There are two types of random variables, namely *discrete random variable* and *continuous random variable*. In general, the probability distribution is expressed as a function of the random variable showing the distribution of probability corresponding to all possible values of random variable. All possible values of a random variable constitute the *support of the random variable*. Generally, the term probability density function (*pdf*) is used for *continuous random variable*, and probability mass function (*pmf*) is used for *discrete random variable*.

The term *univariate* and *bivariate* signifies the number of random variables involved in the distribution function. Univariate probability distributions deal with a single random variable, and bivariate probability distributions deal with two random variables. Similarly, distribution functions involving more than two random variables are called multivariate probability distribution. In the following section, we will explain univariate and bivariate probability distribution for discrete and continuous random variables.

2.5.1 Discrete Random Variable

As stated before, a discrete random variable can take only finite or countably infinite distinct values. Examples may include number of rainy days in a month, number of occurrences of an extreme event during monsoon season, etc.

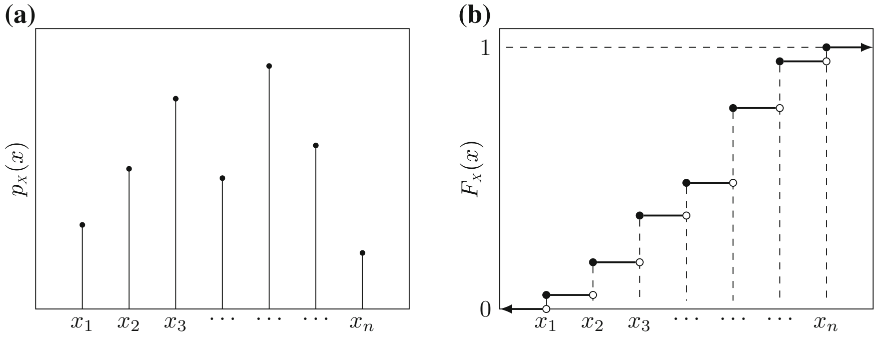


Fig. 2.7 Typical plot of **a** *pmf* and **b** *CDF* of a discrete random variable

Discrete Univariate Probability Distribution

Probability Mass Function (pmf): Let us consider X to be a discrete random variable taking values in a set $\theta = \{x_1, x_2, \ldots, x_n\}$. The probability mass function (*pmf*) of X is $p_X(\bullet)$ satisfying,

(i) $p_X(x_i) \geq 0 \quad \forall\, x_i \in \theta$

(ii) $\sum\limits_{\text{all } i} p_X(x_i) = 1$

A typical plot of a *pmf* is shown in Fig. 2.7a, where filled circles indicate the probability masses concentrated at a point. The vertical lines as such do not indicate anything except showing the position of the values on the x-axis.

Cumulative Distribution Function (CDF): The *CDF* $(F_X(x_i))$ represents the probability that X is less than or equal to x_i. This can be represented as,

$$F_X(x_i) = P\,(X \leq x_i) = \sum_{j=1}^{i} P(X = x_j) \qquad \forall\, x \in \{x_1, x_2, \ldots, x_n\} \tag{2.18}$$

A typical plot of *CDF* for the discrete random variable is shown in Fig. 2.7b. It is a non-decreasing, discontinuous, staircase-like functions with irregular rise. Filled and open circles in this plot indicate inclusive and exclusive boundaries, respectively. The jump at each x_i indicates the value of $p_X(x_i)$ or the probability that $X = x_i$. This probability can be determined from the *CDF* as follows:

$$p_X(x_i) = F_X(x_i) - F_X(x_{i-1}) \tag{2.19}$$

Example 2.5.1
Number of rainy days in the last week of December (traditionally a dry month) at a location is found to follow the following distribution.

$$p_X(x) = \begin{cases} C & \text{for } x = 0 \\ \frac{e^{-1}}{2x} & \text{for } x = 1, 2, \ldots, 7 \\ 0 & \text{elsewhere} \end{cases}$$

Evaluate the value of C for $p_X(x)$ to be a valid *pmf* and the probability of more than two rainy days in the last week of December.

Solution Let us consider X to represent the number of rainy days in the last week of December. The value of C can be evaluated as follows:

$$\sum_{\text{all } i} p_X(x_i) = 1$$

Thereby,

$$C + \sum_{x=1}^{7} \frac{e^{-1}}{2x} = 1$$
$$\Rightarrow C = 1 - 0.477 = 0.523$$

Hence, the complete *pmf* is

$$p_X(x) = \begin{cases} 0.523 & \text{for } x = 0 \\ \frac{e^{-1}}{2x} & \text{for } x = 1, 2, \ldots, 7 \\ 0 & \text{elsewhere} \end{cases}$$

The probability of more than two rainy days can be evaluated as,

$$\begin{aligned} P(X > 2) &= 1 - P(X \le 2) \\ &= 1 - [P(X = 0) + P(X = 1) + P(X = 2)] \\ &= 1 - \left[0.523 + \frac{e^{-1}}{2} + \frac{e^{-1}}{2 \times 2}\right] \\ &= 0.201 \end{aligned}$$

Thereby, the probability of more than two rainy days in the last week of December is 0.201

Discrete Bivariate Probability Distribution

Let us consider X and Y are two discrete random variables, and let $p_{X,Y}(x, y)$ be their joint probability mass function (pmf). For a valid joint *pmf* of two discrete random variables X and Y, following conditions are to be fulfilled:

$$\left.\begin{aligned} &p_{X,Y}(x, y) > 0 && \text{for all } x \text{ and } y \\ &\sum_{\text{all } x}\sum_{\text{all } y} p_{X,Y}(x, y) = 1 && \end{aligned}\right\} \tag{2.20}$$

If $F_{X,Y}(x, y)$ be the corresponding cumulative probability distribution function, then

$$F_{X,Y}(x, y) = \sum\sum p_{X,Y}(x_i, y_j) \qquad \text{for all } (x_i, y_i) \text{ s.t. } x_i < x \text{ and } y_i < y \tag{2.21}$$

Example 2.5.2
The joint *pmf* of two random variables X and Y is given by

$$p_{X,Y}(x, y) = \begin{cases} k(2x + 5y) & \text{for } x = 1, 2;\ y = 1, 2 \\ 0 & \text{otherwise} \end{cases}$$

What is value of k to be a valid joint *pmf*?

Solution From the properties of joint *pmf*,

$$\sum_{(x,y)\in S} p_{X,Y}(x, y) = 1$$

$$\sum_{allx}\sum_{ally} p_{X,Y}(x, y) = \sum_{x=1}^{2}\sum_{y=1}^{2} k(2x + 5y)$$

$$1 = k\{(2 + 5) + (2 + 10) + (4 + 5) + (4 + 10)\}$$

$$\text{Hence, } k = \frac{1}{42}$$

2.5.2 Continuous Random Variable

As stated before, if the set of values a random variable can assume is a continuum, i.e., all possible values within a range of real numbers, is known as *continuous random variable*. In hydrology and hydroclimatology, most of the variables are continuous e.g., streamflow, rainfall depth, evapotranspiration, temperature, wind speed, relative humidity, soil moisture. Support of these variables may be unbounded (e.g., temperature), bounded at one side (e.g., streamflow, rainfall depth), or both sides (e.g., relative humidity, soil moisture).

Univariate Probability Distribution

In case of a continuous random variable, the probability density function (*pdf*) is generally denoted by $f_X(x)$, where the subscript X denotes the random variable and the variable x within the parentheses denotes a specific value of the random variable. For any function to be a valid *pdf*, it has to satisfy two conditions as follows:

(i) $f_X(x) \geq 0$ for all x
(ii) $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

It may be noted that, unlike *pmf*, $f_X(x)$ does not directly provide the value of probability, rather it is probability density. Integration over any range of x provides the probability of X being within that range.

Cumulative Distribution Function (CDF)

The *CDF* ($F_X(x)$) represents the probability that X is less than or equal to a specific value of x, i.e.,

$$F_X(x) = P(X \leq x)$$

The *CDF* is obtained from the *pdf* by integrating it from the left extreme of the support to x. Thus, the expression of *CDF*, $F_X(x)$ is obtained as:

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\,dx$$

To obtain *pdf* from *CDF*, the *CDF* has to be differentiated with respect to x as follows:

$$\frac{d}{dx} F_X(x) = f_X(x) \tag{2.22}$$

The probability that X lies between $[a, b]$ is given by the following equation and illustrated in Fig. 2.8.

$$P(a \leq X \leq b) = \int_{a}^{b} f_X(x)\,dx = F_X(b) - F_X(a) \tag{2.23}$$

Note:

(i) In general, for continuous random variables, probability that the random variable takes a specific value is zero, i.e., $P(X = d) = \int_d^d f_X(x)\,dx = 0$. Thereby, $P(X \leq x) = P(X < x)$. This is not valid for discrete random variables.

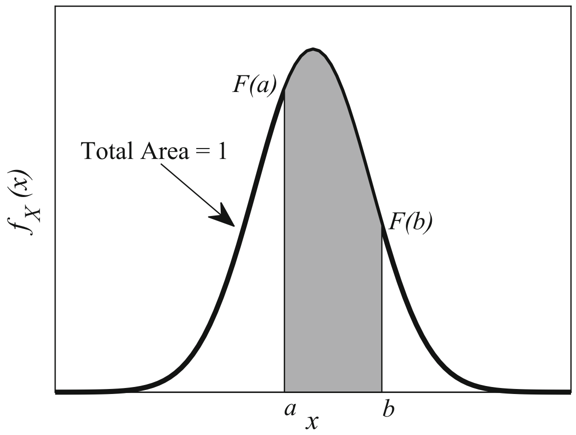


Fig. 2.8 Typical *pdf* for a continuous random variable (X) showing the probability of X lies between $[a, b]$ (shaded area)

(ii) Aforementioned point is also not valid for piecewise continuous distribution or mixed distribution. Without violating the requirements of a valid *pdf*, it is possible that $P\,(X = d)$ is not zero. The *CDF* of such distribution can be defined as follows:

$$F_x\,(x) = \begin{cases} F_1\,(x) & \text{for } X < d \\ F_2\,(x) & \text{for } X \geq d \end{cases} \tag{2.24}$$

where $F_2\,(d) > F_1(d)$, $F_1\,(-\infty) = 0$, $F_2\,(\infty) = 1$, and $F_1\,(x)$ *and* $F_2\,(x)$ are non-decreasing functions of X. For this situation, the $P\,(X = d)$ is equal to the magnitude of the jump ΔF at $X = d$ or is equal to $F_2\,(d) - F_1(d)$. Zero inflated daily rainfall values can be an example of such case where $P\,(X = 0)$ is not zero, and for the range $X > 0$, it is continuous. This situation will be dealt in Chap. 4.

Relative frequency and CDF

Let us consider $f_X\,(x)$ to be the probability density function of X. The probability that X lies between $X = a$ and $X = b$ is given by:

$$P\,(a \leq X \leq b) = \int_a^b f_x(x)dx = F_x\,(b) - F_x\,(a) \tag{2.25}$$

If there are N data available, the expected number of data to fall in the interval $[a, b]$ would be

$$n_{ab} = N\,[F_x\,(b) - F_x\,(a)] \tag{2.26}$$

Thereby, the expected relative frequency of outcomes in the interval $[a, b]$ is

$$f_{ab} = n_{ab}\big/N = F_x\,(b) - F_x\,(a) \tag{2.27}$$

In general, if x_i represents the midpoint of an interval of X given by $x_i - \Delta x_i/2$ to $x_i + \Delta x_i/2$ then the expected relative frequency of the data is given by

$$f_{xi} = F_X\left(x_i + \Delta x_i/2\right) - F_X\left(x_i - \Delta x_i/2\right) \tag{2.28}$$

This equation represents the area under $f_X(x)$ between $x_i - \Delta x_i/2$ and $x_i + \Delta x_i/2$, and it can be approximately written as,

$$f_{xi} = \Delta x_i f_X(x_i) \tag{2.29}$$

Example 2.5.3

The annual maximum discharge at a gauging station follows the following distribution.

$$f_X(x) = \begin{cases} \frac{1}{x^2} & x > 1 \\ 0 & \text{elsewhere} \end{cases}$$

Evaluate the following,

(a) What is the probability of annual maximum discharge greater than 5 units?
(b) What is the probability of annual maximum discharge between 2 and 10 units?

Solution (a) The probability of annual maximum discharge greater than 5 units can be evaluated as follows,

$$P\left(X > 5\right) = 1 - F_X(5) = 1 - \int_1^5 \frac{1}{x^2}dx = \frac{1}{5}$$

The probability of annual maximum discharge greater than 5 units is 0.2.

(b) The probability of annual maximum discharge between 2 and 10 units can be evaluated as follows,

$$\begin{aligned} P\left(2 < X < 10\right) &= F_X(10) - F_X(2) \\ &= \int_1^{10} \frac{1}{x^2}dx - \int_1^2 \frac{1}{x^2}dx \\ &= 0.9 - 0.5 \\ &= 0.4 \end{aligned}$$

The probability of annual maximum discharge between 2 and 10 units is 0.4

Continuous Bivariate Probability Distribution

Let us consider X and Y to be continuous random variables with joint *pdf*, $f_{X,Y}(x, y)$. For a valid joint *pdf*, the following conditions are to be fulfilled:

$$f_{X,Y}(x, y) > 0 \qquad \forall\, x \text{ and } y \tag{2.30a}$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x, y)\, dxdy = 1 \tag{2.30b}$$

The corresponding cumulative probability distribution function is expressed as:

$$F_{X,Y}(x, y) = P(X \le x,\ Y \le y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(t, s)\, dsdt$$

The *pdf* $\left[f_{X,Y}(x, y)\right]$ and the *CDF* $\left[F_{X,Y}(x, y)\right]$ are related as follows:

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) \tag{2.31}$$

Some of the properties of continuous bivariate cumulative distribution are:

(i) $F_{X,Y}(x, \infty)$ is the cumulative marginal probability function of X.
(ii) $F_{X,Y}(\infty, y)$ is the cumulative marginal probability function of Y.
(iii) $F_{X,Y}(\infty, \infty) = 1$
(iv) $F_{X,Y}(-\infty, y) = F_{X,Y}(x, -\infty) = 0$

Example 2.5.4
A storm event occurring at a point in space is characterized by two variables, namely the duration X of the storm and the depth of rainfall Y. The variables X and Y follow following distribution, respectively:

$$F_X(x) = 1 - e^{-x} \qquad x \ge 0$$
$$F_Y(y) = 1 - e^{-2y} \qquad y \ge 0$$

The joint *CDF* of X and Y is assumed to follow the bivariate distribution given as:

$$F_{X,Y}(x, y) = 1 - e^{-x} - e^{-2y} + e^{-x-2y-xy} \qquad x, y \ge 0$$

Find out the cumulative marginal probability function of X and Y. Also, find out the joint *pdf* of X and Y.

Solution From the properties of continuous bivariate cumulative distribution, we know, Marginal CDF of X,

$$\begin{aligned} &F_X(x) = F_{X,Y}(x, \infty) \\ &F_X(x) = 1 - e^{-x} - e^{-2\infty} + e^{-x-2\infty-x\infty} && x \geq 0 \\ &F_X(x) = 1 - e^{-x} && x \geq 0 \end{aligned}$$

Similarly, Marginal CDF of Y,

$$\begin{aligned} &F_Y(y) = F_{X,Y}(\infty, y) \\ &F_Y(y) = 1 - e^{-\infty} - e^{-2y} + e^{-\infty-2y-\infty y} && y \geq 0 \\ &F_Y(y) = 1 - e^{-2y} && y \geq 0 \end{aligned}$$

We know, $f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$

Differentiating the joint CDF w.r.t x, we get

$$\frac{\partial F}{\partial x} = \frac{\partial \left(1 - e^{-x} - e^{-2y} + e^{-x-2y-xy}\right)}{\partial x} = e^{-x} - (1 + y)\, e^{-x-2y-xy}$$

Again differentiating the above equation w.r.t y

$$\begin{aligned} f_{X,Y}(x, y) = \frac{\partial^2 F}{\partial x \partial y} &= \frac{\partial \left(e^{-x} - (1 + y)\, e^{-x-2y-xy}\right)}{\partial y} \\ &= [(1 + y)(2 + x) - 1]\, e^{-x-2y-xy} \end{aligned}$$

Hence, joint pdf of X and Y is

$$f_{X,Y}(x, y) = [(1 + y)(2 + x) - 1]\, e^{-x-2y-xy} \qquad x, y \geq 0$$

2.6 Marginal and Conditional Probability Distribution

Marginal and conditional probability distributions are discussed in the context of multivariate distributions. It is a very useful concept to be used in hydrologic and hydroclimatic prediction and simulation since many variables are associated with each other. The concept of these distributions will be discussed in the context of bivariate distribution (two random variables) and will be extended for the multivariate cases with more than two random variables.

2.6.1 Marginal Probability Distribution

Discrete Random Variables

Let us consider X and Y to be two discrete random variables with their joint *pmf* as $p_{X,Y}(x, y)$. Thus, the joint distribution with $p_{X,Y}\left(x_i, y_j\right) = P\left(X = x_i, Y = y_j\right)$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$ appears as a $m \times n$ two-dimensional table of probability values corresponding to a pair of X and Y values (Table 2.1).

The marginal distribution is the distribution of one of the two random variables, i.e., either X or Y, irrespective of the distribution of the other variable. Thus, the marginal probability of X is obtained by summing up the probability values for all possible values of Y. In other words, the random variable is *marginalized out*. Mathematically, it is obtained as:

$$p_X(x_i) = \sum_{k=1}^{n} p_{X,Y}(x_i, y_k) \qquad \text{for } i = 1, 2, \ldots, m$$

Similarly, the marginal distribution of Y is obtained as:

$$p_Y\left(y_j\right) = \sum_{k=1}^{m} p_{X,Y}\left(x_k, y_j\right) \qquad \text{for } j = 1, 2, \ldots, n$$

The corresponding cumulative marginal distributions are:

$$F_X(x) = \sum_{x_i \le x} p_X(x_i) = \sum_{x_i \le x} \sum_{\text{all } y_j} p_{X,Y}\left(x_i, y_j\right)$$

$$F_Y(y) = \sum_{y_j \le y} p_Y\left(y_j\right) = \sum_{y_j \le y} \sum_{\text{all } x_i} p_{X,Y}\left(x_i, y_j\right)$$

Continuous Random Variables

Let us consider X and Y to be two continuous random variables with their joint *pdf* as $f_{X,Y}(x, y)$. Following the similar concept, the marginal distribution of X can be obtained by *marginalizing Y out*. Mathematically, Y is integrated out to get the marginal distribution of X from the joint *pdf* $f_{X,Y}(x, y)$. Thus,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy \tag{2.32}$$

Similarly, marginal distribution of Y is expressed as:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx \tag{2.33}$$

Table 2.1 Discrete joint distribution between X and Y along with the marginal distributions of X (last column) and Y (last row)

Random variables		Y				Marginal distribution of $X\,[p_X(x_i)]$
		y_1	y_2	$\ldots$	y_n	
X	x_1	$p_{X,Y}\quad(x_1, y_1)$	$p_{X,Y}\quad(x_1, y_2)$	$\ldots$	$p_{X,Y}\quad(x_1, y_n)$	$p_X(x_1) = \sum_{\text{all } j} p_{X,Y}\quad(x_1, y_j)$
	x_2	$p_{X,Y}\quad(x_2, y_1)$	$p_{X,Y}\quad(x_2, y_2)$	$\ldots$	$p_{X,Y}\quad(x_2, y_n)$	$p_X\,(x_2) = \sum_{\text{all } j} p_{X,Y}\quad(x_2, y_j)$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	x_m	$p_{X,Y}\quad(x_m, y_1)$	$p_{X,Y}\quad(x_m, y_2)$	$\ldots$	$p_{X,Y}\quad(x_m, y_n)$	$p_X\,(x_m) = \sum_{\text{all } j} p_{X,Y}\quad(x_m, y_j)$
Marginal distribution of $Y\,(p_Y(y_j))$		$p_Y(y_1) = \sum_{\text{all } i} p_{X,Y}\quad(x_i, y_1)$	$p_Y(y_2) = \sum_{\text{all } i} p_{X,Y}\quad(x_i, y_2)$	$\ldots$	$p_Y(y_n) = \sum_{\text{all } i} p_{X,Y}\quad(x_i, y_n)$	

The corresponding cumulative marginal distributions are:

$$\left.\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(x)\,dx = \int_{-\infty}^{x}\left[\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\right]dx \\ F_Y(y) &= \int_{-\infty}^{y} f_y(y)\,dy = \int_{-\infty}^{y}\left[\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\right]dy \end{aligned}\right\} \tag{2.34}$$

Example 2.6.1

Let X denotes the number of rainy days at station A and Y denotes the number of rainy days at station B. The joint *pmf* of X and Y is given as follows. Find out the marginal distribution of X and Y.

Random		Y			
Variables		0	2	5	7
	0	36/120	18/120	12/120	1/120
X	2	18/120	4/120	9/120	0
	5	12/120	9/120	0	0
	7	1/120	0	0	0

Solution The marginal *pmf* of X can be evaluated using equations shown in Table 2.1.

$$\begin{aligned} P(X = x_i) &= \sum_{\text{all } j} P\left(X = x_i, Y = y_j\right) \\ p_X(0) &= \frac{67}{120} \\ p_X(2) &= \frac{31}{120} \\ p_X(5) &= \frac{21}{120} \\ p_X(7) &= \frac{1}{120} \end{aligned}$$

The marginal *pmf* of Y is as follows,

$$\begin{aligned} P\left(Y = y_j\right) &= \sum_{\text{all } i} P\left(X = x_i, Y = y_j\right) \\ p_Y(0) &= \frac{67}{120} \\ p_Y(2) &= \frac{31}{120} \end{aligned}$$

$$p_Y(5) = \frac{21}{120}$$
$$p_Y(7) = \frac{1}{120}$$

Example 2.6.2
Streamflows at two gauging stations on two nearby tributaries are categorized into four different states, i.e., 1, 2, 3, and 4. These categories are represented by two random variables X and Y, respectively, for two tributaries. Joint *pmf* of streamflow categories (X and Y) are shown in the following table. Calculate the probability of $X > Y$.

Random		*Y*			
Variables		1	2	3	4
X	1	0.310	0.060	0.000	0.000
	2	0.040	0.360	0.010	0.000
	3	0.010	0.025	0.114	0.030
	4	0.010	0.001	0.010	0.020

Solution Let $P(A)$ represent the probability of the event $X > Y$. This will include the set {2, 1}, {3, 2}, {3, 1}, {4, 3}, {4, 2} and {4, 1}.

Thus, probabilities of these sets should be added up to obtain the required probability.

Thus, the probability is given by:

$$\begin{aligned}
P(A) &= P[X > Y] \\
&= \sum_{\text{all possible } x > y} p_{X,Y}(x, y) \\
&= p_{X,Y}(2,1) + p_{X,Y}(3,2) + p_{X,Y}(3,1) + p_{X,Y}(4,3) + p_{X,Y}(4,2) + p_{X,Y}(4,1) \\
&= 0.040 + 0.025 + 0.010 + 0.010 + 0.001 + 0.010 \\
&= 0.096
\end{aligned}$$

Example 2.6.3
The joint *pdf* of two random variables X and Y is given by

$$f_{X,Y}(x, y) = \begin{cases} 2 & 0 \le x \le 1;\ y \le x \\ 0 & \text{otherwise} \end{cases}$$

Determine their marginal *pdf*s.

Solution The marginal *pdf*s are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy = 2\int_0^x dy = 2x \qquad \text{for } 0 \le x \le 1$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx = 2\int_y^1 dx = 2(1 - y) \qquad \text{for } 0 \le y \le 1$$

Example 2.6.4
Two random variables X and Y have joint distribution as follows

$$f_{X,Y}(x, y) = \begin{cases} k(x + y) & 0 < x \le 2 \text{ and } 0 < y \le 4 \\ 0 & \text{otherwise} \end{cases}$$

Find out the value of k and marginal *pdf*s for X and Y.

Solution We know that

$$\int_0^4 \int_0^2 k(x + y)dxdy = 1$$

$$\int_0^4 \left[k\frac{x^2}{2} + kyx\right]_0^2 dy = 1$$

$$\int_0^4 [2k + 2yk]\, dy = 1$$

$$\left[2ky + ky^2\right]_0^4 = 1$$

$$8k + 16k = 1$$

Thus, we obtain $k = \frac{1}{24}$
The marginal distribution of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy = \int_0^4 \frac{x + y}{24} dy = \left[\frac{xy + \frac{y^2}{2}}{24}\right]_0^4 = \frac{x + 2}{6} \qquad 0 < x < 2$$

The marginal distribution of Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx = \int_0^2 \frac{x + y}{24} dx = \left[\frac{\frac{x^2}{2} + yx}{24}\right]_0^2 = \frac{1 + y}{12} \qquad 0 < y < 4$$

Example 2.6.5
A storm event occurring at a point in space is characterized by two variables, namely the duration X of the storm and the depth of rainfall Y, as illustrated in Example 2.5.4. Determine the marginal *pdf* and *CDF* of X and Y.

Solution From Example 2.5.4, we know the joint bivariate *pdf* of X and Y is as follows:

$$f_{X,Y}(x,y) = [(1+cy)(2+cx)-c]\,e^{-x-2y-cxy} \qquad x, y \geq 0$$

Hence, the marginal *pdf* of storm duration X is:

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \\
&= \int_0^{\infty} [(1+y)(2+x)-1]\,e^{-x-2y-xy}dy \\
&= \int_0^{\infty} [(1+y)(2+x)]\,e^{-x-2y-xy}dy - \int_0^{\infty} e^{-x-2y-xy}dy \\
&= \left[-(1+y)\,e^{-x-2y-xy}\right]_0^{\infty} \\
&= e^{-x} \qquad\qquad \text{for } x \geq 0
\end{aligned}$$

So, the marginal *CDF* of storm duration X is:

$$\begin{aligned}
F_X(x) &= \int_{-\infty}^{x} f_X(x)\,dx \\
&= \int_0^{x} e^{-x}(x)\,dx \\
&= 1 - e^{-x} \qquad\qquad \text{for } x \geq 0
\end{aligned}$$

(Result matches with Example 2.5.4)

Similarly, the marginal *pdf* of rainfall depth Y is

$$\begin{aligned}
f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \\
&= \int_0^{\infty} [(1+y)(2+x)-1]\,e^{-x-2y-xy}dx \\
&= \int_0^{\infty} [(1+y)(2+x)]\,e^{-x-2y-xy}dx - \int_0^{\infty} e^{-x-2y-xy}dx \\
&= \left[-(2+x)\,e^{-x-2y-xy}\right]_0^{\infty} \\
&= 2e^{-2y} \qquad\qquad \text{for } y \geq 0
\end{aligned}$$

So, the marginal *CDF* of rainfall depth Y is:

$$\begin{aligned}
F_Y(y) &= \int_{-\infty}^{y} f_Y(y)\,dy \\
&= \int_{0}^{y} 2e^{-2y}(y)\,dy \\
&= 1 - e^{-2y} \qquad\qquad \text{for } x \geq 0
\end{aligned}$$

(Result matches with Example 2.5.4)

2.6.2 Conditional Distribution Function

It may be recalled that marginal distribution of a random variable completely *marginalizes out* the other variable. However, conditional probability distribution of a random variable is the probability distribution of one variable (say X) for a specific value/range of other variable (say Y). For example, the distribution of X given $Y = y_0$, the distribution of Y given $x_1 \leq X \leq x_2$ and so on.

Let us also recall the conditional probability of event A, conditioned on event B expressed as Eq. 2.10,

$$P\left(A/B\right) = \frac{P(A \cap B)}{P(B)} \tag{2.35}$$

This theorem can be utilized to obtain the conditional distribution function of the random variables.

Discrete Random Variable

If X and Y are two discrete random variables and we have the events $(X = x)$ and $(Y = y)$, then using the conditional probability theorem the conditional probability of Y given X under the said conditions can be expressed as:

$$P\left(Y = y/X = x\right) = \frac{P(X = x, Y = y)}{P(X = x)} \tag{2.36}$$

where $P(X = x, Y = y)$ is the joint *pmf* of X and Y, i.e., $f_{X,Y}(X = x, Y = y) = f_{X,Y}(x, y)$ and $P(X = x)$ is the marginal probability distribution of X, i.e., $p_X(X = x) = p_X(x)$. Thereby, using joint *pmf* and marginal probability distribution, the conditional distribution function of Y given X can be written as,

$$p_{Y/X}\left(y/x\right) = \frac{p_{X,Y}(x, y)}{p_X(x)} \tag{2.37}$$

Similarly, the conditional distribution function of X given Y can be expressed as,

$$p_{X/Y}\left(x/y\right) = \frac{p_{X,Y}\left(x, y\right)}{p_Y\left(y\right)} \tag{2.38}$$

Extending this concept, the condition may be over a range of values, such as $X \in \tilde{R}$ ($\tilde{R}$ may contain multiple values of X). In such cases, the expression of the conditional distribution function of Y given that $X \in \tilde{R}$ is expressed as,

$$p_{Y/X}\left(y\Big/X \in \tilde{R}\right) = \frac{\sum\limits_{x_i \in \tilde{R}} p_{X,Y}\left(x_i, y\right)}{\sum\limits_{x_i \in \tilde{R}} p_X\left(x_i\right)} \tag{2.39}$$

Similarly, the expression of the conditional distribution function of X given that $Y \in \tilde{R}$ is expressed as

$$p_{X/Y}\left(x\Big/Y \in \tilde{R}\right) = \frac{\sum\limits_{y_j \in \tilde{R}} p_{X,Y}\left(x, y_j\right)}{\sum\limits_{y_j \in \tilde{R}} p_Y\left(y_j\right)} \tag{2.40}$$

Continuous Random Variable

Concept of conditional distribution in case of continuous random variables remains same as for discrete variables. If X and Y are two continuous random variables with their joint *pdf* as $f_{X,Y}\left(x, y\right)$ and marginal distributions as $f_X\left(x\right)$ and $f_Y\left(y\right)$, the conditional probability distribution of Y given X can be expressed as:

$$f_{Y/X}\left(y/x\right) = \frac{f_{X,Y}\left(x, y\right)}{f_X\left(x\right)} \tag{2.41}$$

In the above expression, the conditioning variable remains as an unspecified value (x) in the conditional distribution function. Most often this is convenient if the conditional distributions are required to derive for different values of conditioning variable.

Similarly, the conditional distribution function of X given Y can be expressed as,

$$f_{X/Y}\left(x/y\right) = \frac{f_{X,Y}\left(x, y\right)}{f_Y\left(y\right)} \tag{2.42}$$

When the conditioning variable belongs to a range $\tilde{R}$ (say $\tilde{R} \in x_1 \leq X \leq x_2$), the conditional distribution function of Y, conditioned on $X \in \tilde{R}$, is expressed as follows,

$$f_{Y/X}\left(y\big/x_1 \leq X \leq x_2\right) = \frac{\int_{x_1}^{x_2} f_{X,Y}(x, y)dx}{\int_{x_1}^{x_2} f_X(x)dx} \tag{2.43}$$

Similarly, the expression of the conditional distribution function of X, conditioned on $Y \in \tilde{R}$ (say $\tilde{R} \in y_1 \leq Y \leq y_2$), is expressed as,

$$f_{X/Y}\left(x/y_1 \leq Y \leq y_2\right) = \frac{\int_{y_1}^{y_2} f_{X,Y}(x, y)dy}{\int_{y_1}^{y_2} f_Y(y)dy} \tag{2.44}$$

Example 2.6.6
Utilizing the data given in Example 2.6.2, estimate the conditional probability of X when $Y \geq 2$.

Solution The conditional probability of X when $Y \geq 2$ can be evaluated using the conditional probability theorem as follows,

$$p_{X/Y \geq y} = P\left(X = x/Y \geq y\right) = \frac{\sum\limits_{y_i \geq y} p_{X,Y}\left(x, y_i\right)}{\sum\limits_{y_i \geq y} p_Y\left(y_i\right)}$$

$$p_{X/Y \geq 2} = \frac{\sum\limits_{y_i \geq 2} p_{X,Y}\left(x, y_i\right)}{\sum\limits_{y_i \geq 2} p_Y\left(y_i\right)}$$

The marginal probability of $Y \geq 2$ (denominator of $p_{X/Y \geq y}$) can be evaluated using the data given in Example 2.6.2 as,

$$\sum_{y_i \geq 2} p_Y\left(y_i\right) = 0.446 + 0.134 + 0.05 = 0.63$$

The numerator of $p_{X/Y \geq y}$ can be evaluated for each x_i (values taken up by the random variable X) using the data provided in Example 2.6.2 as,

X	1	2	3	4
$\sum\limits_{y_i \geq 2} p_{X,Y}\left(x, y_i\right)$	0.0600	0.3700	0.169	0.031

Knowing the denominator and the numerator of $p_{X/Y \geq y}$, the probability of X given $Y \geq 2$ can be evaluated for each x_i is as follows,

X	1	2	3	4
$p_{X/Y}\left(x/y \geq 2\right)$	0.0952	0.5873	0.2683	0.0492

Example 2.6.7
Let X denotes the average rainfall intensity in a particular catchment and Y denotes the peak discharge from the catchment. The joint *pdf* of X and Y is given as follows.

$$f_{X,Y}(x,y) = \begin{cases} x^2 + \frac{xy}{3} & 0 \le x \le 1, 0 \le y \le 2 \\ 0 & \text{elsewhere} \end{cases}$$

(a) Find out the marginal distribution of X and Y.
(b) Find out the probability of peak discharge being greater than 1 unit.
(c) Find out the probability of peak discharge being greater than 1 unit, given that average rainfall intensity is 0.5 units.

Solution (a) The marginal distribution of X can be evaluated as follows,

$$\begin{aligned} f_X(x) &= \int f(x,y)dy \\ &= \int_0^2 \left(x^2 + \frac{xy}{3}\right) dy \\ &= \left[x^2 y + \frac{xy^2}{6}\right]_0^2 \\ &= 2x^2 + \frac{2x}{3} \qquad 0 \le x \le 1 \end{aligned}$$

The marginal distribution of Y can be evaluated as follows,

$$\begin{aligned} f_Y(y) &= \int f(x,y)dy \\ &= \int_0^1 \left(x^2 + \frac{xy}{3}\right) dx \\ &= \left[\frac{x^3}{3} + \frac{x^2 y}{6}\right]_0^1 \\ &= \frac{1}{3} + \frac{y}{6} \qquad 0 \le y \le 2 \end{aligned}$$

(b) The probability of peak discharge being greater than 1 unit can be evaluated from the marginal distribution of Y, as follows,

$$\begin{aligned} P(Y > 1) &= 1 - P(Y \le 1) \\ &= 1 - \int_0^1 \left(\frac{1}{3} + \frac{y}{6}\right) dy \\ &= 1 - \frac{5}{12} \end{aligned}$$

$$= 0.583$$

(c) The probability of peak discharge being greater than 1 unit, given that average rainfall intensity is 0.5 units, can be evaluated using the conditional distribution function. The conditional distribution function of Y given X can be expressed as,

$$\begin{aligned} f_{Y/X}(y/x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &= \frac{x^2 + \frac{xy}{3}}{2x^2 + \frac{2x}{3}} \\ &= \frac{3x^2 + xy}{6x^2 + 2x} \qquad \text{for } 0 \le x \le 1, 0 \le y \le 2 \end{aligned}$$

Thus, the conditional distribution function of Y given $X = 0.5$

$$\begin{aligned} f_{Y/X=0.5}(y/X = 0.5) &= \left.\frac{3x^2 + xy}{6x^2 + 2x}\right|_{x=0.5} \\ &= \frac{3 \times 0.5^2 + 0.5 \times y}{6 \times 0.5^2 + 2 \times 0.5} \\ &= 0.3 + 0.2y \qquad \text{for } 0 \le y \le 2 \end{aligned}$$

Thereby, the probability of peak discharge being greater than 1 unit, given that average rainfall intensity is 0.5 unit, can be evaluated as,

$$\begin{aligned} P(y > 1/x = 0.5) &= 1 - \int_0^1 f_{Y/X=0.5}(y/X = 0.5)\,dy \\ &= 1 - \int_0^1 (0.3 + 0.2y)\,dy \\ &= 1 - \left[0.3y + 0.1y^2\right]_0^1 \\ &= 0.6 \end{aligned}$$

Note: It may be noted that when some information on average rainfall intensity is available, the probability of peak discharge being greater than 1 unit is higher than that without any information of the other variable. That is, we are more confident about the peak discharge with some known value of average rainfall intensity (conditional probability) that without any information (unconditional probability).

2.7 Independence between Random Variables

Let X and Y be two discrete random variables. These are independent if and only if their joint *pmf*, $p_{X,Y}(x, y)$ can be expressed by multiplying the respective marginal distributions as follows:

$$p_{X,Y}(x, y) = p_X(x) \times p_Y(y) \tag{2.45}$$

where $p_X(x)$ and $p_Y(y)$ are the marginal distribution of X and Y, respectively.

Similarly for continuous random variable, $f_{X,Y}(x, y)$ the joint *pdf* between them are independent if and only if,

$$f_{X,Y}(x, y) = f_X(x) \times f_Y(y) \tag{2.46}$$

where $f_X(x)$ and $f_Y(y)$ are the marginal distribution of X and Y, respectively.

In other words, if two random variables are independent, their joint distribution can be obtained by multiplying their respective marginal distributions. However, if the associated random variables are dependent, it is often difficult to ascertain their joint distribution. One possibility is to use copula theory, which is discussed in Chap. 10.

Example 2.7.1
Use the data from Example 2.6.7 to check if X and Y are independent.

Solution X and Y are said to be independent if the given condition is fulfilled.

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$$

As evaluated in Example 2.6.7, the marginal distribution of X is

$$f_X(x) = 2x^2 + \frac{2x}{3} \qquad 0 \le x \le 1$$

and, the marginal distribution of Y is

$$f_Y(y) = \frac{1}{3} + \frac{y}{6} \qquad 0 \le y \le 2$$

Thus,

$$x^2 + \frac{xy}{3} \ne \left(\frac{2}{3}x + 2x^2\right)\left(\frac{1}{3} + \frac{y}{6}\right)$$

Thereby, we can conclude that X and Y are not independent.

2.8 Functions of Random Variables

Sometimes functional relationship between two or more associated variables is known. If the distribution of one variable is known, it is possible/convenient to determine the distribution of other variable. The functional relationship could be of logarithmic, nth root transformation, or simple algebraic functions. The distribution of the transformed function is called derived distribution.

2.8.1 Univariate Random Variable

Let X be a random variable with *CDF* $F_X(x)$ and $y = g(x)$ then,

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(g(x) \le y) \\ &= P\left(X \in g^{-1}((-\infty, y])\right) \end{aligned} \tag{2.47}$$

Let X be a discrete random variable with *pmf* $p_X(x_i)$ and $y = g(x)$ then,

$$\begin{aligned} P(Y = y) &= P(g(x) = y) \\ &= \sum_{g(x)=y} P\left(X = g^{-1}(y)\right) \end{aligned} \tag{2.48}$$

Let X be a continuous random variable with *CDF* $F_X(x)$ and $y = g(x)$ then,

Case 1: Let g be an increasing function then,

$$\begin{aligned} F_Y(y) = P(Y \le y) &= P(g(x) \le y) \\ &= P\left(X \le g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right) \end{aligned} \tag{2.49}$$

So, the *pdf* of Y is as follows,

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| \tag{2.50}$$

Case 2: Let g be a decreasing function then,

$$\begin{aligned} F_Y(y) = P(Y \le y) &= P(g(x) \le y) \\ &= P\left(X \ge g^{-1}(y)\right) = 1 - F_X\left(g^{-1}(y)\right) \end{aligned} \tag{2.51}$$

So, the *pdf* of Y is as follows,

$$f_Y(y) = -f_X\left(g^{-1}(y)\right)\frac{d}{dy}g^{-1}(y) \tag{2.52}$$

However, in this case $\frac{d}{dy}g^{-1}(y)$ is negative. Thus,

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| \tag{2.53}$$

2.8.2 *Bivariate Random Variables*

Let us assume the joint *pdf* of X_1 and X_2 is $f_{X_1,X_2}(x_1, x_2)$. Let us also consider that Y and Z are the functions of X_1 and X_2,

$$Y = H_1(X_1, X_2) \text{ and } Z = H_2(X_1, X_2) \tag{2.54}$$

The joint *pdf* of Y and Z, represented as $f_{Y,Z}(y, z)$, is expressed as,

$$g(y, z) = f_{X_1,X_2}(y, z)|J| \tag{2.55}$$

where $f_{X_1,X_2}(y, z)$ is the *pdf* of X_1 and X_2 represented in terms of y and z, i.e., $x_1 = G_1(y, z)$ and $x_2 = G_2(y, z)$, obtained from Eq. 2.54. J is known as *Jacobian* and expressed as,

$$J = \frac{\partial(G_1, G_2)}{\partial(y, z)} = \begin{vmatrix} \frac{\partial G_1}{\partial y} & \frac{\partial G_1}{\partial z} \\ \frac{\partial G_2}{\partial y} & \frac{\partial G_2}{\partial z} \end{vmatrix} \tag{2.56}$$

Example 2.8.1
Let X follows a distribution shown below,

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

What will be the distribution of $Y = e^{-\lambda x}$?

Solution Given $Y = e^{-\lambda x}$ which can be written as,

$$x = -\frac{1}{\lambda}\ln y = g^{-1}(y)$$

$$\frac{d}{dy}\left(g^{-1}(y)\right) = -\frac{1}{\lambda y}$$

$$f_Y(y) = \lambda y \left| -\frac{1}{\lambda y} \right| = 1$$

$$f_Y(y) = \begin{cases} 1 & 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Example 2.8.2
Two streams A and B meet at a point C. The streamflow for *stream-A* (X) and *stream-B* (Y) follows the given distributions. What is the distribution of streamflow for stream C given by $U = X + Y$. Consider streamflow for stream A and B to be independent.

$$f_X(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$f_Y(y) = \begin{cases} e^{-y} & y > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Solution Given that streamflow for stream A and B are independent thereby,

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y) = e^{-(x+y)}$$

Given $U = X + Y$ and we can assume another function $V = X$ as per convenience. In order to find out the distribution of U first we evaluate the joint *pdf* of U and V. Then we evaluate the marginal distribution of U. We can write $X = V$ and $Y = U - V$.

$$|J| = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = 1$$

Thereby,

$$g(u, v) = f_{X,Y}(x, y)\,|J| = \begin{cases} e^{-u} & 0 < v < u < \infty \\ 0 & \text{elsewhere} \end{cases}$$

The distribution of U is as follows,

$$f_U(u) = \int_0^u e^{-u} dv = \begin{cases} ue^{-u} & u > 0 \\ 0 & \text{elsewhere} \end{cases}$$

2.9 MATLAB Examples

MATLAB (abbreviation for MATrix LABoratory) is a popular mathematical tool used for statistical analysis. A brief introduction to the concepts related to MATLAB is presented here.

In MATLAB, the memory location where data is stored is called workspace. Further, MATLAB GUI provides command window where commands can be issued. The commands can be stored in text files also known as M-files (having extension of '*.m'). These M-files are of two types:

(i) Script file: A script file is collection of instruction/commands which is executed together. The script works on the data in the workspace and cannot have any function definition; i.e., it neither accept any input argument nor return any output. However, any predefined function can be called.

(ii) Function files: Function file on the other hand contains at least one function. There can be multiple function definitions in single function file; however, only the function having same name as file name of M-file can be called from any external script file or MATLAB command window. Function files can accept input argument(s) and return output(s). It should be noted that usually the functions in MATLAB has a separate workspace. Hence, to use any data in main workspace in function, it needs to be transferred to the function as input argument.

MATLAB provides many built-in functions and toolboxes that can be used for hydroclimatological analysis. Toolbox is collection of functions for a particular purpose or domain. Symbolic toolbox, Statistics and Machine Learning Toolbox, Wavelet toolbox, Financial toolbox, etc., are some of the popular toolboxes available in MATLAB. This section (and similar section in other chapter) mostly deals with sample MATLAB script(s) for solving examples in the chapter. Sample function files are presented in Sect. 8.8 of Chap. 8. Some of the commonly used functions/commands in script presented in this book are '`disp`', '`fprintf`', and '`diary`'. The functions '`disp`' and '`fprintf`' are used to display output in MATLAB command window. The command '`diary`' is used for saving the output in command window to text file.

Example 2.6.7 can be solved using the sample script provided in Box 2.1. A brief description of each command line is provided at the end of each line after % symbol.

Box 2.1 Sample MATLAB script for solving Example 2.6.7

```
clear all
clc

%Inputs i.e definition of all the distribution functions
syms x y
joint_fun=(x^2)+(x*y)/3; % Given
```

```

%Evaluation of marginal distribution of X and Y
marg_x=int(joint_fun,y,0,2); % marginal
    distribution of X
marg_y=int(joint_fun,x,0,1); % marginal
    distribution of X

%Evaluation of probability of peak discharge (Y)
    being greater than 1 unit.
prob_y_l1=int(marg_y,y,0,1); % probability oy Y
    less than 1 %unit.
prob_y_g1=eval(1-prob_y_l1);

% Evaluation of probability of peak discharge (Y)
    greater %than 1 unit given
%average rainfall intensity (X) is 0.5 units.
cond_y_x=joint_fun/marg_x; %expression for
    conditional %probability of Y given X
xvalue=0.5;
cond_y_xvalue=subs(cond_y_x,xvalue);
prob_y_l1_xvalue=eval(1-int(cond_y_xvalue,y,0,1));

% Output
disp(['The probability of Y greater than 1 units is
     ' num2str(prob_y_g1)])
disp(['The probability of Y greater than 1 units
    given X is 0.5 is ' num2str(prob_y_l1_xvalue)])
```

The output of the code mentioned in Box 2.1 is as follows:

```
The probability of Y greater than 1 unit is 0.58333.
The probability of Y greater than 1 unit given X is 0.5 is 0.6.
```

The solution obtained using the MATLAB code is same as the conclusions drawn from the solution of Example 2.6.7.

Exercise

2.1 Time length (in months) of uninterrupted functioning of soil moisture measuring sensors until failure follows a distribution, $1/7e^{-x/7}$. The sensors are inspected at every 2 months.

(a) What is the probability that the sensors need to be replaced at the first inspection? **(Ans 0.249)**

(b) What is the probability of proper functioning of the sensors till the second scheduled inspection? **(Ans 0.564)**

2.2 Monthly evaporation at a location is measured for last 10 years. Overall 5% data is erroneous.

(a) What is the probability that none of the measurements are erroneous out of 10 randomly selected data? **(Ans 0.586)**
(b) What is the probability that there will be at least one erroneous data out of 10 randomly selected data? **(Ans 0.414)**

2.3 On an average, five flood events in every 2 years are recorded at a location due to heavy rainfall. Number of occurrences of flood events in a year is found to follow a distribution, $\lambda^x \frac{e^{-\lambda}}{x!}$ where λ is the expected number of flood events in a year. What is the probability of occurring not more than two flood events in a particular year at that location? **(Ans 0.543)**

2.4 Droughts in a region are categorized as severe and moderate based on the last 60 years of record. The number of severe and moderate droughts are noted as 6 and 16, respectively. The occurrence of each type of droughts is assumed to be statistically independent and follows a distribution, $\lambda^x \frac{e^{-\lambda}}{x!}$ where λ is the expected number of droughts over a period.

(a) What is the probability that there will be exactly four droughts in the region over the next decade? **(Ans 0.193).**
(b) Assuming that exactly one drought actually occurred in 2 years, what is the probability that it will be a severe drought? **(Ans 0.164).**
(c) Assuming that exactly three droughts actually occurred in 5 years, what is the probability that all will be moderate droughts? **(Ans 0.104).**

2.5 During summer season number of extremely hot days in a city follows a distribution (*pdf*) shown in the Fig. 2.9.

(a) Determine the value of 'a' as shown in the *pdf*. **(Ans 0.08).**
(b) What is the probability of more than 15 extremely hot days in a particular summer season? **(Ans 0.067).**

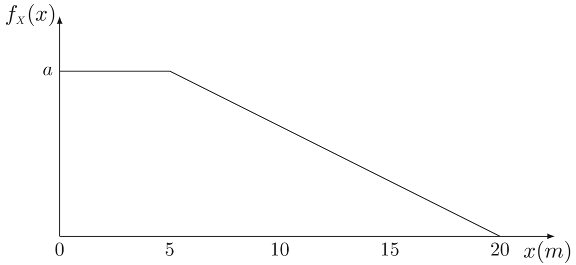


Fig. 2.9 Probability density function for number of floods in a year

2.6 The annual maximum flood level (H) at a river gauging station is approximated to follow a symmetrical triangular distribution over 5–7 m. Values of the *pdf* at the ends and at the midpoint are given in the following table;

Annual maximum flood level (H) in ‘m’	5	6	7
$f_H(h)$	0	1	0

(a) Determine the *pdf* and *CDF* of the flood level.
(b) Determine the maximum flood level that will be exceeded by a probability of 0.05. **(Ans 6.68 m)**.

2.7 A random variable X follows the given distribution,

$$f_X(x) = \begin{cases} Cx^5 & 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Evaluate C such that $f_{X,Y}(x, y)$ is a valid *pdf*. Find the probability that the proportion of X is more than 75%. **(Ans 6, 0.088)**

2.8 The joint *pdf* of random variables X and Y is given as follows,

$$f_{X,Y}(x, y) = 6x^2y \qquad 0 \leq x \leq 1, 0 \leq y \leq 1$$

Evaluate the marginal distributions of X and Y.

2.9 The joint *pdf* of random variables X and Y is given as follows,

$$f_{X,Y}(x, y) = 4xy \qquad 0 < x < 1, 0 < y < 1$$

If random variable $U = X^2$ and $V = XY$, then evaluate the joint *pdf* of U and V. Also evaluate the marginal distribution of U and V.

2.10 The joint *pmf* of X and Y is given as follows. Find out the marginal distribution of X and Y,

Random Variables			*Y*	
		-1	0	1
X	-1	0.00	0.25	0.00
	0	0.25	0.00	0.25
	1	0.00	0.25	0.00

2.11 A random variable X follows the given distribution,

$$f_{X,Y}(x, y) = \begin{cases} 6x & 0 < x < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Evaluate the marginal distributions of X and Y and the expressions for conditional distribution of Y given X.

Chapter 3
Basic Statistical Properties of Data

This chapter starts with some basic exploratory statistical properties from sample data. Concept of moment and expectation, and moment-generating and characteristic functions are considered afterwards. Different methods for parameter estimation build the foundation for many statistical inferences in the field of hydrology and hydroclimatology.

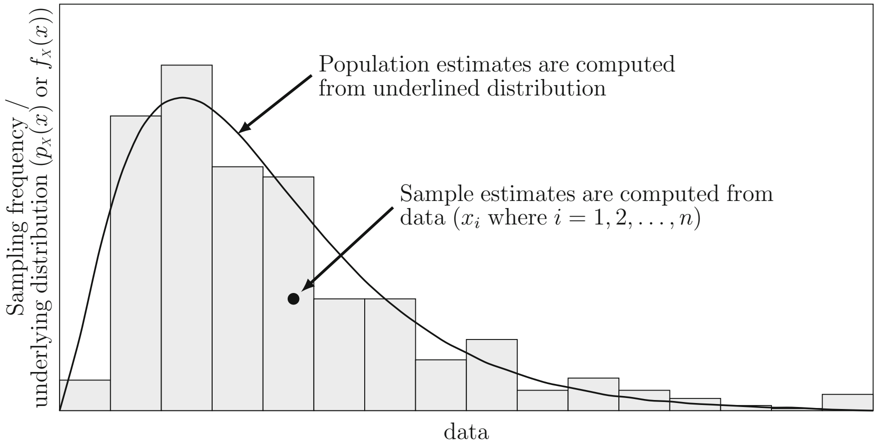
3.1 Descriptive Statistics

The probabilistic characteristics of random variables can be described completely if the form of the distribution function is known and the associated parameters are specified. However, in the absence of knowledge of any parametric distribution, approximate description about the population is assessed through sample statistics. These are also known as descriptive statistics. Some of the most commonly used descriptive statistics are *central tendency*, *dispersion*, *skewness*, and *tailedness*. Respective population parameters are the properties of the underlying probability distribution (Fig. 3.1). Expressions for sample estimates and population parameters are presented simultaneously to facilitate the readers.

3.1.1 Measures of Central Tendency

The measure of central tendency of a random variable can be expressed in terms of three quantities, namely mean, median, and mode. The mean can be further expressed in different forms as discussed in the following sections.

R. Maity, *Statistical Methods in Hydrology and Hydroclimatology*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-10-8779-0_3



Population estimates are computed from underlined distribution

Sample estimates are computed from data (x_i where $i = 1, 2, \ldots, n$)

Fig. 3.1 Frequency plot of a data set with the underlying distribution used to evaluate the sample estimates (from the data) and population parameters (from the underlying distribution)

Arithmetic Mean

Arithmetic mean can be defined as the sum of the observations divided by sample size. Let us consider a sample data set with n observations $x_1, x_2, \ldots, x_n$ for a random variable X. The sample estimate of the population mean (μ) is the arithmetic average $\overline{x}$, calculated as

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.1}$$

In case of grouped data, let us consider k as the number of groups, n as the total number of observations, n_i as the number of observations in the ith group, and x_i as the class mark of the ith group. Class mark is defined as midpoint of the group, i.e., mean of upper and lower bounds of group. For the grouped data, the $\overline{x}$ is given by

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} x_i n_i \tag{3.2}$$

For population, considering $p_X(x_i)$ as the underlying distribution (*pmf*) of a discrete random variable X, the population mean μ is expressed as

$$\mu = \sum_{i=1}^{n} x_i p_X(x_i) \tag{3.3}$$

and considering $f_X(x)$ as the underlying distribution (*pdf*) of a continuous random variable X, the population mean μ is expressed as

$$\mu = \int_{-\infty}^{\infty} x f_x(x)\, dx \tag{3.4}$$

Expressions for population mean are further discussed later with respect to the concept of moment.

Geometric Mean

The geometric mean indicates the central tendency of a data set by using the product of their values. The geometric mean can be defined as the nth root of the product of n observations. The sample geometric mean, $\overline{x}_G$, can be evaluated as

$$\overline{x}_G = \left(\prod_{i=1}^{n} x_i \right)^{1/n} \tag{3.5}$$

where the symbol $\prod$ implies multiplication. The geometric mean can also be expressed as the exponential of the arithmetic mean of logarithms. Thereby, the logarithm of $\overline{x}_G$ is equal to the arithmetic mean of the logarithms of the x_i's. Geometric mean of the population is expressed as:

$$\mu_G = \text{antilog}\left[E\left(\log X \right) \right] \tag{3.6}$$

where $E(\bullet)$ stands for expectation, which is discussed later in Sect. 3.2.

Weighted Mean

The weighted mean is similar to an arithmetic mean except some data points contribute more than others. The calculation of the arithmetic mean of grouped data as explained before is an example of weighted means where n_i / n is the weighted factor. In general, the weighted mean is

$$\overline{x}_w = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i} \tag{3.7}$$

where w_i is the weight associated with the ith observation or group and k is the number of observations or groups.

Median

The median is the value of the random variable at which the values on both sides of it are equally probable. This can be particularly used if one desires to eliminate the effect of extreme values as mean is highly influenced by the extreme values. The median of n observations can be defined as the value of $(n+1)/2$ numbered observation (the observations are arranged in ascending order) in case n is odd and average of two observations in position $n/2$ and $n/2+1$ in case n is an even number. Thereby, we can say that sample median $\overline{x}_{md}$ is the observation such that half of the values lie on either side of $\overline{x}_{md}$.

Considering X to be a discrete random variable, the population median $\mu_{md} = x_d$ where d is determined from

$$\sum_{i=1}^{d} p_X(x_i) = 0.5 \tag{3.8}$$

Considering X to be a continuous random variable, the population median μ_{md} would be the value satisfying

$$\int_{-\infty}^{\mu_{md}} f_X(x)\,dx = 0.5 \tag{3.9}$$

Mode

The mode is the most probable or most frequently occurring value of a random variable. It is the value of the random variable with the highest probability density or the most frequently occurring value. A sample or a population may have none, one, or more than one mode. Thus, the population mode, μ_{mo}, would be a value of X maximizing *pmf* or *pdf*.

Considering X to be a discrete random variable with *pmf* $p_X(x)$, the mode is the value of x_i for which $p_X(x_i)$ is maximum, i.e.,

$$\mu_{mo} = \underset{x_i}{\arg\max}\,[p_X(x_i)] \tag{3.10}$$

Considering X to be a continuous random variable with *pdf* $f_X(x)$, the mode is the value of X that satisfies the following equation

$$\frac{df_X(x)}{dx} = 0 \quad \text{and} \quad \frac{d^2 f_X(x)}{dx^2} < 0 \tag{3.11}$$

3.1.2 *Measure of Dispersion*

The dispersion of a random variable corresponds to how closely the values of a random variable are clustered or how widely it is spread around the central value. Figure 3.2 shows two random variables, X and Y, with same mean but dispersion of Y is more than X.

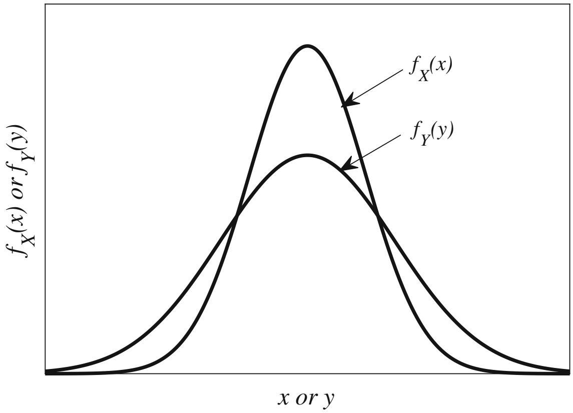


Fig. 3.2 Random variables X and Y with same mean but different dispersion

Range

The range of a sample is the difference between the maximum and the minimum values in the sample. The minimum and the maximum values also convey information about the variability present in data. The range has the disadvantage of not reflecting the frequency or magnitude of values that deviate either positively or negatively from the mean since only the largest and smallest values are used in its determination. Occasionally, the relative range is used which is the range divided by the mean.

Variance

Variance (S^2) is a measure of the dispersion of a random variable taking the mean as the central value. For a sample of size n, the variance is the average squared deviation from the sample mean.

Considering X as a random variable and a sample $x_1, x_2, \ldots, x_n$ with sample mean $\overline{x}$, the differences $x_1 - \overline{x}, x_2 - \overline{x}, \ldots, x_n - \overline{x}$ are called the deviations from the mean. The sample estimate of variance can be defined as the average of the squared deviations from the mean. The sample estimate of population variance σ^2 is denoted by S^2 and is given as

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} \tag{3.12}$$

The reason for dividing by $n - 1$ instead of n is to make the estimator unbiased. Unbiasedness is one of the four properties that an estimator should possess. These properties are explained later in Sect. 3.6. For the time being, readers may note that one degree of freedom is lost while estimating the sample mean ($\overline{x}$) from the data.

For the grouped data with $x_1, x_2, \ldots, x_k$ as the class mark, the variance can be estimated from the following formula

$$S^2 = \frac{\sum_{i=1}^{k} (x_i - \overline{x})^2 n_i}{n - 1} \tag{3.13}$$

where k is the number of groups, n is the total number of observations, x_i is the class mark, and n_i is the number of observations in the ith group.

Standard deviation, another measure of dispersion, is the positive square root of variance, and the unit for standard deviation is the same as the unit of the X. The formula for S is as follows:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n - 1}} \tag{3.14}$$

A dimensionless measure of dispersion is the coefficient of variation defined as the standard deviation divided by the mean. The coefficient of variation is estimated as

$$C_v = \frac{S}{\overline{x}} \tag{3.15}$$

Higher values indicate more dispersed data, i.e., high variability about mean and *vice versa*. Population estimate of variance (denoted as σ^2) is discussed later in Sect. 3.2.1.

3.1.3 Measure of Symmetry

Distributions of data may not be symmetrical with respect to its mean; i.e., they may tail off to the right or to the left. Such distributions are said to be skewed (Fig. 3.3). Skewness of the data is measured using the coefficient of skewness (γ). For positive skewness (coefficient of skewness, $\gamma > 0$), the data is skewed to the right and similarly for negative skewness ($\gamma < 0$) the data is skewed to the left. The difference between the mean and the mode indicates the skewness of the data. The sample estimate skewness is normally made dimensionless by dividing by S^3 to get the coefficient of skewness. A sample estimate of coefficient of skewness (denoted as C_s) is expressed as

$$C_s = \frac{n \sum_{i=1}^{n} (x_i - \overline{x})^3}{(n - 1)(n - 2) S^3} \tag{3.16}$$

Population estimate for coefficient of skewness (denoted by γ) is discussed later in Sect. 3.2.1.

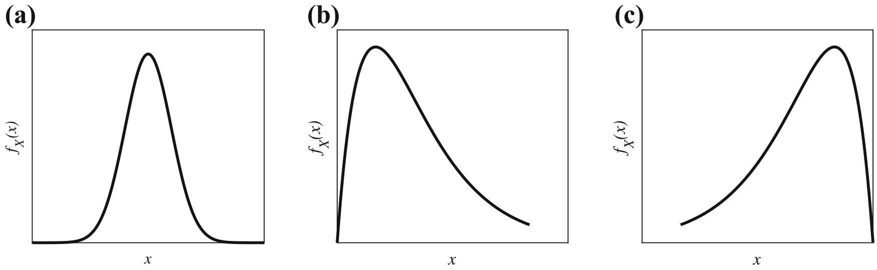


Fig. 3.3 Typical *pdf* plots of a **a** symmetric, **b** positively skewed distribution, and **c** negatively skewed distribution

3.1.4 Measure of Tailedness

The measure of tailedness of a probability distribution function is referred to as kurtosis. Being a measure of tailedness, kurtosis provides important interpretation about the tails, i.e., outlier. For a sample, kurtosis shows the effect of existing outliers. However, for a distribution, kurtosis shows the propensity to produce outliers. The kurtosis is made dimensionless by dividing by S^4 to get the coefficient of kurtosis. Coefficient of kurtosis is a convenient non-dimensional measure of tailedness. The sample estimate of the coefficient of kurtosis is given by

$$k = \frac{n^2 \sum_{i=1}^{n} (x_i - \overline{x})^4}{(n-1)(n-2)(n-3) S^4} \tag{3.17}$$

A particular distribution can be classified on the basis of its tailedness when compared with a standard value. Generally, the standard value taken is the kurtosis of normal distribution that has a value of 3. Thus, sometimes another estimate, $\varepsilon = k - 3$, is also used as a measure of kurtosis. Based on the measure of kurtosis, data or the associated distribution can be divided into three types (Fig. 3.4) as follows:

(i) *Mesokurtic*: If any distribution has same kurtosis as compared to normal distribution, the distribution is called mesokurtic. Thus, for a mesokurtic distribution, $k = 3$ and $\varepsilon = 0$.

(ii) *Leptokurtic*: In case a distribution has a relatively greater concentration of probability near the mean than the normal distribution; the kurtosis will be greater than 3. The value of ε will be positive.

(iii) *Platykurtic*: In case a distribution has a relatively smaller concentration of probability near the mean than the normal distribution; the kurtosis will be less than 3. The value of ε will be negative.

Population estimate for coefficient of kurtosis is discussed later in Sect. 3.2.1.

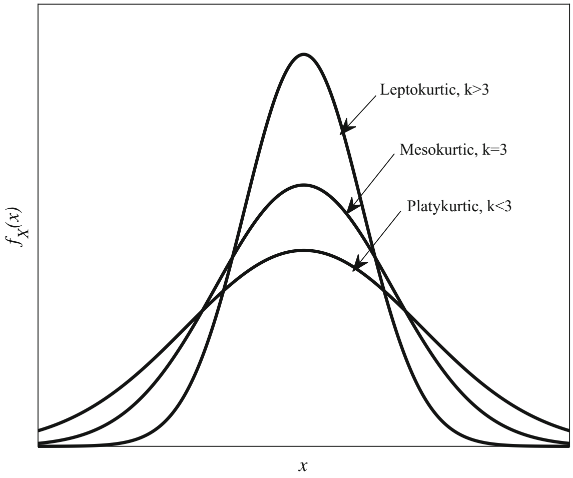


Fig. 3.4 A typical *pdf* plot showing the three zones of kurtosis, namely leptokurtic, mesokurtic, and platykurtic

Example 3.1.1

Consider the following sample data for annual peak discharge (cumec) at a gauging station A. Evaluate the mean, variance, coefficient of skewness, and coefficient of kurtosis for the given sample data. Also, comment regarding the coefficient of skewness and coefficient of kurtosis.

Year	2000	2001	2002	2003	2004	2005	2006	2007
Annual peak discharge (cumec)	4630	2662	1913	3655	3670	4005	4621	1557
Year	2008	2009	2010	2011	2012	2013	2014	2015
Annual peak discharge (cumec)	2405	1625	6216	2602	2157	3120	6403	2934

Solution The mean for the given sample data for peak annual discharge can be evaluated as

$$
\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
&= \frac{4630 + 2662 + \cdots + 2934}{16} \\
&= 3385.93\,\text{cumec}
\end{aligned}
$$

The variance of the sample data can be evaluated as

$$
\begin{aligned}
S^2 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \\
&= \frac{(46330 - 3385.93)^2 + (2662 - 3385.93)^2 + \cdots + (2934 - 3385.93)^2}{15} \\
&= 2214860.86 \approx 2.2 \times 10^6 \text{ cumec}^2
\end{aligned}
$$

The coefficient of skewness of the sample data can be evaluated as

$$
\begin{aligned}
C_s &= \frac{n\sum_{i=1}^{n}(x_i - \overline{x})^3}{(n-1)(n-2)S^3} \\
&= \frac{16\left[(46330 - 3385.93)^3 + (2662 - 3385.93)^3 + \cdots + (2934 - 3385.93)^3\right]}{15 \times 14 \times (2214860.86)^{3/2}} \\
&= 0.745
\end{aligned}
$$

The coefficient of kurtosis of the sample data can be evaluated as

$$
\begin{aligned}
k &= \frac{n^2\sum_{i=1}^{n}(x_i - \overline{x})^4}{(n-1)(n-2)(n-3)S^4} \\
&= \frac{16^2\left[(46330 - 3385.93)^4 + (2662 - 3385.93)^4 + \cdots + (2934 - 3385.93)^4\right]}{15 \times 14 \times 13 \times (2214860.86)^2} \\
&= 2.628
\end{aligned}
$$

As the value of coefficient of skewness is positive so the data is *positively skewed* and coefficient of kurtosis is less than 3, so it is *platykurtic*.

3.2 Concept of Moments and Expectation

In physics, moment is the product of a physical quantity and the distance from a fixed point of reference. While considering mass as the physical quantity, it can be used to locate the center of gravity of any irregularly shaped object. Higher order of moments can also be evaluated. Similar concepts can be utilized to extract some meaningful information from a data set (Fig. 3.5a).

Suppose that the data $x_1, x_2, \ldots, x_n$ is located according to their values on the real line as shown in Fig. 3.5a. Assuming that each data value is equiprobable, the mass of each data can be assumed to be $1/n$, when n is the length of the data. Now, the total moment with respect to origin can be evaluated as $\sum_{\text{all } i} x_i \left(1/n\right)$. We may find out the locations say $\tilde{x}$ of the equivalent total mass, i.e., the mass that will create the

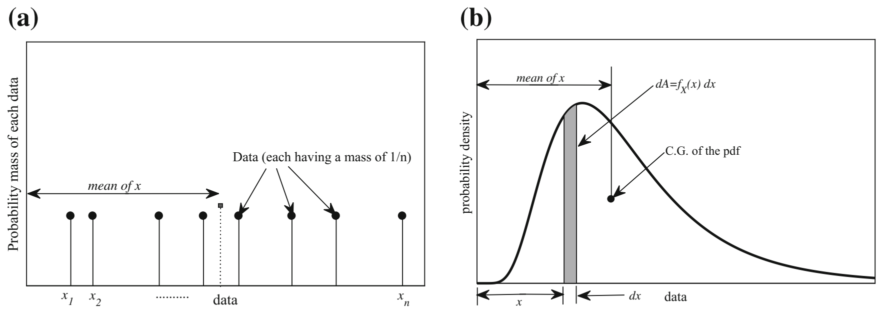


Fig. 3.5 First moment (mean) of the data for **a** discrete data and **b** probability density function of continuous data

same moment (as the total moment) about the origin, expressed as $\left(n \times {}^{1}\!/_{n}\right)\tilde{x} = \tilde{x}$. Equating these two moments, we get

$$\tilde{x} = \sum_{\text{all } i} x_i(1/n) \tag{3.18}$$

This location $(\tilde{x})$ is equivalent to the mean of the data $(\overline{x})$.

In case of the population which is represented by a *pdf*, mean can be identified following the same concept. Referring to Fig. 3.5b, consider a delta width (dx) located at a distance x from the origin. The total probability mass is equal to the area above dx and below the *pdf* (shaded area $= dA$). Total moment for this area with respect to origin is $x.dA = x.f_X(x)dx$. Integrating for the entire range of the data $(-\infty, \infty)$, the total moment can be written as $\int\limits_{-\infty}^{\infty} xf_X(x)dx$. If it is assumed that the total probability mass is located at a distance x from origin that produces same amount of moment, we may write

$$\mu \times \int_{-\infty}^{\infty} f_X(x)dx = \int_{-\infty}^{\infty} xf_X(x)dx$$

Since $\int\limits_{-\infty}^{\infty} f_X(x)dx = 1$

$$\mu = \int_{-\infty}^{\infty} xf_X(x)dx \tag{3.19}$$

Following the same concept, higher order moments with respect to origin can also be evaluated using some power of distance from the origin; for example, x^2 and x^3 can be used to evaluate the second- and third-order moments, respectively. The x in Eq. 3.19 can be replaced with x^i to evaluate the ith moment with respect to origin.

However in probability theory, second moment onwards are calculated with respect to the mean. First moment with respect to the mean is zero. The second-order moment with respect to the mean can be evaluated as

$$E\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 \, f_X(x)\, dx \tag{3.20}$$

In general, the ith-order moment with respect to the mean can be evaluated as

$$E\left[(X-\mu)^i\right] = \int_{-\infty}^{\infty} (x-\mu)^i \, f_X(x)\, dx \tag{3.21}$$

3.2.1 Expectation

The expected value of a random quantity intuitively means the averaged value of the outcome of the corresponding random experiment carried out repetitively for infinite times. Mathematically, the expected value of a random variable (X), represented as $E(X)$, can be defined as the first moment about the origin and represented as follows:

$$E(X) = \mu \tag{3.22}$$

Considering X to be a discrete random variable, the expected value of X is given as

$$E\,(X) = \sum_{all\ j} x_j \, p_X\left(x_j\right) \tag{3.23}$$

and for continuous random variables, the expected value of X is given as

$$E\,(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx \tag{3.24}$$

Any function of X, say $g(X)$, is also a random variable. Thus, the expected value of $g(X)$ is given as

$$E\left[g\,(X)\right] = \sum_{\text{all } j} g\left(x_j\right) p_X\left(x_j\right) \qquad \text{for discrete RV} \tag{3.25}$$

$$E\left[g\,(X)\right] = \int_{-\infty}^{\infty} g\,(x)\, f_X(x)\, dx \qquad \text{for continuous RV} \tag{3.26}$$

Relating the concept of moment with expectation, following points can be noted

(i) The *first moment* about *origin* is the mean, i.e.,

$$E\left(X\right)=\mu=\begin{cases}\sum_{\text{all } j}x_jp_X\left(x_j\right) & \text{for discrete RV}\\ \int_{-\infty}^{\infty}xf_X\left(x\right)dx & \text{for continuous RV}\end{cases} \tag{3.27}$$

(ii) The *second moment* about the *mean* is the *variance*

$$E\left[(X-\mu)^2\right]=\sigma^2=\begin{cases}\sum_{\text{all } j}\left(x_j-\mu\right)^2p_X\left(x_j\right) & \text{for discrete RV}\\ \int_{-\infty}^{\infty}(x-\mu)^2f_X\left(x\right)dx & \text{for continuous RV}\end{cases} \tag{3.28}$$

It can also be shown that

$$V\left(x\right)=E\left(x^2\right)-\left[E\left(x\right)\right]^2 \tag{3.29}$$

(iii) The *third moment* about the *mean* is the *skewness*

$$E\left[(X-\mu)^3\right]=\begin{cases}\sum_{\text{all } j}\left(x_j-\mu\right)^3p_X\left(x_j\right) & \text{for discrete RV}\\ \int_{-\infty}^{\infty}(x-\mu)^3f_X\left(x\right)dx & \text{for continuous RV}\end{cases} \tag{3.30}$$

It can also be shown that

$$E\left[(x-\mu)^3\right]=E\left(x^3\right)-3E\left(x^2\right)E\left(x\right)+2\left\{E\left(x\right)\right\}^3 \tag{3.31}$$

The measure of skewness is non-dimensionalized using variance and termed as *coefficient of skewness* (γ). Thus, γ is expressed as

$$\gamma=\frac{E\left[(X-\mu)^3\right]}{\sigma^3} \tag{3.32}$$

(iv) The *fourth moment* about the *mean* is the *kurtosis* (measure of tailedness)

$$E\left[(X-\mu)^4\right]=\begin{cases}\sum_{all\ j}\left(x_j-\mu\right)^4p_X\left(x_j\right) & \text{for discrete RV}\\ \int_{-\infty}^{\infty}(x-\mu)^4f_X\left(x\right)dx & \text{for continuous RV}\end{cases} \tag{3.33}$$

It can also be shown that

$$E\left[(x-\mu)^4\right]=E\left(x^4\right)-4E\left(x^3\right)E\left(x\right)+6E\left(x^2\right)\left(E(x)\right)^2-3\left\{E\left(x\right)\right\}^4 \tag{3.34}$$

The measure of tailedness (*kurtosis*) is also non-dimensionalized using variance and termed as *coefficient of kurtosis* (κ). Thus, κ is expressed as

$$\kappa=\frac{E\left[(X-\mu)^4\right]}{\sigma^4} \tag{3.35}$$

Table 3.1 Population parameters and sample statistics

Property	Parameter name	Population parameter	Sample statistic
Central tendency	Arithmetic mean	For discrete case, $\mu = \sum x p_X(x)$ For continuous case, $\mu = \int_{-\infty}^{\infty} x f_X(x) dx$	$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
	Geometric mean	$\mu_G = \text{antilog}\left[E\left(\log X\right)\right]$	$\overline{x}_G = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$
	Median	X such that $F(x) = 0.5$	50th percentile value of data
	Mode	For discrete case, $\mu_{mo} = \underset{x_i}{\arg\max}\left[p_X(x_i)\right]$ For continuous case, μ_{mo} is the root of $\frac{df_X(x)}{dx} = 0$ and $\frac{d^2 f_X(x)}{dx^2} < 0$	Most frequently occurring data
Variability	Variance	$\sigma^2 = E\left[(X-\mu)^2\right]$	$S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
	Standard deviation	$\sigma = E\left[(X-\mu)^2\right]^{1/2}$	$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$
	Coefficient of variation	$c_v = \frac{\sigma}{\mu}$	$CV = \frac{S}{\overline{x}}$
Symmetry	Coefficient of skewness	$\gamma = \frac{E\left[(X-\mu)^3\right]}{\sigma^3}$	$C_s = \frac{n\sum_{i=1}^{n}(x_i-\overline{x})^3}{(n-1)(n-2)S^3}$
Tailedness	Coefficient of kurtosis	$\kappa = \frac{E\left[(X-\mu)^4\right]}{\sigma^4}$	$k = \frac{n^2\sum_{i=1}^{n}(x_i-\overline{x})^4}{(n-1)(n-2)(n-3)S^4}$

Population parameters and corresponding sample estimates of different descriptive statistics are shown in Table 3.1.
Some useful information on the expected values is:

(i) Expectation of a constant is same as that constant, i.e., $E(C) = C$.
(ii) Expectation of a modified random variable obtained by multiplying with a constant is equal to the product of the constant and the expectation of the original random variable, i.e., $E(CX) = C\,E(X)$.
(iii) Expectation of a random variable obtained by addition/subtraction of two random variables is equal to the sum/difference of their individual expectations, i.e., $E(X \pm Y) = E(X) \pm E(Y)$.

Some useful information on the variance values is:

(i) Variance of a constant is zero, i.e., $V(C) = 0$.
(ii) Variance of a modified random variable obtained by multiplying with a constant is equal to the product of the square of constant and the variance of the original random variable, i.e., $V(CX) = C^2\,V(X)$.

(iii) Variance of a modified random variable obtained by multiplying with a constant (a) and addition to another constant (b) is equal to the product of the square of constant (a) and the variance of the original random variable, i.e., $V\,(aX + b) = a^2\,V\,(X)$.

Example 3.2.1

The number of thunderstorms per year (X) and its *pmf* obtained from the historical data are shown in the following table:

X	0	1	2	3
pmf $(p_X(x))$	0.3	0.4	0.2	0.1

What is the mean and variance of the number of thunderstorms in a year?

Solution The mean of number of thunderstorms per year can be evaluated as

$$\begin{aligned} E(x) &= \sum x p_X(x) \\ &= 0 \times 0.3 + 1 \times 0.4 + 2 \times 0.2 + 3 \times 0.1 \\ &= 1.1 \end{aligned}$$

The variance of storms can be evaluated as

$$\begin{aligned} V(x) &= E(x^2) - \{E(x)\}^2 \\ &= [1^2 \times 0.4 + 2^2 \times 0.2 + 3^2 \times 0.1] - 1.1^2 \\ &= 0.89 \end{aligned}$$

Example 3.2.2

The time (T) between two successive floods follows following *pdf*

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t} & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases}$$

Find the mean, mode, median, and the coefficient of variation of T.

Solution The mean time between successive floods is given by

$$E(T) = \int_0^\infty t\lambda e^{-\lambda t} dt = -\int_0^\infty t\ d\left(e^{-\lambda t}\right)$$

Integrating by parts (i.e., $\int udv = uv - \int vdu$), we get

$$E(T) = -te^{-\lambda t}\Big|_0^\infty + \int_0^\infty e^{-\lambda t}dt = \left[-\frac{e^{-\lambda t}}{\lambda}\right]_0^\infty = \frac{1}{\lambda}$$

Hence, the mean time between successive floods is $\overline{t} = 1/\lambda$.

The mode is the value of t with the maximum value of *pdf*. From the *pdf*, it can be observed that the probability density is highest at $t = 0$. Thus, the mode is $\mu_{mo} = 0$ The median can be evaluated as

$$\int_0^{\mu_{md}} \lambda e^{-\lambda t} dt = 0.5$$

$$\text{or, } 1 - e^{-\lambda \mu_{md}} = 0.5$$

$$\text{or, } \mu_{md} = \frac{-\ln(0.5)}{\lambda} = \frac{0.693}{\lambda}$$

Therefore, median is $\mu_{md} = \frac{0.693}{\lambda}$.

The variance can be evaluated as

$$\sigma_T^2 = \int_0^\infty \left(t - \frac{1}{\lambda}\right)^2 \lambda e^{-\lambda t} dt = \int_0^\infty \left(\lambda t^2 - 2t + \frac{1}{\lambda}\right) e^{-\lambda t} dt$$

Integrating by parts (as done for $E(T)$), we get

$$\text{First term, } \int_0^\infty t^2 \lambda e^{-\lambda t} dt = -\int_0^\infty t^2 d\left(e^{-\lambda t}\right) = -\frac{2}{\lambda}$$

$$\text{Second term, } -2\int_0^\infty te^{-\lambda t}dt = 2\int_0^\infty \frac{t}{\lambda} d\left(e^{-\lambda t}\right) = \frac{2}{\lambda}$$

$$\text{Third term, } \int_0^\infty \frac{1}{\lambda} e^{-\lambda t} dt = \frac{1}{\lambda^2}$$

$$\text{Hence, } \sigma_T^2 = \frac{2}{\lambda} - \frac{2}{\lambda} + \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

The standard deviation is given by $\sigma_T = \frac{1}{\lambda}$.

The coefficient of variation of the given distribution is $c_v = \frac{\sigma_T}{\mu_T} = \frac{1/\lambda}{1/\lambda} = 1$.

Example 3.2.3

The rainfall depth (in cm) received during thunderstorms at a place (X) is a random variable with the following density function

$$f_X(x) = \begin{cases} \frac{3}{2500}(x-10)(x-20) & 0 \le x \le 10 \\ 0 & \text{elsewhere} \end{cases}$$

Determine the following

(a) Mean value of X;
(b) Median of X;
(c) Mode of X;
(d) Standard deviation of X;
(e) Coefficient of variation of X; and
(f) Skewness coefficient.

Solution The density function is $f_x\left(x\right)=\frac{3}{2500}\left(x-10\right)\left(x-20\right)$ for $0\leq x\leq 10$.

(a) Mean (μ)

$$\mu=\int_0^{10} xf_x(x)dx=\frac{3}{2500}\int_0^{10} x\left(x-10\right)\left(x-20\right)dx$$
$$=\frac{3}{2500}\left[\frac{x^4}{4}-10x^3+100x^2\right]_0^{10}=3$$

(b) Median (μ_{md})

$$\int_0^{\mu_{md}} f_x(x)dx=0.5$$
$$\text{or, }\frac{3}{2500}\int_0^{\mu_{md}}\left(x-10\right)\left(x-20\right)dx=0.5$$
$$\text{or, }\frac{3}{2500}\left[\frac{x^3}{3}-15x^2+200x\right]_0^{\mu_{md}}=0.5$$
$$\text{or, }\mu_{md}=2.5398$$

(c) Standard deviation σ

$$\sigma^2=\int_0^{10}(x-\overline{x})^2 f_x(x)dx$$
$$\text{or, }\sigma^2=\frac{3}{2500}\int_0^{10}(x-3)^2(x-10)(x-20)dx$$
$$\text{or, }\sigma^2=\frac{1}{12500}\left[3x^5-135x^4+1945x^3-11025x^2+27000x\right]_0^{10}$$
$$\text{Hence, }\sigma=\sqrt{5}=2.2361$$

(d) Coefficient of variation is calculated as

$$c_v=\frac{\sigma}{\mu}=\frac{2.236}{3}=0.7454\approx 74.5\%.$$

(e) Coefficient of skewness is obtained as

$$\begin{aligned}
\gamma &= \left(\int_0^{10} (x-\mu)^3 f_x(x)dx\right) \Big/ \sigma^3 \\
&= \frac{3}{2500 \times 5\sqrt{5}} \int_0^{10} (x-3)^3\,(x-10)\,(x-20)\,dx \\
&= \frac{1}{250000\sqrt{5}} \Big[x(10x^5 - 468x^4 + 7455x^3 - 52740x^2 + 186300x - 324000)\Big]_0^{10} \\
&= 0.7155
\end{aligned}$$

3.3 Moment-Generating Functions

Moment-generating function of a random variable is generally treated as an alternative to its probability distribution. Though all the random variables may not have moment-generating functions, however, if available, these are sometimes easier to compute moments of the random variables of any desired order.

Expectation of e^{tX}, which is a function of the random variable X, is known as moment-generating function of the random variable X. It can be represented as

$$M_X\,(t) = E\left(e^{tX}\right) \tag{3.36}$$

In case of discrete random variable, the moment-generating function can be evaluated as

$$M_X\,(t) = \sum_{all\ j} e^{tx_j}\,p_x\left(x_j\right) \tag{3.37}$$

In case of continuous random variable, the moment-generating function can be evaluated as

$$M_X\,(t) = \int_{-\infty}^{\infty} e^{tx}\,f_x\,(x)\,dx \tag{3.38}$$

We can show that the Taylor series expansion of $M_X\,(t)$ is

$$M_X\,(t) = 1 + \mu\,t + \mu_2^{'}\frac{t^2}{2} + \cdots + \mu_k^{'}\frac{t^k}{k} + \cdots \tag{3.39}$$

The kth moment about origin is then found to be the kth derivative of $M_X(t)$ with respect to t and evaluated at $t = 0$.

$$\mu_k^t = \left.\frac{d^k M_X\,(t)}{dt^k}\right|_{t=0} \tag{3.40}$$

Usefulness of the moment generation function can be explored by evaluating the derivatives of the function. First derivative of $M_X(t)$, evaluated at $t = 0$, results in the expected value, which is first moment of the random variable with respect to origin. Mathematically,

$$\left.\frac{dM_X(t)}{dt}\right|_{t=0} = \int_{-\infty}^{\infty} x f_x(x)\,dx \tag{3.41}$$

Similarly, second derivative of $M_X(t)$, evaluated at $t = 0$, results in second moment of the random variable with respect to origin. Thus,

$$\left.\frac{d^2M_X(t)}{dt^2}\right|_{t=0} = \int_{-\infty}^{\infty} x^2 f_x(x)\,dx \tag{3.42}$$

In general, nth derivative of $M_X(t)$, evaluated at $t = 0$, results in nth moment of the random variable with respect to origin.

$$\left.\frac{d^nM_X(t)}{dt^n}\right|_{t=0} = \int_{-\infty}^{\infty} x^n f_x(x)\,dx \tag{3.43}$$

Example 3.3.1

Consider a data set to follow the given distribution where λ is a constant. Evaluate the first moment with respect to origin, the second moment with respect to the mean and the moment-generating function.

$$p_x(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0, 1, 2, \ldots$$

Solution Calculation for the first moment with respect to origin otherwise known as mean

$$E(x) = \sum_0^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} = \lambda\sum_1^{\infty}\frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda}\sum_1^{\infty}\frac{\lambda^{x-1}}{(x-1)!} = \lambda$$

Calculation for the second moment with respect to mean otherwise known as variance

$$V(x) = E\left(x^2\right) - [E(x)]^2$$

In above expression, $E\left(x^2\right)$ can also be expressed as

$$E\left(x^2\right) = E[x(x-1)] + E(x)$$

$$E[x(x-1)] = \sum_0^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} = \lambda^2\sum_1^{\infty}\frac{e^{-\lambda}\lambda^{x-2}}{(x-2)!} = \lambda^2 e^{-\lambda}\sum_1^{\infty}\frac{\lambda^{x-1}}{(x-1)!} = \lambda^2$$

$$V(x) = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Calculation for the moment-generating function

$$E\left(e^{tx}\right) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\left(\lambda e^t\right)^x}{x!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda\left(e^t - 1\right)}$$

3.4 Characteristic Functions

Similar to moment-generating function, characteristic function of a random variable may also serve as another alternative to its probability distribution. It is the Fourier transform of the probability density function of the random variable.

The expectation of e^{itX} (where $i = \sqrt{-1}$), which is a complex function of the random variable X, is known as characteristic function of that random variable X. It can be defined as

$$\phi_X(t) = E\left(e^{itX}\right) = M_X(it) \tag{3.44}$$

The characteristic function for X can be expressed as

$$\phi_X(t) = \sum_{\text{all } j} e^{itx_j} p_X\left(x_j\right) \qquad \text{for discrete RV} \tag{3.45}$$

$$\phi_X(t) = \int_{-\infty}^{\infty} f_X(x)\, e^{itx} dx \qquad \text{for continuous RV} \tag{3.46}$$

Using the characteristic function, the nth moment of X can be expressed as

$$E\left(X^n\right) = \frac{1}{i^n} \left.\frac{d^n \phi_X(t)}{dt^n}\right|_{t=0} \tag{3.47}$$

Example 3.4.1
Consider a random variable X that follows the given distribution. Evaluate the mean, variance, skewness, and kurtosis.

X	0	25	60	75	100
$p_X(x)$	0.5	0.24	0.12	0.08	0.06

Solution Mean can be evaluated as

$$E\left(x\right) = 0 \times 0.5 + 25 \times 0.24 + 50 \times 0.12 + 75 \times 0.08 + 100 \times 0.06 = 24$$

Variance can be evaluated as

$$\begin{aligned}
V(x) &= E\left[(x-\mu)^2\right] = E\left(x^2\right) - \{E\left(x\right)\}^2\\
E\left(x^2\right) &= 0^2 \times 0.5 + 25^2 \times 0.24 + 50^2 \times 0.12 + 75^2 \times 0.08 + 100^2 \times 0.06 = 1500\\
V(x) &= 1500 - 24^2 = 924
\end{aligned}$$

Skewness can be evaluated as

$$\begin{aligned}
E\left[(x-\mu)^3\right] &= E\left(x^3\right) - 3E\left(x^2\right)E\left(x\right) + 2\{E\left(x\right)\}^3\\
E\left(x^3\right) &= 0^3 \times 0.5 + 25^3 \times 0.24 + 50^3 \times 0.12 + 75^3 \times 0.08 + 100^3 \times 0.06 = 112500\\
E\left[(x-\mu)^3\right] &= 112500 - 3 \times 1500 \times 24 + 2 \times 24^3 = 32148
\end{aligned}$$

Kurtosis can be evaluated as

$$\begin{aligned}
E\left[(x-\mu)^4\right] &= E\left(x^4\right) - 4E\left(x^3\right)E\left(x\right) + 6E\left(x^2\right)(E(x))^2 - 3\{E\left(x\right)\}^4\\
E\left(x^4\right) &= 0^4 \times 0.5 + 25^4 \times 0.24 + 50^4 \times 0.12 + 75^4 \times 0.08 + 100^4 \times 0.06 = 9375000\\
E\left[(x-\mu)^4\right] &= 9375000 - 4 \times 112500 \times 24 + 6 \times 1500 \times 24^2 - 3 \times 24^4 = 2763672
\end{aligned}$$

Example 3.4.2

Consider a continuous random variable X having the following marginal distribution. Evaluate the mean, variance, and median.

$$f_X\left(x\right) = \begin{cases} \frac{2}{x^3} & \text{for } x > 1\\ 0 & \text{elsewhere} \end{cases}$$

Solution Mean can be evaluated as

$$\begin{aligned}
E\left(x\right) &= \int x\, f_x(x)dx\\
&= \int_1^{\infty} x \times \frac{2}{x^3}dx\\
&= \left[\frac{-2}{x}\right]_1^{\infty}\\
&= 2
\end{aligned}$$

Variance can be evaluated as

$$
\begin{aligned}
V(x) &= E\left(x^2\right) - \{E(x)\}^2 \\
&= \int x^2 f_X(x)\,dx - \{E(x)\}^2
\end{aligned}
$$

$$
\text{where } \int x^2 f_X(x)\,dx = \int_1^\infty x^2 \times \frac{2}{x^3}dx
$$

The integral does not exist; thereby, the variance does not exist.

For the calculation of median, we first need to calculate the *CDF* of X.

$$
F_X(x) = \int_1^x f_X(x)dx = 1 - \frac{1}{x^2}
$$

$$
\text{Hence, } F_X(x) = \begin{cases} 1 - \frac{1}{x^2} & x \geq 1 \\ 0 & \text{elsewhere} \end{cases}
$$

Median can be evaluated as

$$
\begin{aligned}
F_X(\mu_{md}) &= 0.5 \\
\text{or, } 1 - \frac{1}{\mu_{md}^2} &= \frac{1}{2} \\
\text{or, } \mu_{md} &= \sqrt{2}
\end{aligned}
$$

Example 3.4.3

Evaluate the coefficient of variation, coefficient of skewness, and coefficient of kurtosis for the data supplied in Example 3.4.1. Also, provide an insight into the skewness and tailedness of the distribution.

Solution Calculation for coefficient of variation:

$$
c_v = \frac{\sigma}{\mu} = \frac{\sqrt{924}}{24} = 1.266
$$

Calculation for coefficient of skewness:

$$
\gamma = \frac{E\left[(x-\mu)^3\right]}{\sigma^3} = \frac{32148}{924^{3/2}} = 1.144
$$

Calculation for coefficient of kurtosis:

$$
\kappa = \frac{E\left[(x-\mu)^4\right]}{\sigma^4} = \frac{2763672}{924^2} = 3.237
$$

As γ is positive, so the distribution is positively skewed, and κ-3 is positive so the distribution is leptokurtic.

Example 3.4.4
The 30 years of monthly rainfall data (mm) at rain gauge stations A and B are found to follow the given distribution.

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

If the probability of rainfall exceeding 50 mm is 0.135 for station A and 0.188 for station B, which station receives higher mean rainfall?

Solution In order to determine the station receiving higher mean rainfall, we first have to evaluate the mean for the above-mentioned distribution, i.e., $E(X)$. Hence, from Example 3.2.2,

$$E(X) = \int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

Further, the probability that rainfall exceeds 50 mm is given by

$$P(x > 50) = \int_{50}^\infty \lambda e^{-\lambda x} dx = e^{-50\lambda}$$

For station A:

$$\begin{aligned} &e^{-50\lambda} = 0.135 \\ &\lambda = 0.04 \\ &\text{Thus, } f_X(x) = 0.04e^{-0.04x} \\ &\mu = \int_0^\infty x f_X(x) dx = 25 \end{aligned}$$

Similarly, for station B:

$$\begin{aligned} &e^{-50\lambda} = 0.188 \\ &\lambda = 0.033 \\ &\text{Thus, } f_X(x) = 0.033e^{-0.033x} \\ &\mu = \int_0^\infty x f_X(x) dx = 30.30 \end{aligned}$$

Therefore, station B receives higher mean rainfall.

3.5 Statistical Properties of Jointly Distributed Random Variables

3.5.1 *Expectation*

If X and Y are considered to be jointly distributed continuous random variable and U is some function of X and Y, $U = g(X, Y)$, then expectation of U, $E(U)$ can be written as

$$E\left(U\right) = E\left[g\left(X, Y\right)\right] = \int u\, f_U\left(u\right)\, du \tag{3.48}$$

In case of continuous random variables,

$$E\left[g\left(X, Y\right)\right] = \int\int g\left(x, y\right) f_{X,Y}\left(x, y\right)\, dx\, dy \tag{3.49}$$

In case of discrete random variables,

$$E\left[g\left(X, Y\right)\right] = \sum_i \sum_j g\left(x_i, y_j\right) p_{X,Y}\left(x_i, y_j\right) \tag{3.50}$$

In all the cases, the result is the average value of the function $g(X, Y)$ weighted by the probability that $X = x$ and $Y = y$ or the mean of the random variable U.

3.5.2 *Moment about the Origin*

A general expression for the (r, s)th moment of the jointly distributed continuous random variable X and Y is

$$\mu^1_{r,s} = \int\int x^r y^s f_{X,Y}\left(x, y\right)\, dx\, dy \qquad \text{for continuous RV} \tag{3.51}$$

$$\mu^1_{r,s} = \sum_i \sum_j x_j^r\, y_i^s\, p_{X,Y}\left(x_i, y_j\right) \qquad \text{for discrete RV} \tag{3.52}$$

3.5.3 *Moment about the Mean (Central Moment)*

The central moment for jointly distributed continuous random variables X and Y is given by

$$\mu_{r,s} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_X)^r \, (y - \mu_Y)^s \, f_{X,Y}(x, y)\, dx\, dy \qquad \text{for continuous RV} \tag{3.53}$$

$$\mu_{r,s} = \sum_i \sum_j (x_i - \mu_X)^r \left(y_j - \mu_Y\right)^s p_{X,Y}(x, y) \qquad \text{for discrete RV} \tag{3.54}$$

3.5.4 Moment-Generating Function

Similar to moment-generating function of a single random variable defined in previous section, the moment-generating function for two random variables is defined for discrete and continuous cases. The moment-generating function for two continuous random variables can be obtained as

$$M_{X,Y}(t, u) = E\left(e^{tX+uY}\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{tx+uy} f_{X,Y}(x, y) \tag{3.55}$$

The moment-generating function for two discrete random variables can be obtained as

$$M_{X,Y}(t, u) = E\left(e^{tX+uY}\right) = \sum_{\text{all } x}\sum_{\text{all } y} e^{tx+uy} p_{X,Y}(x, y) \tag{3.56}$$

Example 3.5.1

A reservoir has two inflow points A and B. The streamflow gauging records at station A and B show that inflow at station A (designated by X) and the same at station B (designated by Y) follow the given distributions.

$$f_X(x) = \begin{cases} \frac{1}{50}(10 - x) & 0 \le x \le 10 \\ 0 & \text{elsewhere} \end{cases}$$

$$f_Y(y) = \begin{cases} \frac{1}{300}(25 - y) & 0 \le y \le 20 \\ 0 & \text{elsewhere} \end{cases}$$

Considering the inflow at station A and B to be independent, evaluate the mean of total inflow to the reservoir and the moment-generating function for the same.

Solution As given, X designates the inflow at station A and Y designates the inflow at station B. The total inflow to the reservoir can be designated by another random variable, say Z. Thus, Z is a function of random variables X and Y such that $Z = g(X, Y) = X + Y$.

As the inflows at station A and B are independent, their joint *pdf* can be evaluated as the product of their individual, i.e., $f_{X,Y}(x, y) = f_X(x) f_Y(y)$. The mean of the total inflow to the reservoir can be evaluated as

$$
\begin{aligned}
E(Z) &= \int_0^{20} \int_0^{10} (x + y) \left(\frac{10 - x}{50}\right) \left(\frac{25 - y}{300}\right) dx\, dy \\
&= \int_0^{20} -\frac{(3y + 10)(y - 25)}{900} dy \\
&= \frac{100}{9} \qquad\qquad\qquad\qquad = 11.11
\end{aligned}
$$

The moment-generating function can be written as

$$
\begin{aligned}
M_Z(t, u) &= E(e^{tX+uY}) \\
&= \int_0^{20} \int_0^{10} e^{tx+uy} \left(\frac{10 - x}{50}\right) \left(\frac{25 - y}{300}\right) dx\, dy \\
&= \frac{(10t - e^{10t} + 1)(25u - e^{20u} - 5ue^{20u} + 1)}{15000t^2u^2}
\end{aligned}
$$

3.5.5 *Covariance*

The covariance of jointly distributed random variables X and Y can be written as the expected value of the product of their deviations from their respective mean values as follows:

$$
\operatorname{Cov}(X, Y) = \sigma_{X,Y} = E\left[(X - \mu_X)(Y - \mu_Y)\right] \tag{3.57}
$$

By using the linearity property of expectations, *r.h.s.* of Eq. 3.57 can be transformed to a simpler form, which describes as the expected value of their product minus the product of their expected values, as shown in Eq. 3.58.

$$
E\left[(X - \mu_X)(Y - \mu_Y)\right] = E(XY) - E(X)E(Y) \tag{3.58}
$$

For continuous random variables, covariance can be expressed as

$$
\sigma_{X,Y} = \iint (x - \mu_X)(y - \mu_Y) f_{X,Y}(x, y)\, dx\, dy \tag{3.59}
$$

For discrete random variables, covariance can be expressed as

$$\sigma_{X,Y} = \sum_{\text{all } x} \sum_{\text{all } y} (x - \mu_X)(y - \mu_Y)\, p_{X,Y}(x, y) \tag{3.60}$$

If X and Y are independent, then $f_{X,Y}(x, y) = f_X(x)\ f_Y(y)$ for continuous random variable and $p_{X,Y}(x, y) = p_X(x)\ p_Y(y)$ for discrete random variable.

Thus, covariance for independent continuous random variables can be expressed as

$$\begin{aligned} \sigma_{X,Y} &= \iint (x - \mu_X)(y - \mu_Y)\, f_{X,Y}(x, y)\, dx\, dy \\ &= \int (x - \mu_X)\, f_X(x)\, dx \int (y - \mu_Y)\, f_Y(y)\, dy = 0 \end{aligned} \tag{3.61}$$

Thus, covariance for independent discrete random variables can be expressed as

$$\begin{aligned} \sigma_{X,Y} &= \sum_{\text{all } x} \sum_{\text{all } y} (x - \mu_X)(y - \mu_Y)\, p_{X,Y}(x, y) \\ &= \sum_{all\ x} (x - \mu_X)\, p_X(x) \sum_{all\ y} (y - \mu_Y)\, p_Y(y) = 0 \end{aligned} \tag{3.62}$$

since first central moment with respect to mean is 0. This implies covariance of two independent variables is always 0. However, the reverse is not true, i.e., zero covariance does not necessarily indicate that the variables are independent.

The sample estimate for the covariance $\sigma_{X,Y}$ is $S_{X,Y}$ computed as

$$S_{X,Y} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{(n - 1)} \tag{3.63}$$

Example 3.5.2

The joint distribution of two random variables X_1 and X_2 is given as follows. Find out the covariance of X_1 and X_2.

$$f_{X_1,X_2}(x_1, x_2) = \begin{cases} 6x_1 & 0 < x_1 < x_2 < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Solution The marginal distributions of X_1 and X_2 are as follows:

$$f_X(x_1) = \int_{x_1}^{1} 6x_1 dx_2 = [6x_1 x_2]_{x_1}^{1} = 6x_1(1 - x_1) \qquad 0 < x_1 < 1$$

$$f_X(x_2) = \int_0^{x_2} 6x_1 dx_1 = \left[6\frac{x_1^2}{2}\right]_0^{x_2} = 3x_2^2 \qquad 0 < x_2 < 1$$

The covariance of X_1 and X_2 can be calculated as follows:

$$\text{Cov}\left(X_1, X_2\right) = E\left(X_1, X_2\right) - E\left(X_1\right) E\left(X_2\right)$$

Expectation for X_1 and X_2 can be calculated as follows:

$$E\left(X_1\right) = \int_0^1 x_1 6\left(x_1\right)\left(1 - x_1\right) dx_1 = \frac{1}{2}$$
$$E\left(X_2\right) = \int_0^1 x_2\left(3x_2^2\right) dx_2 = \frac{3}{4}$$

Expectation of joint distribution of X_1 and X_2 can be evaluated as

$$E\left(x_1 x_2\right) = \int_0^1 \int_0^{x_2} x_1 x_2 6 x_1 dx_1 dx_2 = \frac{2}{5}$$

Thereby, the covariance can be evaluated as

$$\text{Cov}\left(X_1, X_2\right) = \frac{2}{5} - \frac{1}{2} \times \frac{3}{4} = \frac{1}{40}$$

3.5.6 *Correlation Coefficient*

Correlation coefficient is a normalized form of covariance which is obtained by dividing the covariance by the product of standard deviation of X and Y.

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{3.64}$$

The range of $\rho_{X,Y}$ is $-1 \leq \rho_{X,Y} \leq 1$. Actually, $\rho_{X,Y}$ is the measure of linear dependence between X and Y. Thereby, if $\rho_{X,Y} = 0$, and X and Y are linearly independent, however, they might be related by some nonlinear functional form. In this case, X and Y are said to be uncorrelated. A value of $\rho_{X,Y}$ equal to ± 1 implies that X and Y are perfectly related by $Y = a + bX$. In this case, X and Y are said to be correlated. The sample estimate of the population correlation coefficient $\rho_{X,Y}$ is $r_{X,Y}$ computed from

$$r_{X,Y} = \frac{S_{X,Y}}{S_X S_Y} \tag{3.65}$$

where S_X and S_Y are the sample estimates of σ_X and σ_Y, respectively, and $S_{X,Y}$ is the sample covariance.

Example 3.5.3
Let X units denote the rainfall intensity in a particular catchment and Y units denote the runoff from the catchment. The joint *pdf* of X and Y is given as follows. Evaluate the covariance and the correlation coefficient.

$$f_{X,Y}(x,y) = \begin{cases} x^2 + \frac{xy}{3} & 0 \le x \le 1; 0 \le y \le 2 \\ 0 & \text{elsewhere} \end{cases}$$

Solution Evaluation of the marginal *pdf* of X and Y is carried out in Example 3.5.2. In order to evaluate the correlation coefficient, we have to evaluate the variance of X, variance of Y, and covariance of X and Y.

$$\text{Cov}(XY) = E(XY) - E(X)E(Y)$$

$$E(X) = \int_0^1 x\left(\frac{2}{3}x + 2x^2\right)dx = \left[\frac{2}{9}x^3 + \frac{1}{2}x^4\right]_0^1 = \frac{13}{18}$$

$$E\,(Y) = \int_0^2 y\left(\frac{1}{3} + \frac{y}{6}\right)dy = \left[\frac{1}{6}y^2 + \frac{1}{18}y^3\right]_0^2 = \frac{10}{9}$$

$$\begin{aligned} E(X,Y) &= \int_0^1\int_0^2 xy\left(x^2 + \frac{xy}{3}\right)dy\,dx \\ &= \int_0^1 \left[\frac{1}{2}x^3y^2 + \frac{1}{9}x^2y^3\right]_0^2 dx = \int_0^1 2x^3 + \frac{8}{9}x^2\,dx = \left[\frac{1}{2}x^4 + \frac{8}{27}x^3\right]_0^1 = \frac{43}{54} \end{aligned}$$

$$\text{Cov}(XY) = \frac{43}{54} - \left(\frac{13}{18}\right)\left(\frac{10}{9}\right) = -\frac{1}{162}$$

As $\text{Cov}(X,Y) \neq 0$, thereby, X and Y are correlated.

Calculation of variance of X and Y,

$$\text{Var}(X) = E(X^2) - [E\,(X)]^2$$

$$\text{Var}(Y) = E(Y^2) - [E\,(Y)]^2$$

$$E(X^2) = \int_0^1 x^2\left(\frac{2}{3}x + 2x^2\right)dx = \left[\frac{1}{6}x^4 + \frac{2}{5}x^5\right]_0^1 = \frac{17}{30}$$

$$E\,(Y^2) = \int_0^2 y^2\left(\frac{1}{3} + \frac{y}{6}\right)dy = \left[\frac{1}{9}y^3 + \frac{1}{24}y^4\right]_0^2 = \frac{14}{9}$$

$$\mathrm{Var}\,(X) = \frac{17}{30} - \left(\frac{13}{18}\right)^2 = 0.045$$

$$\mathrm{Var}\,(Y) = \frac{14}{9} - \left(\frac{10}{9}\right)^2 = 0.321$$

Calculation of correlation coefficient,

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \frac{-1/162}{\sqrt{0.045}\sqrt{0.321}} = -0.051$$

The correlation coefficient is −0.051.

3.5.7 *Further Properties of Moments*

If Z is a linear function of two random variables X and Y such that $Z = aX + bY$, then

$$E\,(Z) = E\,(aX + bY) = aE\,(X) + bE\,(Y) \tag{3.66}$$

$$\mathrm{Var}\,(Z) = a^2 Var\,(X) + b^2 Var\,(Y) + 2ab\,Cov\,(X, Y) \tag{3.67}$$

We can generalize the above equations considering Y as a linear function of n random variables such that $Y = \sum_{i=1}^{n} a_i X_i$, then,

$$E\,(Y) = E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E\,(X_i) \tag{3.68}$$

$$\mathrm{Var}\,(Y) = \sum_{i=1}^{n} a_i^2\; Var\,(x_i) + 2\sum_{i<j} a_i a_j\; Cov\,\left(X_i, X_j\right) \tag{3.69}$$

Now for a special case considering $a_i = 1/n$ in Y, we get $Y = \overline{X}$. Since x_i form a random sample, the $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$ and $\mathrm{Var}(X_i) = \mathrm{Var}(X)$. Thereby,

$$\mathrm{Var}\,(Y) = \mathrm{Var}\left(\overline{X}\right) = \sum_{i=1}^{n} \frac{1}{n^2}\mathrm{Var}\,(X) = \frac{n}{n^2} Var\,(X)$$

or,

$$\mathrm{Var}\left(\overline{X}\right) = \frac{\mathrm{Var}\,(X)}{n} \tag{3.70}$$

If we consider X and Y to be independent random variables, then the variance of their product XY is given by:

$$\text{Var}\,(XY) = E\,(XY)^2 - E^2\,(XY) \tag{3.71}$$

Now, $E\,(XY)^2 = E\left(X^2\right) E\left(Y^2\right) = \left(\mu_X^2 + \sigma_X^2\right)\left(\mu_Y^2 + \sigma_Y^2\right)$.
And $E^2\,(XY) = E^2\,(X)\;E^2\,(Y) = \mu_X^2\,\mu_Y^2$.
Thus, variance of the product X and Y can also be expressed as

$$\text{Var}\,(XY) = \mu_X^2\sigma_Y^2 + \mu_Y^2\sigma_X^2 + \sigma_X^2\sigma_Y^2 \tag{3.72}$$

3.6 Properties of the Estimator

In general, the probability distribution functions are the functions of a set of parameters and the random variable. To use the probability distribution for the estimation of probability, it is important to calculate the values of the parameters. The general procedure for estimating a parameter is to obtain a random sample from the population and use it to estimate the parameters. Now if we consider $\hat{\theta}_i$ as the estimate for the parameter θ_i, then $\hat{\theta}_i$ is a function of the random variables since $\hat{\theta}_i$ is itself a random variable possessing mean, variance and probability distribution. An ideal estimator should possess the following four characteristics, namely unbiasedness, consistency, efficiency, and sufficiency.

3.6.1 Unbiasedness

An estimator $(\hat{\theta})$ of a parameter (θ) is said to be unbiased if the expected value of the estimate is equal to the parameter $\left(E(\hat{\theta}) = \theta\right)$. As unbiased, estimator implies that an average of many independent estimators for the parameter will be equal to the parameter itself. In case the estimate is biased, the bias can be evaluated as $E(\hat{\theta}) - \theta$.

3.6.2 Consistency

An estimator $(\hat{\theta})$ of a parameter (θ) is said to be consistent if the probability that the estimator differs from the parameter $(\hat{\theta} - \theta)$ by more than a constant (ε) approaches to 0 as the sample size approaches infinity.

3.6.3 Efficiency

An estimator $(\hat{\theta})$ is said to be more efficient estimator for a parameter (θ) if the estimator is unbiased and its variance is at least as small as that of another unbiased estimator $\hat{\theta}_1$. The relative efficiency (RE) of $\hat{\theta}$ with respect to another estimator $\hat{\theta}_1$ can be evaluated as follows:

$$RE = \frac{V\left(\hat{\theta}\right)}{V\left(\hat{\theta}_1\right)} \tag{3.73}$$

If the value of the relative frequency is less than 1, then $\hat{\theta}$ is a more efficient estimator of θ than $\hat{\theta}_1$.

3.6.4 Sufficiency

An estimator $(\hat{\theta})$ is said to be a sufficient estimator for a parameter (θ) if the estimator utilizes all of the information contained in the sample and is relevant to the parameter.

Example 3.6.1
Consider a random variable X such that $X \sim N\left(\mu, \sigma^2\right)$. Check if the estimators of mean $\overline{X} = \frac{1}{n}\sum_i X_i$ and variance $S^2 = \frac{1}{n-1}\sum_i \left(X_i - \overline{X}\right)^2$ are biased or unbiased.

Solution Estimator of mean (μ) is given as follows:

$$\overline{X} = \frac{1}{n}\sum_i X_i$$

Expectation of the estimator $E\left(\overline{X}\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mu_i = \mu$, which is equal to population mean. Therefore, $\overline{X}$ is an unbiased estimator of μ.

Estimator of variance $\left(\sigma^2\right)$ is given as follows:

$$S^2 = \frac{1}{n-1}\sum_i \left(X_i - \overline{X}\right)^2$$

Expectation of the estimator can be evaluated as

$$
\begin{aligned}
E\left(S^2\right) &= \frac{1}{n-1}\sum_i \left(X_i-\mu+\mu-\overline{X}\right)^2 \\
&= \frac{1}{n-1}\sum_i (X_i-\mu)^2+\left(\mu-\overline{X}\right)^2+2\left(X_i-\mu\right)\left(\mu-\overline{X}\right) \\
&= \frac{1}{n-1}\sum_i (X_i-\mu)^2+\left(\mu-\overline{X}\right)^2+2n\left(\overline{X}-\mu\right)\left(\mu-\overline{X}\right) \\
&= \frac{1}{n-1}\sum_i (X_i-\mu)^2-n\left(\mu-\overline{X}\right)^2 \\
&= \frac{1}{n-1}\left(n\sigma^2-\sigma^2\right) \\
&= \sigma^2
\end{aligned}
$$

Therefore, S^2 is an unbiased estimator of σ^2.

3.7 Parameter Estimation

3.7.1 Method of Moments

The method of moments is a popular method of estimation of population parameters. It considers that a good estimate of a probability distribution parameter is that for which central moments of population equal with corresponding central moment of the sample data. Finally, an equation is derived that relates the population moments to the parameters of interest. For this purpose, a sample is drawn and the population moments are estimated from the sample. Then, the equations are solved for the parameters of interest, after replacing (unknown) population moments by sample moments. In case of a distribution with m parameters, the first m moments of the distribution are equated to the sample moments to obtain m equations which can be solved for the m unknown parameters. In other words, let us consider a random variable X that follows a distribution function $f_X\left(x;\theta_1,\ldots,\theta_k\right)$, with parameters $\theta_1,\ldots,\theta_k$ and a random sample $x_1,\ldots,x_n$, and then as per the assumptions of the method of moment, the rth population moment can be equated to the rth sample moment. Thus, we finally get the estimates of that parameter (see Example 3.7.1).

Example 3.7.1
Consider an exponential distribution whose *pdf* is given by $f_X(x)=\lambda e^{-\lambda x}$ for $x>0$. Determine the estimate of the parameter λ.

Solution Equating the first-order central moment of population to that of sample, we get

$$\mu = E(X) = \int_{-\infty}^{\infty} x f_x(x) dx = \int_{0}^{\infty} x \lambda e^{-\lambda x} dx$$

Using integration by parts (Example 3.2.2)

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

That yields, $\lambda = 1\big/\mu$, and thus the corresponding sample estimate is $\lambda = 1\big/\overline{x}$.

3.7.2 *Maximum Likelihood*

Maximum-likelihood (ML) method assumes that the best estimator of a parameter of a distribution should maximize the *likelihood* or the joint probability of occurrence of a sample. Let us consider, $x = (x_1, \ldots, x_n)$ is a set of n independent and identically distributed observed sample and $f\ (x,\ \theta)$ is the probability distribution function with parameter θ. The likelihood function can be written as follows:

$$L = \prod_{i=1}^{n} f_x(x_i) \tag{3.74}$$

where the symbol $\prod$ indicates multiplication. Sometimes, it becomes convenient to work with logarithmic of likelihood function, i.e,

$$\ln L = \sum_{i=1}^{n} \ln\left[f_x(x_i)\right] \tag{3.75}$$

In this case, $\hat{\theta}$ is said to be the maximum-likelihood estimator (MLE) of θ if $\hat{\theta}$ maximizes the function L or $\ln(L)$.

Example 3.7.2 Consider $x_1, \ldots, x_n$ to follow the following distribution

$$f_X\ (x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad -\infty < x < \infty$$

Evaluate MLE for μ and σ^2.

Solution The likelihood function is to be evaluated as follows:

$$\begin{aligned} L &= L\left(\mu, \sigma^2 \left| x_1, \ldots, x_n \right.\right) \\ &= \prod_{i=1}^{n} f_{X_i}\left(x_i\right) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \\ &= \frac{1}{\left(\sqrt{2\pi}\right)^n \left(\sigma^2\right)^{n/2}} e^{-1/2\sigma^2 \sum_{i=1}^{n}(x_i-\mu)^2} \end{aligned}$$

Thereby, the log-likelihood function can be evaluated as follows:

$$\log\, L = -\frac{n}{2}\log\, 2\pi - \frac{n}{2}\log\, \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

The estimator of μ can be evaluated by maximizing the log-likelihood function

$$\begin{aligned} &\frac{\partial \log L}{\partial \mu} = 0 \\ &\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \qquad\qquad \mu = \frac{\sum_{i=1}^{n} x_i}{n} = \hat{\mu} \end{aligned}$$

Therefore, the estimator of the mean is $\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$.

$$\begin{aligned} &\frac{\partial \log L}{\delta \sigma^2} = 0 \\ &\left(-\frac{n}{2\sigma^2}\right) + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \\ &\frac{1}{2\sigma^2}\left(-n + \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i-\hat{\mu}\right)^2\right) = 0 \\ &\sigma^2 = \frac{\sum_{i=1}^{n}\left(x_i-\hat{\mu}\right)^2}{n} = \hat{\sigma}^2 \end{aligned}$$

Therefore, the estimator of the variance is $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}$.

3.8 Chebyshev Inequality

Certain general statements about random variables can be made without fitting a specific distribution to the random variable. One such statement can be provided by the Chebyshev inequality. It ensures that not more than a certain fraction of values can be away from the mean by certain distance. The Chebyshev inequality states that the probability of getting a value which is away from μ by atleast $k\sigma$ is at most $1\big/k^2$, where μ is the population mean and σ is the population standard deviation. Thus,

$$P\left(|X-\mu| \geq k\sigma\right) \leq \frac{1}{k^2} \tag{3.76}$$

The Chebyshev inequality provides an upper limit for the probability of a deviation of a specific value from the mean.

3.9 Law of Large Number

Chebyshev's inequality can be written in terms of sample mean (sample size n) as follows:

$$P\left(|\overline{x}-\mu| \geq \frac{k\sigma}{\sqrt{n}}\right) \leq \frac{1}{k^2} \tag{3.77}$$

where $\overline{x}$ is the sample mean for a sample of size n. For the above inequality, if $1\big/k^2 = \delta$ and the sample size is selected such that $n \geq \sigma^2\big/\varepsilon^2$ where $\varepsilon < 0$ and $0 < \delta < 1$, then we get the law of large number. It can be stated as

$$P\left(\left|\overline{X}-\mu\right| \geq \varepsilon\right) \leq \delta \tag{3.78}$$

The law of large number ensures that by selecting a large enough sample, we can estimate the population mean with the desired accuracy.

3.10 MATLAB Examples

For solving examples in this chapter, symbolic toolbox of MATLAB is required. Some of the important function/commands are listed below.

- `syms`: This command is used for defining new algebraic symbol. For example, `syms x` will define an algebraic symbol `x`.
- `[output1,...,outputN] = eval(expr)`: This function evaluates the expression (`expr` argument). In case of symbolic expressions, this function can be used for simplifying them.

- `[y1,...,yN] = solve(eqns,vars)`: This function is used for solving univariate or multivariate equations (`eqns` argument) for variables `vars`. The variables argument is optional. In case of multiple equations, they are passed as string separated by comma like `[x_value,y_value] = solve(‘x+y = 7,x-y = 3’)` yields `x_value = 5` and `y_value = 2`.
- `output_expr = int(expr,var)`: This function is used for indefinite integration of expression (`expr` argument) with respect to variable (`var` argument). Further, `int(expr,var,a,b)` is used for definite integration of expression (`expr` argument) between the variable (`var` argument) value `a` and `b`.
- `output_expr = diff(expr,var)`: This function is used for symbolic differentiation of expression (`expr` argument) with respect to variable (`var` argument).

Using the functions discussed above, sample MATLAB script for solving Example 3.5.3 is provided in Box 3.1.

Box 3.1 Sample MATLAB script for solving Example 3.5.3

```
clear all;clc

%% Inputs, i.e, definition of all the distribution functions.
syms x y
x_fun=(2/3)*x+2*(x^2);
y_fun=(1/3)+(y/6);
joint_fun=(x^2)+(x*y)/3;

%% Evaluation of expectation of x, y and the joint distribution %of x and y.
exp_x=int(x*x_fun,x,0,1); % Expectation of x within the  %defined support
exp_y=int(y*y_fun,y,0,2); % Expectation of y within the %defined support
exp_joint=int(int(x*y*joint_fun,y,0,2),x,0,1);
cov_xy=exp_joint-(exp_y*exp_x); % Covariance of a and y

%% Evaluation of the correlation coefficient
exp_x2=int((x^2)*x_fun,x,0,1);
exp_y2=int((y^2)*y_fun,y,0,2);
var_x=exp_x2-(exp_x^2); %Evaluation of variance of x
var_y=exp_y2-(exp_y^2); %Evaluation of variance of y
cc_xy=eval(cov_xy/(sqrt(var_x)*sqrt(var_y))); % Evaluation of %the correlation coefficient

%% Display Results
output_file=['output' filesep() 'code_1_result.txt'];
delete(output_file);diary(output_file);diary on;
% Output stating if the variables are correlated
```

```
if cov_xy==0
    disp('The random variables X and Y are not
        correlated.');
else
    disp('The random variables X and Y are
        correlated.');
end
fprintf('The correlation coefficient of X and Y is
    %2.3f.\n',cc_xy)
diary off;
```

The output of the code mentioned in Box 3.1 is provided in Box 3.2. The solution obtained using the MATLAB code is same as the conclusions drawn from the solution of Example 3.5.3.

Box 3.2 Results for Box 3.1

```
The random variables X and Y are correlated.
The correlation coefficient of X and Y is -0.051.
```

Exercise

3.1 Considering the number of storms in an area for the month of June to follow the following distribution

$$p_X(x) = \begin{cases} \frac{2^x e^{-2}}{x!} & x = 1, 2, \ldots, 5 \\ 0.152 & x = 0 \end{cases}$$

Evaluate the mean and median for the number of storms in the given month. (Ans: 0.848; median lies between 1 and 2)

3.2 Soil samples are collected from 15 vegetated locations in a particular area. The moisture content of the samples as obtained from the laboratory tests is shown in the following table. Evaluate the arithmetic mean, geometric mean, range, variance, coefficient of skewness, and coefficient of kurtosis of the soil moisture data. Comment regarding the skewness and kurtosis of the data. (Ans: 0.3207; 0.2926; 0.490; 0.018; 0.4136; 3.496)

Sample no	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
SMC	0.25	0.40	0.11	0.45	0.36	0.24	0.26	0.31	0.50	0.60	0.39	0.28	0.19	0.14	0.33

3.3 The maximum temperature (in °C) at a city in the month of May follows the distribution as given below

$$f_X(x) = \frac{1}{\beta - \alpha} \qquad 40 \le x \le 45$$

Evaluate the mean, variance, and coefficient of variation of the maximum temperature in the city. (Ans: 42.5 °C; 2.083; 0.034)

3.4 The discharge at a gauging station follows the given distribution

$$f_X(x) = 5e^{-5x} \qquad x \ge 0$$

Determine the nature of the distribution in terms of its coefficient of variation, skewness, and tailedness. (Ans: 1/5; 2; 6)

3.5 A city supplied water from two sources. The joint *pdf* of discharge from two sources is as follows:

$$f_{X,Y}(x, y) = \begin{cases} x^2 + \frac{xy}{3} & 0 \le x \le 1; 0 \le y \le 2 \\ 0 & \text{elsewhere} \end{cases}$$

Evaluate the marginal probability density of each source and the mean discharge from the two sources. Also, evaluate the covariance and the forms of conditional distribution of X given $Y = y$. (Ans: 13/18 units; 30/27 units; $-1/162$)

3.6 Consider a random variable X to follow a two-parameter distribution. The population mean (μ) and standard deviation (σ) are the parameters of the distribution. Evaluate an unbiased estimation of μ and unbiased and biased estimation of σ.

3.7 Let $x_1, x_2, \ldots, x_n$ be a random sample for a distribution with *pdf*

$$f_X(x) = \frac{e^{-x/\beta} \times x^{\alpha-1}}{\beta\alpha\,\Gamma(\alpha)} \qquad \alpha > 1; x, \beta > 0$$

Find estimators for α and β using method of moments.

3.8 Let $x_1, x_2, \ldots, x_n \sim U(0, \theta)$. Find the maximum-likelihood estimate of θ?

3.9 If $x_1, x_2, \ldots, x_n \sim \frac{e^{-\lambda}\lambda^x}{x!}$. Find the maximum-likelihood estimate of λ?

3.10 Considering the peak annual discharge at a location to have a mean of 1100 cumec and standard deviation of 260 cumec. Without making any distributional assumptions regarding the data, what is the probability that the peak discharge in any year will deviate more than 800 cumec from the mean? (Ans: 0.106)

3.11 The random variable X can assume the values 1 and -1 with probability 0.5 each. Find (a) the moment-generating function and (b) the first four moments about the origin. (Ans: (a) $E(e^{tX}) = \frac{1}{2}(e^{t} + e^{-t})$, (b) 0, 1, 0, 1)

3.12 A random variable X has density function given by

$$f_X(x) = \begin{cases} 2e^{-2x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Find (a) the moment-generating function and (b) the first four moments about the origin. (Ans: (b) 1/2, 1/2, 3/4, 3/2)

3.13 Find the first four moments (a) about the origin and (b) about the mean, for a random variable X having density function

$$f_X(x) = \begin{cases} 4x(9 - x^2)/81 & 0 \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

(Ans: (a) 8/5, 3, 216/35, 27/2 (b) 0, 11/25, 32/875, 3693/8750)

3.14 Find (a) $E(X)$, (b) $E(Y)$, (c) $E(X, Y)$, (d) $E(X^2)$, (e) $E(Y^2)$, (f) $\mathrm{Var}(X)$, (g) $\mathrm{Var}(Y)$, (h) $\mathrm{Cov}(X, Y)$ if the joint *pdf* of random variables X and Y is given as

$$f_{X,Y}(x, y) = \begin{cases} c(2x + y) & 2 < x < 5; 0 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

Use $c = 1/210$. (Ans: (a) 268/63, (b) 170/63, (c) 80/7, (d) 1220/63, (e) 1175/126, (f) 5036/3969, (g) 16225/7938, (h) $-200/3969$)

3.15 Joint distribution between two random variables X and Y is given as follows:

$$f_{X,Y}(x, y) = \begin{cases} 8xy & 2 \leq x \leq 1; 0 \leq y \leq x \\ 0 & \text{otherwise} \end{cases}$$

Find the conditional expectation of (a) Y given X and (b) X given Y. (Ans: (a) $\frac{2x}{3}$ (b) $\frac{2(1+y+y^2)}{3(1+y)}$)

3.16 The density function of a continuous random variable X is

$$f_X(x) = \begin{cases} 4x(9 - x^2)/81 & 0 \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

Find the (a) mean, (b) median, and (c) mode. (Ans: (a) 1.6 (b) 1.62 (c) 1.73)

3.17 Find the coefficient of (a) skewness and (b) kurtosis of the standard normal distribution which is defined by

$$f_x(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad - \infty < x < \infty.$$

(Ans: (a) 0, (b) 3).

Chapter 4
Probability Distributions and Their Applications

Use of probability distributions in hydrology and hydroclimatology is inevitable. This is mostly due to the presence of uncertainty and lack of complete knowledge from the data. In this chapter, two types of probability distributions, namely discrete probability distribution and continuous probability distribution are discussed elaborately. Commonly used distributions with their parameters, properties of the distribution supported by graphical representation, and their plausible applications in hydrology and hydroclimatology are explained. Each distribution is explained in the following order—basics, interpretation of the random variable, parameters, probability mass/density function, description, potential applications, and illustrative examples. This order is expected to help the readers to understand the distribution and to develop the knowledge base for its further applications.

4.1 Discrete Probability Distributions

A discrete probability distribution describes the probability of occurrence of each value of a discrete random variable. A discrete random variable is a random variable that can only take finite or countably infinite specific values (see Sect. 2.5.1). For example, let X be a random variable representing the number of rainy days in a month at a location. It is obvious that X can take up values belonging to the set of nonnegative integers only. Thus, X follows a discrete probability distribution. In this book, discrete probability distribution functions are referred as probability mass function (*pmf*) and denoted as $p_X(x)$. In the following sections, we will explain some of the most commonly used discrete probability distributions in the field of hydrology and hydroclimatology. Mathematical details of all the distributions are summarized at the end of the chapter in Table 4.2 to facilitate the readers.

R. Maity, *Statistical Methods in Hydrology and Hydroclimatology*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-10-8779-0_4

4.1.1 Binomial Distribution

Basics: Binomial distribution is a discrete probability distribution of the number of occurrences of an event in a sequence of n independent trials of a random experiment with a probability of occurrence of the event be p in each trial. Thus, the interpretation could be as follows:

> *Random Variable*: The number of occurrences (X) in a sequence of n independent experiments.
> *Parameters*: n and p, where n is the number of trials and p is the probability of occurrences in each trial.
> *Probability mass function*: $p_X(x; n, p) = {}^nC_x p^x (1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n$

Descriptions: Any process that may occur with the probability p at discrete points in time or space or individual trials may be a *Bernoulli Process* if the following assumptions hold:

(i) There are two and only two possible outcomes in each trial. These outcomes may be termed as ‘occurrence’ and ‘non-occurrence.’[1]
(ii) The probability of occurrence is the same for each trial. This implies that the probability of non-occurrence is also same for each trial.
(iii) The outcome of any trial is independent of the history of any prior occurrence or non-occurrence.

Let us consider X to be a random variable that represents the number of occurrence in the sequence of n number of trials of a *Bernoulli Process*. X is supposed to follow the *Binomial Distribution.* Occurrence of flood in a year, number of rainy days in a month, failure of embankment in a year might be possible Bernoulli processes if the aforementioned assumptions are satisfied.

To obtain the probability concerning X, we proceed as follows: If p and $(1-p)$ are the probability of occurrence and non-occurrence for each trial, then the probability of getting x occurrences (i.e., $(n-x)$ non-occurrences) in any order is $p^x(1-p)^{(n-x)}$. This is by the virtue of the generalized multiplication rule for more than two independent events. Now, the number of different orders in which x number of occurrences can happen is ${}^nC_x$, i.e., the number of combinations of x objects selected from a set of n objects. Thus, the probability of x occurrences out of n trials can be expressed as:

$$p_X\,(x; n, p) = {}^nC_x p^x\,(1-p)^{n-x} \qquad x = 0,\ 1,\ 2,\ \ldots,\ n \tag{4.1}$$

This is the *pmf* of binomial distribution for X with parameters n and p. The *cumulative binomial distribution* is expressed as,

[1] Some mathematical textbook may refer these outcomes as ‘success’ and ‘failure.’ We prefer to use ‘occurrence’ and ‘non-occurrence’ since it may be embarrassing to refer some of the extreme events, such as ‘floods,’ ‘droughts,’ ‘overtopping an embankment’ as success.

$$F_X\left(x; n, p\right) = \sum_{i=0}^{x} p_X\left(i; n, p\right) \tag{4.2}$$

The mean, variance, and coefficient of skewness of the binomial distribution are as follows:

$$E\left(X\right) = np \tag{4.3}$$

$$Var\left(X\right) = npq \tag{4.4}$$

$$\gamma = \frac{(q - p)}{\sqrt{npq}} \tag{4.5}$$

where $q = 1 - p$. Hence, the distribution is symmetric for $p = q$, positively skewed for $p < q$, and negatively skewed for $p > q$. The probabilities p and q may also referred as exceedance and non-exceedance probabilities.

Applications:

Probability of exceedance: An extreme event, such as heavy rainfall, high river discharge or flood, is said to have occurred if $X \geq x_T$, where X is the random variable and x_T is a fixed level. The probability of occurrence of such an extreme event is known as probability of exceedance. Binomial distribution is often used to compute the probability of occurrences for such extreme events.

Design return period: The time between the occurrences of two events is known as recurrence interval or return period of that event. Theoretically, return period (T) is the inverse of the probability (p) that the magnitude of event (x_T) will be equalled or exceeded in any year $(T = 1\big/p)$. Concept of return period is discussed in Sect. 5.1 of Chap. 5. The design return period of an extreme event should be much greater than the design life of a hydraulic structure such as a dam or an embankment. Reasonably high design life assures that an exceedance may not occur within the life span of a structure. The fact however remains that no matter the value of design return period considered to design a hydraulic structure; there remains a chance for an exceedance to occur. Several statistical assessments regarding design return period can be done using binomial distribution.

Example 4.1.1

Find the average occurrence of a 10-year flood (return period of the flood is 10 years) in a 100 year period? What is the probability that exactly this number of 10-year flood will occur in a 100-year period?

Solution The probability of occurrence of 10-year flood in any year $= 1/10 = 0.1$.

Thus, the average number of occurrences in 100 years $= E(X) = np = 100 \times 0.1 = 10$

The probability of 10 occurrences of 10-year flood in 100 years can be evaluated using binomial distribution,

$$p_X(x; n, p) = {}^{n}C_x p^x (1-p)^{n-x}$$

$$p_X(10\,; 100,\ 0.1) = {}^{100}C_{10}(0.1)^{10}(1-0.1)^{100-10} = 0.1319$$

Example 4.1.2

A hydrologist has two possible proposals to consider for the construction of an embankment. The details of the two proposals are given as follows. Which proposal should be considered for an economic design?

Design Parameters/Information	*Proposal 1*	*Proposal 2*
Return period	5 years	10 years
Flood magnitude	1400 m^3/s	2200 m^3/s
Time period of occurrence of the event once, such that the facility can be repaired with the revenue earned without any loss	Once in 8 years	Once in 15 years

Solution Let X be the number of occurrences of flood.

Proposal 1: If flood of the given magnitude occurs once or does not occur at all, then there will be no loss. So X can take up values 0 and 1. Now as this is a Bernoulli process, we can use the binomial distribution where $p = 1/5 = 0.2$ and $n = 5$. The probability that there is no loss,

$$\begin{aligned} p_X\,(x=0;\ 8,\ 0.2) + p_X\,(x=1;\ 8,\ 0.2) &= {}^{8}C_0\,(0.2)^0\,(0.8)^8 + {}^{8}C_1\,(0.2)^1\,(0.8)^7 \\ &= 0.168 + 0.335 \\ &= 0.503 \end{aligned}$$

Therefore, the probability of loss,
$= (1 - 0.503) = 0.497.$

Proposal 2: Similarly, in this case $p = 1/10 = 0.1$ and $n = 10$. The probability that there is no loss,

$$\begin{aligned} p_X\,(x=0;\ 15,\ 0.1) + p_X\,(x=1;\ 15,\ 0.1) &= {}^{15}C_0\,(0.1)^0\,(0.9)^{15} + {}^{15}C_1\,(0.1)^1\,(0.9)^{14} \\ &= 0.206 + 0.343 \\ &= 0.549 \end{aligned}$$

Therefore, the probability of loss,
$= (1 - 0.549) = 0.451.$

For the 2nd proposals, probability of loss is lower than the 1st proposal. Therefore, the 2nd proposal is more economic.

Example 4.1.3

If the probability of a design flood not exceeding in 20 years is 0.8, what should be the return period of the design storm?

Solution Using binomial distribution,

$$\begin{aligned} p_x\left(0;\ 20,\ p\right) &= {}^{20}C_0 p^0 \left(1-p\right)^{20} \\ \text{or, } 0.8 &= \left(1-p\right)^{20} \\ \text{hence, } p &= 1-\left(0.8\right)^{1/20} = 0.0111 \\ T &= 1\big/p = 90\,\text{years} \end{aligned}$$

The return period of the design flood is 90 years.

4.1.2 *Negative Binomial Distribution*

Basics: Negative binomial distribution is another discrete probability distribution of the random variable that denotes the number of trials in a Bernoulli process before a specific number (denoted by j) of occurrences. Thus, the interpretation could be as follows:

> *Random variable*: The number of occurrence (X) in a sequence of independent and identically distributed Bernoulli trials before a specific number of non-occurrences occurs.
> *Parameters*: j and p, where j is the number of non-occurrences and p is the probability of occurrence in each independent trial.
> *Probability mass function*: $p_x\left(x;\, j,\, p\right) = {}^{x-1}C_{j-1} p^j \left(1-p\right)^{x-j} \qquad x = j,\ j+1,\ \ldots$

Description: The probability that the jth occurrence happens at the Xth (X is the random variable here) trial can be calculated by noting that there must be $(j-1)$ occurrences in the $x-1$ trials preceding the Xth trial. The probability of $(j-1)$ occurrences in $x-1$ trials can be computed from the binomial distribution (explained before) as $p_x\left(x;\, j-1,\, p\right) = {}^{x-1}C_{j-1} p^{j-1} \left(1-p\right)^{x-j}$, where p is the probability of occurrence in each trial as defined in binomial distribution.

Next, the probability of occurrence in Xth trial is p. As all the trials are independent, the joint probability distribution function is obtained by multiplying these probabilities (${}^{x-1}C_{j-1} p^{j-1} \left(1-p\right)^{x-j}$ and p). Thus, probability of $X = x$, i.e., *pmf* of the *negative binomial distribution* is given by,

$$p_x\left(x;\, j,\, p\right) = {}^{x-1}C_{j-1} p^j \left(1-p\right)^{x-j} \qquad x = j,\ j+1,\ \ldots \tag{4.6}$$

Thus, different functional forms will result for different values of j. The *CDF* is expressed as,

$$F_x\left(x;\, j,\, p\right) = \sum_{i=j}^{x} p_x\left(i;\, j,\, p\right) \tag{4.7}$$

The mean, variance, and coefficient of skewness of the negative binomial distribution are

$$E\left(X\right) = \frac{j}{p} \tag{4.8}$$

$$Var\left(X\right) = \frac{j\left(1-p\right)}{p^2} \tag{4.9}$$

$$\gamma = \frac{1+p}{\sqrt{pj}} \tag{4.10}$$

Applications: The number of occurrences of extreme events within the life span of a hydraulic structure can be determined using the negative binomial distribution.

Rare Events Probabilities: Number of rare events like thunderstorm and hail days over certain period may fit the negative binomial distribution.
Tropical cyclone frequency distributions: The occurrence of cyclones and hurricanes in a year is identified as a rare event. Negative binomial distribution may be used for the annual frequencies of these events.

Example 4.1.4
What is the probability that the 10th occurrence of a 10-year flood will be on the 100th year?

Solution Using negative binomial distribution,

$$p_x\left(100;\ 10,\ 0.1\right) = {}^{99}C_9\left(0.1\right)^{10}\left(0.9\right)^{90} = 0.013$$

The probability that the 10th occurrence of a 10-year flood will occur on the 100th year is 0.013.

Example 4.1.5
The probability of non-occurrence of a hurricane in the state of Orissa once in 20 years is 0.05. Determine the probability of 5th occurrence of the hurricane in the 50th year?

Solution Using the binomial distribution with $n = 20$ and $x = 0$, the probability of occurrence of hurricane in a year can be evaluated as follows,

$$p_x\left(0;\ 20,\ p\right) = {}^{20}C_0 p^0\left(1-p\right)^{20}$$
$$0.05 = \left(1-p\right)^{20}$$
$$p = 1-\left(0.05\right)^{1/20} = 0.139$$

Thus, the return period,

$$T = \frac{1}{p} = 7.19 \approx 7 \text{ years}$$

The probability of 5th occurrence in 50th year can be evaluated using negative binomial distribution,

$$p_x\,(50;\,5\,,\;0.139) = {}^{49}C_4 0.139^5\,(0.861)^{45} = 0.013$$

Thus, the probability of five occurrences of a hurricane with 7-year return period in a span of 50 years is 0.013.

4.1.3 Multinomial Distribution

Basics: Multinomial distribution is the generalized form of a binomial distribution by assuming each trial to have more than two (i.e., k) possible outcomes. The interpretation could be as follows:

> *Random Variable*: The number of occurrences $(X_1, \ldots, X_k)$ in a sequence of n independent experiments.
> *Parameters*: n and p_i $(i = 1,\ \ldots\ ,\ k)$, where n is the number of trials and p_i is the probability of occurrences of the ith outcome (X_i) in each experiment.
> *Probability mass function*: $p\,(x_1,\ x_2,\ \ldots,\ x_k) = \frac{n\,!}{x_1\,!x_2\,!\ldots.x_k\,!} p_1^{x_1} p_2^{x_2}\ \cdots\ p_k^{x_k}$
>
> for $x_i = 0, 1,\ \ldots,\ n$ and $\sum_{i=1}^{k} x_i = n$

Descriptions: Let us consider n independent trials, with each trial permitting k mutually exclusive outcomes whose respective probabilities are $p_1,\ \ldots,\ p_k$ such that $\sum_{i=1}^{k} p_i = 1$. Considering the outcomes of the first kind, second kind, and so on, we are interested in the probability $p\,(x_1,\ \ldots,\ x_k)$ of getting x_1 outcomes of the first kind, x_2 outcomes of the second kind, and so on. Using the arguments similar to the ones in Sect. 4.1.1, probability mass function can be developed. The *pmf* can also be expressed using gamma function as,

$$p\,(x_1,\ x_2, \ldots, x_k) = \frac{\Gamma\left(\sum_i x_i + 1\right)}{\Pi_i \Gamma\,(x_i + 1)} \prod_{i=1}^{k} p_i^{x_i} \qquad \text{for } x_i = 0, 1, \ldots, n \tag{4.11}$$

where $\Gamma\,(\bullet)$ is the gamma function (refer Sect. 4.2.5). The *CDF* is expressed as,

$$F_X\,(x_i) = \sum_{X_i < x_i} p\,(x_1,\ x_2,\ \ldots,\ x_k) \tag{4.12}$$

The mean and the variance of the multinomial distribution are,

$$E\left(X_i\right) = np_i \tag{4.13}$$

$$Var\left(X\right) = np_i\left(1 - p_i\right) \tag{4.14}$$

When the parameters are $n = 1$ and $k = 2$, the multinomial distribution becomes the Bernoulli distribution, i.e.,

$$p\left(x_1, x_2\right) = \frac{1}{x_1\,! \times x_2\,!} p_1^{x_1} \times p_2^{x_2} \tag{4.15}$$

When $n > 1$ and $k = 2$, the multinomial distribution becomes the binomial distribution (presented before).

When $n = 1$, and $k > 2$, it becomes the categorical distribution.

Applications: Categorical variables targeting multiple outcomes including rainfall, streamflow can be considered to follow multinomial distribution. For example, if the amount of rainfall at a particular region is divided into five categories based on the depth of rainfall such as very low, low, normal, high, and very high and we wish to evaluate the probability of a particular category of rainfall in that region, the multinomial distribution can be used.

Example 4.1.6

The probability of the annual maximum peak discharge less than 140 m^3/s is 0.4, and the probability for the same between 140 and 280 m^3/s is 0.3. What is the probability of occurrence of 4 peak flows less than 140 m^3/s and 2 peak flows between 140 and 280 m^3/s for a 10-year period?

Solution The first outcome (discharge less than 140 m^3/s) and the second outcomes (discharge between 140 and 280 m^3/s) are fixed as per the example. The third outcome can be considered as the peak discharge is greater than 280 m^3/s.

The probability of occurrence of the third outcome = $1 - 0.4 - 0.3 = 0.3$.

The event must occur = $10 - 4 - 2 = 4$ times.

Now, using multinomial distribution,

$$p\left(4,\ 2,\ 4;\ 10,\ 0.4,\ 0.3,\ 0.3\right) = \frac{10! \times \left((0.4)^4\,(0.3)^2\,(0.3)^4\right)}{(4!\ 2!\ 4!)} = 0.059$$

The probability of occurrence is therefore 0.059 which is basically very low.

Example 4.1.7

At a given location, years are considered to be below-normal if their respective annual total rainfall depths are lower than 500 mm, normal if it lies between 500 and 1100 mm, and above-normal if it lies above 1100 mm. Frequency analysis of annual rainfall record shows that the probability of normal, below-normal, and above-normal rainfall is 0.6, 0.3, and 0.1, respectively. Considering a randomly selected 20 years, determine the probability that 13 normal, 5 below-normal, and 2 above-normal rainfall years will occur.

Solution This example defines three outcomes, namely normal (r_1), below-normal (r_2), and above-normal (r_3) rainfall. The sought probability can be evaluated using multinomial distribution,

$$p\,(13,\ 5,\ 2;\ 20,\ 0.6,\ 0.3,\ 0.1) = \frac{20! \times \left((0.6)^{13}\,(0.3)^{5}\,(0.1)^{2}\right)}{(13!\ 5!\ 2!)} = 0.052$$

The probability of occurrences of 13 normal, 5 below-normal, and 2 above-normal rainfall years is therefore 0.052.

4.1.4 Hypergeometric Distribution

Basics: Hypergeometric distribution is a discrete probability distribution. The interpretation is as follows:

> *Random variable*: The number (X) of occurrences of an event in a sample of size n (drawn without replacement), from a population of size N containing k-specific possibilities of occurrences.
> *Parameters*: N, n, and k, where N is the size of the population, n is the size of the sample to be selected, and k is the number of specific events in the population, occurrence of which is calculated.
> *Probability mass function*: $p_X(x;\ N,\ n,\ k) = \frac{{}^{k}C_x \times {}^{N-k}C_{n-x}}{{}^{N}C_n}$
> for $x = \max\,(0, n + k - N)\,,\ \ldots,\ \min(n, k)$

Descriptions: Let us consider a sample of size n selected from a population of size N. The total possible outcome of the selection is ${}^{N}C_n$. The number of ways x occurrences may happen is ${}^{k}C_x$, k being the specific possibilities of occurrences. The number of ways $(n - x)$ non-occurrences may happen is ${}^{N-k}C_{n-x}$, where $(N - k)$ is the total number of possible non-occurrences. Thus, considering all the possibilities to be equally likely and for sampling without replacement, the probability of getting 'x occurrences in a sample size of n' is as follows:

$$p_X\,(x;\ N,\ n,\ k) = \frac{{}^{k}C_x \times {}^{N-k}C_{n-x}}{{}^{N}C_n} \qquad \text{for } x = \max\,(0, n + k - N)\,,\ \ldots,\ \min(n, k) \tag{4.16}$$

where x cannot exceed k and $(n - x)$ cannot exceed $(N - k)$.The *CDF* is expressed as,

$$F_X(x; N, n, k) = \sum_{x=\max(0,n+k-N)}^{x} p_X(i; n, p) \tag{4.17}$$

The mean, variance, and coefficient of skewness of the hypergeometric distribution are,

$$E(X) = \frac{nk}{N} \tag{4.18}$$

$$Var(X) = \frac{nk(N-k)(N-n)}{N^2(N-1)} \tag{4.19}$$

$$\gamma = \frac{(N-2k)(N-1)^{1/2}(N-2n)}{[nk(N-k)(N-n)]^{1/2}(N-2)} \tag{4.20}$$

Applications: Applications of hypergeometric distribution are general in nature. Generally, wherever total number of events/cases (N) with the number of total favorable cases (k) in it and a sample size of (n) are known, and it is required to calculate the probability of favorable cases in the sample, hypergeometric distribution is used. Sometimes, significance of relationship between climate indices and hydrologic variables is tested with hypergeometric distribution.

Example 4.1.8

Assume that during the month of July, 20 rainy days occurred. The occurrence of rain on a particular day is independent of occurrence of rain on any other day.

(a) What is the probability that 8 out of any 10 days are rainy days?
(b) What is the probability that less than 8 out of any 10 days are rainy days?

Solution

(a) The month of July has 31 days. So we are selecting 10 days out of 31 days. It is also given that the number of rainy days is 20 days. Using the hypergeometric distribution considering $N = 31$, $n = 10$ and $k = 20$.

$$p_X(8;\ 31,\ 10,\ 20) = \frac{{}^{20}C_8\,{}^{11}C_2}{{}^{31}C_{10}} = 0.156$$

Therefore, probability that 8 of these days are rainy is 0.156.

(b) Using the cumulative hypergeometric distribution considering $N = 31$, $n = 10$ and $k = 20$.

$$F_X(7;\ 31,\ 10,\ 20) = \frac{{}^{20}C_0\,{}^{11}C_{10} + {}^{20}C_1\,{}^{11}C_9 + \cdots + {}^{20}C_7\,{}^{11}C_3}{{}^{31}C_{10}} = 0.798$$

Therefore, probability that less than 8 of these days are rainy is 0.798.

Example 4.1.9

From a record of annual rainfall data for a particular station, 24 years are found to be above-normal. Among those 24 years, flood was observed for 6 years. Now if 10 above-normal annual rainfall data are chosen out of 24 years, what is the probability that 2 of the years will be flood years?

Solution For the given situation, hypergeometric distribution can be applied. Let us define a random variable X as number of observed flood years, which follows hypergeometric distribution with the following *pmf*

$$P\left(X=x\right)=\frac{{}^{k}C_{x}\,{}^{(N-k)}C_{(n-x)}}{{}^{N}C_{n}}$$

In the example, following data are given,

Total above-normal rainfall years $(N)=24$
Total no of flood years $(k)=6$
Above-normal rainfall years chosen as sample $(n)=10$
Number of observed flood years out of this sample $(x)=2$

Hence, probability of observing 2 flood years out of 10 above-normal annual rainfall years is

$$P\left(X=2\right)=\frac{{}^{6}C_{2}\,{}^{(24-6)}C_{(10-2)}}{{}^{24}C_{10}}=\frac{{}^{6}C_{2}\,{}^{18}C_{8}}{{}^{24}C_{10}}=\frac{15\times 43758}{1961256}=0.335$$

Example 4.1.10

Assume over a 100 years of record, 23 and 20 years were recorded as El Niño and La Niña years, respectively, out of which 20 and 13 years were found to have above-normal and below-normal rainfall at a region respectively. Overall, out of 100 years, 32 and 31 years were found to receive above-normal and below-normal rainfall respectively, at that region. Fifteen random above-normal and below-normal years are selected. To establish that El Niño and La Niña events are associated with above-normal and below-normal rainfall for that region respectively, what should be the number of selected El Niño and La Niña years in the sample? Assume 0.95 as the threshold probability to establish the fact.

Solution Let us define a random variable X as number of El Niño/La Niña years in the randomly selected 15 years. Thus, for the given situation, hypergeometric distribution can be applied, for which the *pmf* and *CDF* are as follows:

$$p\left(X=x\right)=\frac{{}^{k}C_{x}\,{}^{(N-k)}C_{(n-x)}}{{}^{N}C_{n}}$$

$$P\left(X\leq x\right)=\sum_{x}\frac{{}^{k}C_{x}\,{}^{(N-k)}C_{(n-x)}}{{}^{N}C_{n}}$$

For the first part of the example, following data are given,

Total above-normal rainfall years $(N) = 32$,

Total no of El Niño years $(k) = 20$,

Number of above-normal rainfall years chosen as sample $(n) = 15$,

Minimum number of observed El Niño years out of this sample to establish the fact that El Niño events are associated with above-normal rainfall $(x) = ?$

Now according to the example, the threshold probability to establish the fact that El Niño events are associated with above-normal rainfall is given as 0.95, hence

$$\begin{aligned} &P\,(X \le x) \ge 0.95 \\ \text{or, } &\sum_{x=1}^{x} \frac{{}^{20}C_x\,{}^{(32-20)}C_{(15-x)}}{{}^{32}C_{15}} \ge 0.95 \\ \text{or, } &\sum_{x=1}^{x} \frac{{}^{20}C_x\,{}^{12}C_{(15-x)}}{{}^{32}C_{15}} \ge 0.95 \end{aligned}$$

By solving the above equation by trial and error, $x = 12$.

For the second part of the example, following data are given,

Total below-normal rainfall years $(N) = 31$,

Total no of La Niña years $(k) = 13$,

Number of below-normal rainfall years chosen as sample $(n) = 15$,

Minimum number of observed La Niña years out of this sample to establish the fact that La Niña events are associated with above normal rainfall $(x) =?$

Now according to the example, the threshold probability to establish the fact that La Niña events are associated with above-normal rainfall is given as 0.95, hence

$$\begin{aligned} &P\,(X \le x) \ge 0.95 \\ \text{or, } &\sum_{x=1}^{x} \frac{{}^{13}C_x\,{}^{(31-13)}C_{(15-x)}}{{}^{31}C_{15}} \ge 0.95 \\ \text{or, } &\sum_{x=1}^{x} \frac{{}^{13}C_x\,{}^{18}C_{(15-x)}}{{}^{31}C_{15}} \ge 0.95 \end{aligned}$$

By solving the above equation by trial and error, $x = 9$.

4.1.5 Geometric Distribution

Basics: Geometric distribution is another discrete probability distribution of a random variable that defines the number of trials to get the first occurrence of a particular event in a Bernoulli process. Thus, the interpretation could be as follows:

Random variable: The number of trials (X) in the sequence of a Bernoulli process to get the first occurrence.
Parameters: p, where p is the probability of occurrence.
Probability mass function: $p_X\,(x : p) = p\,(1-p)^{x-1}$ for $x = 1,\ 2,\ \ldots,\ n$

Descriptions: The probability that the first success of a Bernoulli trial occurs on the xth trial can be found using the geometric distribution. In order to attain the first occurrence on the xth trial, there must be $(x-1)$ preceding trials whose outcome is non-occurrence. Since the successive outcomes in the Bernoulli process are independent, the desired probability distribution is given by:

$$p_X(x : p) = p(1-p)^{x-1} \qquad \text{for } x = 1,\ 2,\ \ldots,\ n \tag{4.21}$$

The *CDF* is expressed as,

$$F_X\,(x; p) = \sum_{i=1}^{x} p_X\,(i; p) \tag{4.22}$$

The mean, variance, and coefficient of skewness of the geometric distribution are as follows:

$$E\,(X) = \frac{1}{p} \tag{4.23}$$

$$Var\,(X) = \frac{(1-p)}{p^2} \tag{4.24}$$

$$\gamma = \frac{2-p}{\sqrt{1-p}} \tag{4.25}$$

Applications: Application of geometric distribution is also general. Wherever the calculation involves, consecutive non-occurrences and/or first occurrence of any hydrologic events, such as embank overtopping, cyclones, extreme rainfall, geometric distribution is used.

Example 4.1.11
A dam is constructed across a river to prevent the flooding in the downstream region. What is the probability that a 20-year flood will occur for the first time in the 10th year after the completion of the project? What is the probability that the same will not occur at least within 10 years?

Solution Using the geometric distribution, the probability that the first occurrence is in the tenth 10th year is,

$$p_x = (10, 0.05) = (0.05)\,(0.95)^9 = 0.031$$

This is explained as nine consecutive non-occurrences followed by 1 occurrence. These events are independent to each other, so the probability is obtained by multiplying these individual probabilities.

The probability that it will not occur at least within 10 years can also be interpreted as non-occurrence in the first 10 years.

$$(0.95)^{10} = 0.599$$

4.1.6 Poisson Distribution

Basics: Poisson distribution is a discrete probability distribution of a random variable that describes the probability of a particular number of events occurring within a fixed time interval. Thus, the interpretation could be as follows:

Random variable: The number of occurrences (X) of an event (outcomes of a Bernoulli Process) in a fixed interval of time.
Parameters: λ, also known as the shape parameter, indicates the average number of events per unit time interval or the expected number of occurrences of the event.
Probability mass function: $p_X\,(x;\,\lambda) = \lambda^x \frac{e^{-\lambda}}{x\,!}$ for $x = 0,\ 1,\ \ldots\ ;\ \lambda > 0$

Descriptions: Let us consider a Bernoulli process defined over an interval of time, and let p be the probability of occurrence of an event in a particular interval of time. If the time interval becomes shorter, the probability of occurrence of the event (p) in the interval also becomes smaller; on the other hand, the number of trials (n) increases. As a result, np (denoted by λ) remains constant, i.e., the expected number of occurrences in a time interval remains the same. In such case, the binomial distribution approaches to a Poisson distribution and is given by:

$$p_X\,(x;\,\lambda) = \lambda^x \frac{e^{-\lambda}}{x\,!} \qquad \text{for } x = 0,\ 1,\ \ldots\ ;\ \lambda > 0 \qquad (4.26)$$

The mean, variance, and coefficient of skewness of the Poisson distribution are as follows:

$$E(X) = \lambda \tag{4.27}$$

$$Var(X) = \lambda \tag{4.28}$$

$$\gamma = \lambda^{-1/2} \tag{4.29}$$

A process is defined as a *Poisson process* if the events occurring over time/area/space satisfy the three assumptions;

(i) The number of events occurring in disjoint time intervals is independent.
(ii) The probability of a single occurrence in a small time interval is proportional to the length of the interval.
(iii) Probability of more than one occurrences in a small interval is negligible.

Applications:

Thunderstorm and Hail days Probabilities: Number of occurrences of the rare events like thunderstorm and hail days during certain period may fit the Poisson distribution. Whether the occurrences of such events are changed over time can be checked through parameters of this distribution.
Tropical cyclone frequency distributions: The occurrence of cyclones and hurricanes in a year is identified as a rare event. Poisson distribution shows good statistical fit with the annual frequencies of these events.
Number of rainy days in a particular monsoon month: Number of rainy days in a particular month can be modeled using Poisson distribution. Sometimes, the characteristics of monsoon with respect to number of rainy days may change at a location over time due to climate change. Such investigation can be done through the distributional properties over two time periods using Poisson distribution.

In fact, any such similar application as mentioned above can be modeled using Poisson distribution.

Example 4.1.12
What is the probability that a flood with return period 10 years will occur once in 4 years?

Solution Probability of occurrence of a flood with return period $T = 10$ year is $1/10 = 0.1$

This example can be solved assuming two distributions—Binomial and Poisson distributions.

Using Binomial distribution

The probability of single occurrence ($x = 1$) of 10-year flood ($p = 0.1$) in 4 years ($n = 4$),

$$p_x(x = 1) = {}^nC_x p^x (1 - p)^{n-x} = {}^4C_1 (0.1)^1 (1 - 0.1)^3 = 0.292$$

Using Poisson distribution

Expected number of 10-year flood ($p = 0.1$) in 4 years ($n = 4$) is $\lambda = np = 4 \times 0.1 = 0.4$

The probability of single occurrence ($x = 1$) of 10-year flood,

$$p_X(x = 1) = \frac{\lambda e^{-\lambda}}{x!} = \frac{0.4e^{-0.4}}{1!} = 0.268$$

It can be noted that both the distributional assumptions provide approximately same answer.

Example 4.1.13

What is the probability of fewer than 2 occurrences of a 10-year storm in a 50-year period?

Solution Using the Poisson distribution, expected number of 10-year storm ($p = 0.1$) in 50 years ($n = 50$), $\lambda = np = 50 \times 0.1 = 5$

Thus, probability of fewer than 2 occurrences of 10-year storm in a 50 year

$$= \text{Prob}\,(x < 2) = \text{Prob}\,(x \leq 1) = \sum_{x=0}^{1} \frac{\lambda^x e^{-\lambda}}{x!} = \sum_{x=0}^{1} \frac{5^x e^{-5}}{x!} = \frac{5^0 e^{-5}}{0!} + \frac{5^1 e^{-5}}{1!} = 0.04$$

Therefore, the probability is 0.04.

4.2 Continuous Probability Distributions

If a random variable can take any possible real value from the range of real numbers, its probability distribution is called a continuous probability distribution (see Sect. 2.5.2). Let X be a random variable representing the annual streamflow at a particular station. It can take any possible value from 0 to ∞. Such random variables (X) will follow a continuous probability distribution. In this book, continuous probability distribution functions are referred as probability density function (*pdf*) and denoted as $f_X(x)$. In the following section, we will explain some of the most commonly used continuous probability distributions.

4.2.1 Uniform Distribution

Basics: Uniform distribution is the simplest and symmetric continuous probability distribution function. It is defined over a range (known as *support*) such that its

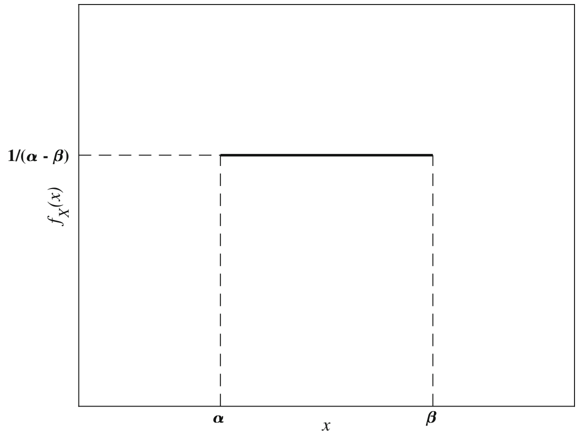


Fig. 4.1 Probability density function of uniform distribution with parameters α and β

occurrence is equally possible (equiprobable) over any subinterval of same length within the support. Thus, the interpretation could be as follows:

Random Variable: X that is equiprobable over any subinterval of same length within its support.
Parameters: α and β, where α and β are the minimum and maximum limit of the support respectively.
Probability density function: $f_x(x) = \frac{1}{\beta-\alpha} \qquad \alpha \le x \le \beta$

Descriptions: Let us consider a continuous random process restricted to a finite interval $[\alpha, \beta]$, and the probability of an outcome lying within a subinterval of $[\alpha, \beta]$ is proportional to the length of the subinterval. Such processes are said to be uniformly distributed over the interval α to β as shown in Fig. 4.1. The probability density function for the uniform distribution is as follows:

$$f_x(x) = \frac{1}{\beta - \alpha} \qquad \alpha \le x \le \beta \tag{4.30}$$

The cumulative density function for the continuous uniform distribution is as follows:

$$F_x(x) = \frac{x - \alpha}{\beta - \alpha} \qquad \alpha \le x \le \beta \tag{4.31}$$

The mean, variance, and coefficient of skewness of the uniform distribution are,

$$E(X) = \frac{(\beta + \alpha)}{2} \tag{4.32}$$

$$Var(X) = \frac{(\beta - \alpha)^2}{12} \tag{4.33}$$

$$\gamma = 0 \tag{4.34}$$

Applications:

General application: Many a times, random numbers are generated in hydrologic simulation. A random number is uniformly distributed over 0–1.

Statistical test: In statistical analysis, p-value is commonly utilized to assess the significance of a statistical test (refer to Chap. 6). The p-value is uniformly distributed between 0 and 1 if the null hypothesis is true and distribution of the test statistic is continuous.

Example 4.2.1

What is the probability of getting a number between 50 and 60 from a uniformly distributed series with support 0 to 100?

Solution Interval of probability distribution is 0–100. Thereby, density of probability is,

$$f_X(x) = \frac{1}{100 - 0} = \frac{1}{100}$$

Interval of probability distribution of success event is 50–60.

The probability ratio is thereby

$$P(50 \leq x \leq 60) = \frac{10}{100} = 0.1$$

Hence, probability of getting a number between 50 and 60 is 0.1.

Example 4.2.2

Number of hurricanes at a location per year is found to vary between 0 and 10 over last 50 years. If it is assumed to be uniformly distributed between these two limits, what is the probability of getting more than six hurricanes at that location in a particular year?

Solution Interval of probability distribution is 0–10. Therefore, the probability density function,

$$f_X(x) = \frac{1}{10 - 0} = \frac{1}{10}$$

And the cumulative probability distribution function is

$$F_X(x) = \frac{x - 0}{10 - 0} = \frac{x}{10}$$

Interval of probability distribution of success event is 7–10.

Thus, the probability of getting more than six hurricanes at the location in a particular year is

$$P\,(x > 6) = 1 - P\,(x \leq 7) = 1 - F_x\,(7) = 1 - \frac{7}{10} = \frac{3}{10} = 0.3$$

4.2.2 Exponential Distribution

Basics: Exponential distribution is a continuous probability distribution that may take any value between 0 and ∞, with higher probability of occurrence for lower values. It is an asymmetric distribution. The interpretation could be as follows:

Random Variable: The time (X) between two successive events, occurrences of which follow a Poisson process. It can also be spatial length (X) between two events if the events occur over space.
Parameters: λ, also known as the rate parameter, which is the average interarrival time (or space) between two successive events.
Probability density function: $f_x\,(x) = \lambda e^{-\lambda x}$ for $x > 0,\ \lambda > 0$

Descriptions: Let us assume that the interarrival times of an event are being noted. The event follows a Poisson process as discussed in Sect. 4.1.6

Since the probability that the event occurs during a certain time interval is proportional to the length of that time interval, it follows an exponential distribution. The continuous probability distribution of the interarrival time, i.e., the time between the occurrences of two successive events, can be evaluated by noting the $P\,(X \leq t)$ is equal to $1 - P\,(X > t)$. Thus, the CDF is

$$F_x(x) = 1 - e^{-\lambda x} \qquad \text{for } x > 0 \tag{4.35}$$

and the corresponding probability density function is given by:

$$f_x\,(x) = \frac{d}{dx} F_x\,(x) = \lambda e^{-\lambda x} \qquad \text{for } x \geq 0,\ \lambda > 0 \tag{4.36}$$

The mean, variance, and coefficient of skewness of the exponential distribution are as follows:

$$E\,(X) = \frac{1}{\lambda} \tag{4.37}$$

$$Var\,(X) = \frac{1}{\lambda^2} \tag{4.38}$$

$$\gamma = 2 \tag{4.39}$$

Applications:

Temporal: The interarrival time of hydrologic and other natural events like rainy day (>2.5 mm of rainfall in a day), earthquake, hurricane.

Categorical: Rainfall depth over different categories (0–10 mm, 10–20 mm, and so on).
Spatial: Many a times, variation of rainfall intensity from a rain gauge to any radial direction is considered to follow exponential distribution.

Example 4.2.3
Daily rainfall was recorded at a particular location for a period of 1 year. The data for rainy days are grouped into magnitude and number of days. The grouped data is presented in the following table. Plot a relative frequency histogram of the grouped data. Fit the exponential distribution to the histogram. Estimate the probability that a day selected in random will have rainfall greater than 45 mm.

Rainfall (mm)	Rainy Days	Rainfall (mm)	Rainy Days
0–10	90	50–60	5
10–20	49	60–70	3
20–30	34	70–80	2
30–40	17	80–90	1
40–50	13	90–100	1

Solution The relative frequency can be calculated by dividing the number of rainy days in each class with the total number of rainy days. These are the observed relative frequencies.

The best-fitted exponential curve can be fitted by the following method. The expected relative frequency in each class can be calculated as,

$$f_{x_i} = \Delta x_i \, p_X(x_i)$$

Here, $\Delta x_i = 10$ and x_i is the midpoint of each class interval.

Using exponential distribution,

$$p_X(x_i) = \lambda e^{-\lambda x}$$

with $\lambda = 1/\overline{x}$. The magnitude of $\overline{x}$ can be calculated using the expression to evaluate the mean for grouped data.

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i$$

where n in the total number of events and k is the total number of class intervals. Here, $n = 215$ and $k = 10$. Therefore, $\overline{x} = 18.674$ and corresponding $\lambda = 0.054$. The expression for $p_X(x_i)$ is

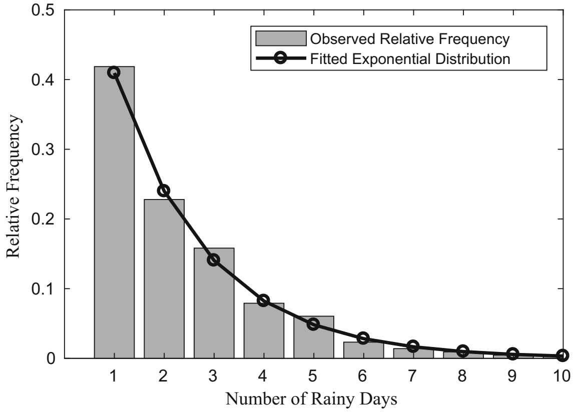


Fig. 4.2 Histogram plot of observed relative frequency of the data and the best-fitted exponential distribution

$$p_X(x_i) = 0.054 \times e^{(-0.0535 x_i)}$$

and the expression of expected relative frequency is

$$f_{x_i} = 10 \times 0.054 \times e^{(-0.0535 x_i)}$$

Histogram plot of observed relative frequency of the data and the best-fitted exponential distribution is shown in Fig. 4.2.

The estimated probability that a day will have rainfall greater than 45 mm is,

$$P_X(X > 45) = 1 - P_X(X \le 45) = 1 - \left(1 - e^{-0.054 \times 45}\right) = 0.088$$

4.2.3 *Normal Distribution*

Basics: Normal distribution, also known as Gaussian distribution or bell curve, is a continuous probability distribution. The interpretation could be as follows:

Random Variable: A continuous variable (X) that can take any value in the real line with a symmetrical (with respect to its mean) bell-shaped distribution of probability.
Parameters: μ and σ^2, where μ is the mean and σ^2 is the variance.
Probability density function: $f_X\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty$

Descriptions: Normal distribution is the most frequently used continuous probability distribution function. When mean is zero and variance is 1, the distribution is called as *standard normal distribution*. A *pdf* of standard normal distribution is shown in Fig. 4.3. It can be noticed that it is symmetrical with respect to mean and the typical

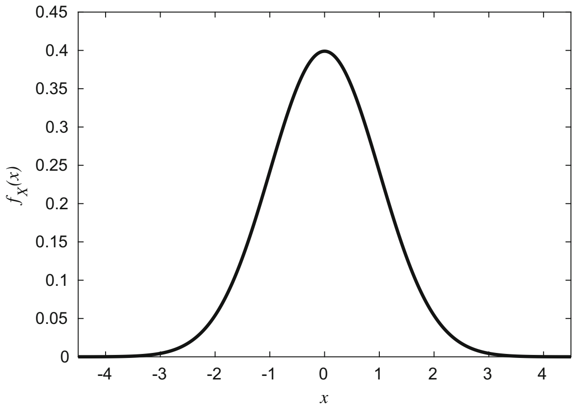


Fig. 4.3 Bell-shaped *pdf* of standard normal distribution

shape is known as a bell-shaped curve. The line of symmetry and the shape will change depending on the values of mean and variance, respectively.

The *pdf* of the *normal distribution* is given by:

$$f_X\left(x;\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}\qquad -\infty<x<\infty \tag{4.40}$$

The *CDF* of the *Normal Distribution* is given by:

$$F_X\left(x;\mu,\sigma^2\right)=\int\limits_{-\infty}^{x}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}dx\qquad -\infty<x<\infty \tag{4.41}$$

As stated before, the mean and the variance of the distribution are μ and σ^2 respectively, and the coefficient of skewness is 0, as it is a symmetric distribution.

Some properties:

(i) If X is normally distributed with mean μ and standard deviation σ, $Y = aX + b$ is another random variable that also follows normal distribution. However, the mean and standard deviation of Y are $a\mu + b$ and $a\sigma$.

(ii) **Central Limit Theorem**: The central limit theorem (*CLT*) specifies the conditions under which a random variable might be expected to follow a normal distribution. Under general conditions, it can be stated that if X_i $(i = 1, 2, \ldots, n)$ are n different independent and identically distributed (popularly known as *iid*) random variables with $\mathrm{E}(X_i) = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2$, then the sum of the random variables $S_n = X_1 + X_2 + \cdots + X_n$ approaches to a normal distribution as n approaches infinity. The mean and variance of S_n are as follows:

$$\mu_{S_n} = \sum_{i=1}^{n} \mu_i \tag{4.42a}$$

$$\sigma_{S_n} = \sum_{i=1}^{n} \sigma_i^2 \tag{4.42b}$$

> This generalized theorem is applicable irrespective of the parent distribution, i.e., distribution of X_i's. This is an attractive property to apply CLT in many fields of applications including hydrology and hydroclimatology.
>
> In practical cases, if X_i are independently and identically distributed, n does not have to be very large for S_n to approximately follow normal distribution. If interest lies in the central part of the distribution, even small number of values can result into normal distribution producing reasonable approximations to the true distribution of S_n. If interest lies in the tail of the distribution of S_n, a large number of values are required.

Evaluation of probability for Normal distribution:

Since the normal probability distribution cannot be integrated in closed form (Eq. 4.41) between two limits, say a to b, the probabilities related to normal distribution are generally computed numerically. These values are provided in Table B.1 in Appendix B at the end of the book (p. 434). This table pertains to *standard normal distribution*, i.e., the normal distribution with $\mu = 0$ and $\sigma = 1$, and its entries are the values of,

$$F_Z(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt \tag{4.43}$$

where $z = \frac{x-\mu}{\sigma}$. From the properties 1 (mentioned above), if X follows normal distribution with mean μ and standard deviation σ, Z follows standard normal distribution. As stated in Eq. 4.43, the table provides cumulative probability value from left extreme $(-\infty)$ to z.

To find the probability that a random variable having the standard normal distribution will take on a value between a and b, we use the equation $P(a < Z \leq b) = F(b) - F(a)$ as shown in the Fig. 4.4.

If a random variable X follows normal distribution with mean μ and variance σ^2, then a random variable Z given below follows normal distribution with mean 0 and variance 1, i.e., the standard normal distribution.

$$Z = \frac{X - \mu}{\sigma} \tag{4.44}$$

Thus, when X follows normal distribution with mean μ and variance σ^2, the probability of X being between p and q is given by,

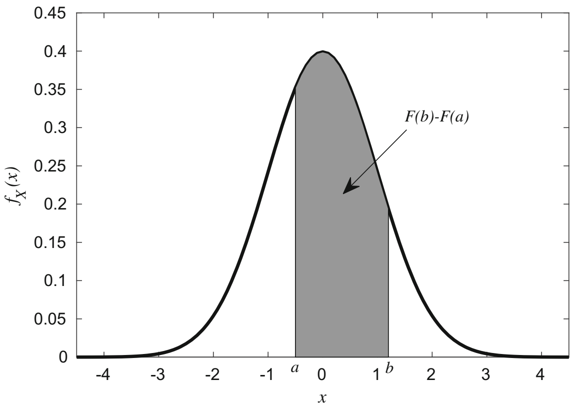


Fig. 4.4 The standard normal probability density function showing the shaded area as the probability $P(a < Z < b)$

$$\begin{aligned} P(p < X \leq q) &= P\left(\frac{p-\mu}{\sigma} < \frac{X-\mu}{\sigma} \leq \frac{q-\mu}{\sigma}\right) \\ &= P\left(\frac{p-\mu}{\sigma} < Z \leq \frac{q-\mu}{\sigma}\right) \\ &= F\left(\frac{q-\mu}{\sigma}\right) - F\left(\frac{p-\mu}{\sigma}\right) \end{aligned} \tag{4.45}$$

This probability is indicated by the shaded area in Fig. 4.4 if $\left(\frac{q-\mu}{\sigma}\right) = b$ and $\left(\frac{p-\mu}{\sigma}\right) = a$. An example is shown later on how to use the standard normal table to compute the values.

Applications: In hydrology and hydroclimatology, many variables may be found to follow normal distribution, e.g., temperature, relative humidity, wind velocity. The distribution of long duration river discharge or rainfall, e.g., monthly and yearly totals, is often found to follow normal distribution. Moreover, many statistical methods are developed under the assumption that the sample data follows normal distribution. However, many of the hydrologic variables may not follow normal distribution. Sometimes, some transformation techniques (discussed in Chap. 9, Sect. 9.6) are applied on these data so that transformed data follows normal distribution. Statistical inferences are made based on these transformed data.

Example 4.2.4
Considering mean daily temperature (X) at a location to follow normal distribution with mean 10 °C and standard deviation 5 °C,

(a) What is the probability of the mean daily temperature to be between 15 °C and 24 °C?
(b) What is the probability of the mean daily temperature is greater than 5 °C?

(c) What is the probability of the mean daily temperature is less than 20 °C?

Solution Using standard normal distribution, we can transform the limits of X to limits of Z and then use standard normal tables.

(a) $x = 15$ transforms to $z = \frac{x-\mu}{\sigma} = \frac{15-10}{5} = 1$
Similarly, $x = 24$ transforms to $z = \frac{24-10}{5} = 2.8$
From Table B.1, it can be seen that $F_z(1) = 0.841$ and $F_z(2.8) = 0.997$. Thus, the probability is,

$$P\left(15 \le X \le 24\right) = P\left(1 \le Z \le 2.8\right) = F_z\left(2.8\right) - F_z\left(1\right) = 0.997 - 0.841 = 0.156$$

(b) $x = 5$ transforms to $z = \frac{5-10}{5} = -1$
Thus, the desired probability is,

$$\begin{aligned} P\left(X > 5\right) &= 1 - P\left(X \le 5\right) \\ &= 1 - P\left(\frac{X-\mu}{\sigma} \le \frac{5-10}{5}\right) \\ &= 1 - P\left(Z \le -1\right) \\ &= 1 - F_z\left(-1\right) \\ &= 1 - 0.159 = 0.841 \end{aligned}$$

(c) $x = 20$ transforms to $z = \frac{20-10}{5} = 2$

$$\begin{aligned} P\left(X < 20\right) &= P\left(\frac{X-\mu}{\sigma} \le \frac{20-10}{5}\right) \\ &= P\left(Z \le 2\right) \\ &= F_z\left(2\right) \\ &= 0.977 \end{aligned}$$

4.2.4 Lognormal Distribution

Basics: Lognormal distribution is a continuous probability distribution of a random variable which is such that its logarithmic transformation follows a normal distribution. Thus, the interpretation could be as follows:

Random Variable: A random variable (X) that can take only positive values, asymmetric (positively skewed) and its logarithmic transformation ensures a normal distribution.

Parameters: α and β, where α is the mean and β is the variance of the logarithmic transformation of the random variable.

Probability density function: $f_X(x) = \begin{cases} \frac{1}{x\sqrt{2\pi\beta^2}} e^{-\frac{(\ln x-\alpha)^2}{2\beta^2}} & \text{for } x > 0,\ \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$

Descriptions: The product of many independent random variables each of which is positive may result in a lognormal distribution. This is justified by considering the central limit theorem (as discussed earlier) in the logarithmic domain. The probability distribution of lognormal distribution is as follows:

$$f_X(x) = \begin{cases} \frac{1}{x\sqrt{2\pi\beta^2}} e^{-\frac{(\ln x-\alpha)^2}{2\beta^2}} & \text{for } x > 0,\ \alpha, \beta > 0 \\ 0 & \text{elsewhere} \end{cases} \tag{4.46}$$

where $\ln x$ is the natural logarithm of x. The probability that a random variable having a lognormal distribution will lie between a and b $(0 < a < b)$ is given by,

$$P\left(a \le x \le b\right) = \int_a^b \frac{1}{x\sqrt{2\pi\beta^2}} e^{-\frac{(\ln x-\alpha)^2}{2\beta^2}} dx \tag{4.47}$$

Now, considering $y = \ln(x)$ and identifying the integrand as the normal density with $\mu = \alpha$ and $\sigma = \beta$, the desired probability is given by,

$$\begin{aligned} P\left(a \le X \le b\right) &= P\left(\ln a \le \ln X \le \ln b\right) \\ &= P\left(\ln a \le Y \le \ln b\right) \\ &= \int\limits_{\ln a}^{\ln b} \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(y-\alpha)^2}{2\beta^2}} dy \\ &= F\left(\frac{\ln(b)-\alpha}{\beta}\right) - F\left(\frac{\ln(a)-\alpha}{\beta}\right) \end{aligned} \tag{4.48}$$

where F is the cumulative distribution function of standard normal distribution. Typical *pdf* curves of the lognormal distribution with different combinations of α and β are shown in Fig. 4.5. It is very clear from the graph that the distribution is positively skewed.

The mean, variance, and coefficient of skewness of the lognormal distribution are as follows:

$$\mu = e^{(\alpha+\beta^2/2)} \tag{4.49}$$

$$\sigma^2 = \left(e^{\beta^2} - 1\right)\ e^{(2\alpha+\beta^2)} \tag{4.50}$$

$$\gamma = \left(e^{\beta^2} + 2\right)\sqrt{e^{\beta^2} - 1} \tag{4.51}$$

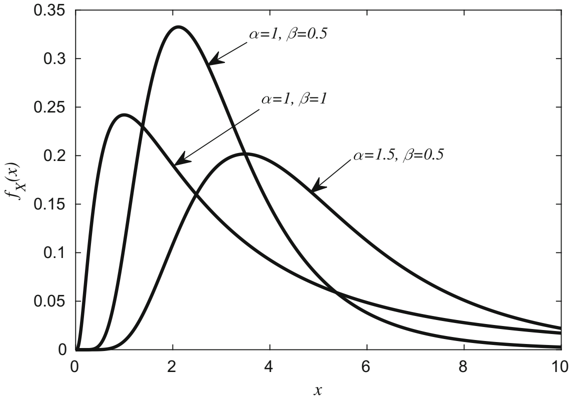


Fig. 4.5 Probability distribution functions of lognormal distribution for different combinations of α and β

Applications: The lognormal distribution is mostly applicable for hydrologic variables like monthly rainfall depth, river discharge volumes. The lognormal distribution is used to determine the extremes of variables at monthly and annual scales.

Example 4.2.5

Peak discharge at a particular river gauging station is found to have a mean of 130 m^3/s and standard deviation of 30 m^3/s. Considering the peak discharge to follow lognormal distribution evaluated the following,

(a) Probability of peak discharge being greater than 180 m^3/s.
(b) Probability of peak discharge lying in between 120 and 150 m^3/s.

Solution Given $\overline{x} = 130$ and $S_X = 30$.

As the peak discharge follows lognormal distribution, the parameters ($\overline{y}$ and S_Y) can be evaluated from the sample statistics ($\overline{x}$ and S_x) as follows,

$$C_v = \frac{S_X}{\overline{x}} = 0.231$$

$$\overline{y} = \frac{1}{2} \ln\left[\frac{\overline{x}^2}{C_v^2 + 1}\right] = 4.841$$

$$S_Y = \sqrt{\ln\left(C_v^2 + 1\right)} = 0.228$$

(a) For $x = 180$, the reduced variate is,

$$Z = \frac{y - \overline{y}}{S_Y} = \frac{\ln x - \overline{y}}{S_Y} = \frac{\ln 180 - 4.841}{0.228} = 1.544$$

The probability of peak discharge being greater than 180 m^3/s can be evaluated as follows: $P\,(Y > \ln\ 180) = 1 - P(Z < 1.544) = 1 - 0.939 = 0.061$

(b) For $x = 120$ and 150, the corresponding reduced variate are,

$$Z_1 = \frac{y - \overline{y}}{S_Y} = \frac{\ln\ x - \overline{y}}{S_Y} = \frac{\ln\ 120 - 4.841}{0.231} = -0.232$$

Similarly,

$$Z_2 = \frac{\ln\ 150 - 4.841}{0.231} = 0.734$$

The probability of peak discharge lying in between 120 and 150 m^3/s can be evaluated as follows,

$$P\,(\ln\ 120 < Y < \ln\ 150) = P\,(-0.232 < Z < 0.734) = 0.360$$

4.2.5 Gamma Distribution

Basics: Gamma distribution is a continuous probability distribution that is positively skewed over the positive side of the real line. The interpretation could be as follows:

Random Variable: A continuous, positively skewed random variable (X) that takes nonnegative values only.

Parameters: α and β are the shape and rate parameters respectively.

Probability density function: $f_X(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & \text{for } x \geq 0,\ \alpha > 0,\ \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$

Descriptions: The gamma distribution can be treated as the sum of exponentially distributed random variables each with the same parameter. The parameter α is the number of random variables following exponential distribution and β is the parameter of the exponential distributions. Gamma distribution has the probability density function as follows:

$$f_X(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & \text{for } x \geq 0,\ \alpha > 0,\ \beta > 0 \\ 0 & \text{elsewhere} \end{cases} \tag{4.52}$$

where $\gamma(\alpha)$ is the value of the gamma function defined by,

$$\gamma\,(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \tag{4.53}$$

integrating by parts results in,

Table 4.1 Values of gamma function, $\gamma\,(\alpha)$ for $\alpha \in [0, 1]$

(a) for $\alpha \in [0.1, 1]$

α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	∞	99.43	49.44	32.78	24.46	19.47	16.15	13.77	12.00	10.62
0.1	9.51	8.61	7.86	7.23	6.69	6.22	5.81	5.45	5.13	4.85
0.2	4.59	4.36	4.15	3.96	3.79	3.63	3.48	3.34	3.22	3.10
0.3	2.99	2.89	2.80	2.71	2.62	2.55	2.47	2.40	2.34	2.28
0.4	2.22	2.16	2.11	2.06	2.01	1.97	1.93	1.88	1.85	1.81
0.5	1.77	1.74	1.71	1.67	1.64	1.62	1.59	1.56	1.54	1.51
0.6	1.49	1.47	1.45	1.42	1.40	1.38	1.37	1.35	1.33	1.31
0.7	1.30	1.28	1.27	1.25	1.24	1.23	1.21	1.20	1.19	1.18
0.8	1.16	1.15	1.14	1.13	1.12	1.11	1.10	1.09	1.09	1.08
0.9	1.07	1.06	1.05	1.05	1.04	1.03	1.02	1.02	1.01	1.01

(b) for $\alpha \in [0, 0.1]$

α	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.0	∞	999.42	499.42	332.76	249.43	199.43	166.10	142.29	124.43	110.54
0.01	99.43	90.34	82.77	76.36	70.87	66.10	61.94	58.26	55.00	52.07
0.02	49.44	47.06	44.90	42.92	41.11	39.45	37.91	36.49	35.16	33.93
0.03	32.78	31.71	30.70	29.76	28.87	28.03	27.24	26.49	25.77	25.10
0.04	24.46	23.85	23.27	22.72	22.19	21.69	21.21	20.74	20.30	19.88
0.05	19.47	19.08	18.70	18.34	17.99	17.66	17.33	17.02	16.72	16.43
0.06	16.15	15.87	15.61	15.35	15.11	14.87	14.64	14.41	14.19	13.98
0.07	13.77	13.57	13.38	13.19	13.00	12.83	12.65	12.48	12.32	12.15
0.08	12.00	11.84	11.69	11.55	11.40	11.27	11.13	11.00	10.87	10.74
0.09	10.62	10.49	10.38	10.26	10.15	10.04	9.93	9.82	9.72	9.61

$$\gamma\,(\alpha) = (\alpha - 1)\,\gamma\,(\alpha - 1) \tag{4.54}$$

and $\gamma\,(\alpha) = (\alpha - 1)\,!$ when α is a positive integer excluding 1. The value of $\gamma\,(1) = 1$ and $\gamma\,(0.5) = \sqrt{\pi}$. For α between 0 and 1, values of $\gamma(\alpha)$ can be found from any standard gamma function table (Table 4.1).

Shape of gamma distribution is shown in Fig. 4.6 for different combinations of α and β. The graphs shown in the figure exhibit that gamma distributions are positively skewed and the skewness decreases as α increases for any fixed value of β. It may also be noticed that the exponential distribution is a special case of gamma distribution with $\alpha = 1$. The mean, variance, and coefficient of skewness of the gamma distribution are given as follows:

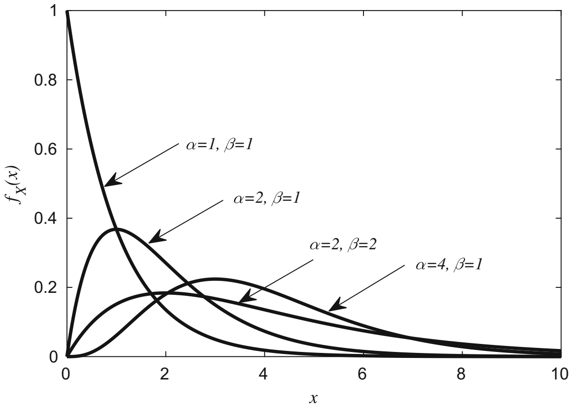


Fig. 4.6 Probability distribution functions of gamma distribution for different combinations of α and β

$$\mu = \alpha\beta \tag{4.55}$$

$$\sigma^2 = \alpha\beta^2 \tag{4.56}$$

$$\gamma = \frac{2}{\sqrt{\alpha}} \tag{4.57}$$

Applications: In hydrology, the gamma distribution has the advantage of having only positive values, since hydrological variables such as rainfall and runoff are always positive and lower bounded by zero.

Example 4.2.6

Engineers designed a hydroelectric power station with two pumps—one active and the other in reserve. If the primary pump malfunctions, the second is automatically brought to use. Suppose in a typical day, it is expected that the pump runs for 10 h. According to the specification of the manufacturer, the pumps are expected to fail once every 100 h. What are the chances that such a pump system fails to last for 8 days, i.e., 80 h?

Solution The average number of failures in a 100 h interval is 1. Therefore, the mean of interarrival time between two failures is $1/\lambda$ or 100 h. Interarrival time between two successive failures for each pump is expected to follow an exponential distribution with $\lambda = 1/100$. Since the system failure indicates the simultaneous failure of both the pumps, the interarrival time for the system failure can be assumed to follow gamma distribution with $\alpha = 2$ and $\beta = 100$. Knowing this, let Y denote the time elapsed until the system failure (failure of both the pumps). The probability density function of Y is as follows:

$$f_Y(y) = \frac{1}{100^2\Gamma(2)} e^{-y/100} y^{2-1} = \frac{1}{10000} y e^{-y/100}$$

Therefore, the probability that the system fails to last 80 h is,

$$P\left(Y \leq 80\right) = F_Y\left(y = 80\right) = \int_0^{80} \frac{1}{10000} y \times e^{-y/100} dy$$

Solving it by integration by parts

$$\begin{aligned}
\frac{1}{10000} \int y \times e^{-y/100} dy &= \frac{1}{10000} \left[y \int e^{-y/100} dy - \int \frac{dy}{dy} \left(\int e^{-y/100} dy \right) dy \right] \\
&= \frac{1}{10000} \left[y \frac{e^{-y/100}}{-1/100} - \int 1 \left(\frac{e^{-y/100}}{-1/100} \right) dy \right] \\
&= \frac{1}{10000} \left[y \frac{e^{-y/100}}{-1/100} - \frac{e^{-y/100}}{(-1/100)^2} \right]
\end{aligned}$$

Thus, $P\left(Y \leq 80\right) = \frac{1}{10000} \left[y \frac{e^{-y/100}}{-1/100} - \frac{e^{-y/100}}{(-1/100)^2} \right]_0^{80} = 0.191$

4.2.6 Extreme Value Distribution

Extreme value distribution is a continuous probability distribution used for the analysis of extreme values. The extreme values from a set of random variables can also be assumed to be random. The probability distribution of these extreme values depends on the size of the sample (n) and the distribution from which the sample is drawn. Considering a random sample of size n, let Y be the largest of the sample values.

Now, $P(Y \leq y) = F_Y(y)$ and $P(X_i \leq x) = F_{X_i}(x)$

Hence, $F_Y(y) = P(Y \leq y) = P(\text{all possible values of } x \leq y)$

If the x's are independently and identically distributed, we have,

$$F_Y\left(y\right) = \left[F_X\left(y\right)\right]^n \tag{4.58}$$

$$f_Y\left(y\right) = \frac{dF_Y\left(y\right)}{dy} = n\left[F_X\left(y\right)\right]^{n-1} f_X\left(y\right) \tag{4.59}$$

However, the parent distribution from which the extreme value is observed is not known and cannot be determined. In such cases, if the sample size is large, we can use certain general asymptotic results that depend on limited assumptions concerning the parent distribution of extreme values. Three types of asymptotic distributions have been developed based on different parent distributions, and they are as follows:

(i) Type I—Parent distribution unbounded in direction of the desired extreme, and all the moments of the distribution exist.

(ii) Type II—Parent distribution unbounded in direction of the desired extreme, and all the moments of the distribution do not exist.
(iii) Type III—Parent distribution bounded in the direction of the desired extreme.

In the field of hydrology, many times interest exists in the extreme values of a particular event especially in the cases of flood and drought. The extreme value distribution is specifically used for description of such tail-risk values. Some of such frequently used distributions in hydrology and hydroclimatology are discussed in the following sections.

Example 4.2.7
Assume that time between rains follows an exponential distribution with a mean of 5 days. Also assume that time between rains is independent from one rain to the next. Irrigators might be interested in the maximum time between rains. Over a period of 15 rains, what is the probability that the maximum time between rains is 9 days?

Solution Since the parent distribution is known, we may use Eq. 4.58.

Fifteen rain events mean 14 inter-rain periods or $n = 14$. From Eq. 4.58, the probability that the maximum inter-rain time is less than 9 days is,

$$P\,(Y \leq 9) = F_{Y}\,(y) = [F_{X}\,(y)]^{n}$$

Using exponential distribution, i.e., $\lambda = \frac{1}{\overline{X}} = \frac{1}{5}$

$$F_{X}\,(y) = 1 - e^{-y\lambda} \Rightarrow F_{X}\,(9) = 1 - e^{-\frac{9}{5}}$$

Thus,

$$P\,(Y \leq 9) = [F_{X}\,(y)]^{n} = \left[1 - e^{-9/5}\right]^{14} = 0.08$$

Therefore, the probability that the maximum inter-rain time will be greater than 9 is $= 1 - 0.08 = 0.92$.

Extreme Value Type I (Gumbel Distribution)

Basics: Extreme value type I (EV-I) distribution, also known as Gumbel distribution, is a limiting probability distribution which is used to model the maximum or minimum values from a sample of independent, identically distributed random variables, as the size of the sample increases. Thus, the interpretation could be as follows:

Random Variable: A continuous random variable (X) which is the maximum/minimum of a number of samples of various distribution (e.g., normal or exponential).
Parameters: α and β are the scale and location parameters, where $\beta - \alpha \ln\,(\ln\,2)$ is the median of the distribution and β is the mode of the distribution.

Probability density function: $f_x(x) = \exp\left[\mp(x-\beta)/\alpha - \exp\left(\mp(x-\beta)/\alpha\right)\right]/\alpha$ where $-\infty < x < \infty$; $-\infty < \beta < \infty$; $\alpha > 0$. *The –ve sign implies maximum value, and the +ve sign implies minimum value.*

Descriptions: The EV-I distribution for maximum/minimum values is the limiting model as n approaches infinity for the distribution of the maximum/minimum of n independent values from an initial distribution whose right/left tail is unbounded, that is, the initial cumulative distribution approaches unity (zero) with increasing/decreasing values of the random variable at least as fast as the exponential distribution approaches infinity. The normal, lognormal, exponential, and gamma distribution all meet the requirement for the maximum values, whereas only normal distribution satisfies the conditions for minimum values.

The probability density function of the EV-I distribution is as follows:

$$f_x(x) = \frac{1}{\alpha}\exp\left[\mp(x-\beta)/\alpha - \exp\left(\mp(x-\beta)/\alpha\right)\right] \tag{4.60}$$

where $-\infty < x < \infty; -\infty < \beta < \infty; \alpha > 0$. The –ve sign implies maximum value, and the +ve sign implies minimum value. The *CDF* of the EV-I is as follows:

$$F_x(x) = \exp\left[\mp\exp\left(\mp(x-\beta)/\alpha\right)\right] \tag{4.61}$$

where $-\infty < x < \infty; -\infty < \beta < \infty; \alpha > 0$. The –ve sign implies maximum value, and the +ve sign implies minimum value. The parameters α and β are scale and location parameters with β being the mode of the distribution. The mean, variance, and the skewness coefficient are as follows:

$$E(X) = \beta \pm 0.5772\alpha \tag{4.62}$$

$$Var(X) = 1.645\alpha^2 \tag{4.63}$$

$$\gamma = \pm 1.1396 \tag{4.64}$$

where +ve sign implies maximum, and –ve sign implies minimum.

Applications: In hydrology, the Gumbel distribution is used to analyze variables such as monthly and annual maximum values of daily rainfall or river discharge volumes. It is also used in the frequency analysis of floods.

Example 4.2.8

In a certain stream, the annual maximum daily discharge follows Gumbel distribution with mean value of 12000 m^3/s and standard deviation of 4000 m^3/s. What is the probability that the annual maximum daily discharge will exceed 16000 m^3/s? What is the magnitude of annual maximum daily discharge with a return period of 100 years?

Solution As given, annual maximum daily discharge (X) follows Gumbel distribution. The mean and the standard deviation of the distribution are given as 12000 and 4000 m^3/s, respectively. The parameters α and β can be calculated as follows,

We know $Var\,(X) = 1.645\alpha^2$

$$\Rightarrow \alpha = \sqrt{\frac{4000^2}{1.645}} = 3118.7$$

Also, $E\,(x) = \beta + 0.5772\alpha$

$$12000 = \beta + 0.5772 \times 3118.7$$

$$\beta = 10199.88$$

The required probability can be evaluated as,

$$F_x\,(x > 16000) = 1 - F_x\,(x \leq 16000) = 1 - \exp\left(-\exp\left(-(16000 - 10199.88)\,/3118.7\right)\right) = 0.144$$

The probability that the annual maximum daily discharge will exceed 16000 m^3/s is 0.144.

Let the magnitude with return period 100 years be x.

Then, the $P\,(X > x) = 1/100 = 0.01$, hence,

$$\begin{aligned} &P\,(X > x) = 1 - P\,(X \leq x) \\ \text{or, } &1 - \exp\left(-\exp\left(-\frac{(x - 10199.88)}{3118.7}\right)\right) = 0.01 \\ \text{or, } &x = 24546\,\text{m}^3/\text{s} \end{aligned}$$

Extreme Value Type III (Weibull Distribution)

Basics: In general, extreme value type III (EV-III) distribution can be utilized for the extremes in the direction toward which the parent distribution is limited. It is generally used for minimum values in hydrology and hydroclimatology. EV-III for minimum values is also known as Weibull distribution . The interpretation could be as follows:

Random Variable: A continuous random variable (X) which is the minimum of a sample from an asymmetric distribution and takes nonnegative values.
Parameters: α and β are the scale and location parameters.
Probability density function: $f_X\,(x) = \alpha x^{\alpha-1}\beta^{-\alpha}\exp\left[-\left(x/\beta\right)^{\alpha}\right] \quad x \geq 0;\ \alpha, \beta > 0$

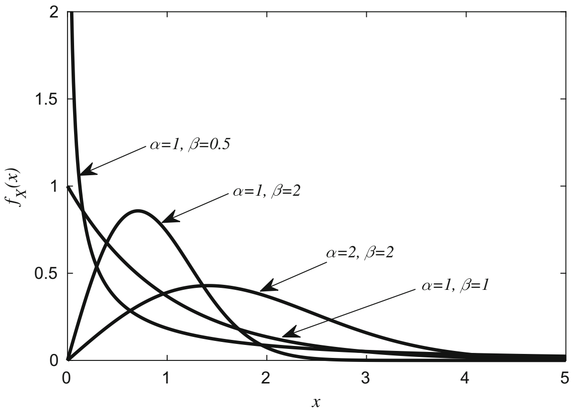


Fig. 4.7 Probability distribution functions of Weibull distribution for different combinations of α and β

Descriptions: The nature of the distribution varies with the change in the shape and scale parameters. Figure 4.7 shows the variation of the nature of the distribution keeping the value of α constant and varying value of β.

The *pdf* and *CDF* of the Weibull distribution are given as follows:

$$f_x(x) = \alpha x^{\alpha-1}\beta^{-\alpha}\exp\left[-\left(x/\beta\right)^{\alpha}\right] \qquad x \geq 0;\, \alpha,\, \beta > 0 \tag{4.65a}$$

$$F_x\,(x) = 1 - \exp\left[-\left(x/\beta\right)^{\alpha}\right] \qquad x \geq 0;\, \alpha,\, \beta > 0 \tag{4.65b}$$

The mean, variance, and the coefficient of skewness are as follows:

$$E\,(X) = \beta\Gamma\left(1 + 1/\alpha\right) \tag{4.66}$$

$$Var\,(X) = \beta^2\left[\Gamma\left(1 + 2/\alpha\right) - \Gamma^2\left(1 + 1/\alpha\right)\right] \tag{4.67}$$

$$\gamma = \frac{\Gamma(1 + 3/\alpha) - 3\Gamma(1 + 2/\alpha)\Gamma(1 + 1/\alpha) + 2\Gamma^3(1 + 1/\alpha)}{\left[\Gamma(1 + 2/\alpha) - \Gamma^2(1 + 1/\alpha)\right]^{3/2}} \tag{4.68}$$

where $\Gamma\,(\bullet)$ is the gamma function as described before (Sect. 4.2.5). Sometimes, in a few applications, the lower bound may not be zero. In such cases, a displacement parameter (ε) must be added to the EV-III distribution for minimums, and the density function becomes:

$$f_x\,(x) = \alpha\,(x - \varepsilon)^{\alpha-1}\,(\beta - \varepsilon)^{-\alpha}\exp\left[-\left\{(x - \varepsilon)/(\beta - \varepsilon)\right\}^{\alpha}\right] \tag{4.69a}$$

$$F_x\,(x) = 1 - \exp\left[-\left\{(x - \varepsilon)/(\beta - \varepsilon)\right\}^{\alpha}\right] \tag{4.69b}$$

Equations 4.69a and 4.69b are also known as three-parameter Weibull distribution. The corresponding mean and variance are as follows:

$$E\left(X\right)=\varepsilon+\left(\beta-\varepsilon\right)\,\Gamma\left(1+1\big/\alpha\right) \tag{4.70}$$

$$Var\left(X\right)=\left(\beta-\varepsilon\right)^{2}\left[\Gamma\left(1+2\big/\alpha\right)-\Gamma^{2}\left(1+1\big/\alpha\right)\right] \tag{4.71}$$

The coefficient of skewness is again given by Eq. 4.68.

Applications: The Weibull distribution can be used most efficiently in hydrology for analysis of low flows in the rivers, as the low flows are naturally lower bounded by zero.

4.2.7 Beta Distribution

Basics: Beta distribution is a continuous probability distribution that represents outcomes for percentages or proportions over an interval, parameterized by two shape parameters. Thus, the interpretation could be as follows:

Random Variable: A continuous random variable (X) which is generally defined within the interval [0, 1].
Parameters: α and β are the shape parameters.
Probability density function:

$$f_X\left(x\right)=\begin{cases}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}\left(1-x\right)^{\beta-1} & \text{for } 0<x<1, \alpha>0,\ \beta>0\\ 0 & \text{elsewhere}\end{cases}$$

Descriptions: Beta distribution has both upper and lower bounds. Thus, if a random variable takes values specifically in the interval (0,1), one choice of probability density can be beta distribution. However, the beta distribution can also be transformed to any interval (a, b). The shape parameters of the distribution vary with the nature of the distribution and are shown in Fig. 4.8. Sometimes, if the limits of the distribution are unknown, it becomes a four-parameter distribution.

Considering the usual case of limits as 0 and 1, the density function is as follows:

$$f_X\left(x\right)=\begin{cases}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}\left(1-x\right)^{\beta-1} & \text{for } 0<x<1, \alpha>0,\ \beta>0\\ 0 & \text{elsewhere}\end{cases} \tag{4.72}$$

The mean, variance, and coefficient of skewness of the beta distribution are given by,

$$E\left(X\right)=\frac{\alpha}{\left(\alpha+\beta\right)} \tag{4.73}$$

$$Var\left(X\right)=\frac{\alpha\beta}{\left(\alpha+\beta\right)^{2}\left(\alpha+\beta+1\right)} \tag{4.74}$$

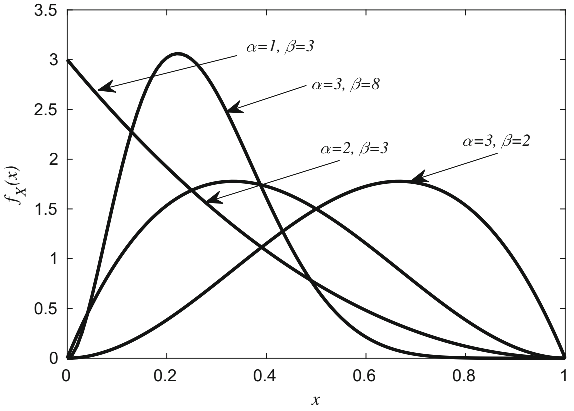


Fig. 4.8 Probability distribution functions of beta distribution for different combinations of α and β

$$\gamma = \frac{2(\beta - \alpha)\sqrt{\alpha + \beta + 1}}{(\alpha + \beta + 2)\sqrt{\alpha\beta}} \tag{4.75}$$

Applications: The beta distribution has been applied to model the behavior of random variables limited to intervals of finite length, for example, volumetric soil moisture content that varies between 0 and 1.

Example 4.2.9

Fifty soil samples are collected from a region and tested in the laboratory for soil moisture content. The data is found to have a mean value of 0.375 and standard deviation of 0.361. If the data follows a beta distribution, develop the probabilistic model for the data. What is the probability of the soil moisture content being below permanent wilting point (PWP), which is 0.11 for that location?

Solution As given, soil moisture (X) follows beta distribution. The mean and the variance are given as 0.375 and 0.361 respectively. The shape parameters (α , β) of the beta distribution can be evaluated as follows,

$$\begin{aligned}
E(X) &= \frac{\alpha}{(\alpha + \beta)} \\
\Rightarrow 0.375 &= \frac{\alpha}{(\alpha + \beta)} \\
\text{Var}\ (X) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \\
0.361^2 &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}
\end{aligned}$$

Solving these equations simultaneously, we get $\alpha = 0.3$ and $\beta = 0.5$.

Thus, the probabilistic model for the data can be written as follows:

$$f\left(x\right)=\begin{cases}\frac{\Gamma(0.3+0.5)}{\Gamma(0.3)\Gamma(0.5)}x^{0.5-1}\left(1-x\right)^{0.3-1} & \text{for } 0<x<1,\ \alpha>0,\ \beta>0\\ 0 & \text{elsewhere}\end{cases}$$

Next, solving numerically,

$$P\left(X\leq 0.11\right)=F_x\left(0.11\right)=\int_0^{0.11}\frac{\Gamma\left(0.3+0.5\right)}{\Gamma\left(0.3\right)\Gamma\left(0.5\right)}x^{0.3-1}\left(1-x\right)^{0.5-1}dx=0.382$$

So, the probability of soil moisture being below the PWP is 0.382.

4.2.8 Pearson and Log-Pearson Type III Distribution

Basics: Pearson type III distribution is a continuous probability distribution. The interpretation could be as follows:

Random Variable: A continuous random variable (X) is such that the distribution is skewed and the mode of the data is at zero.

Parameters: α and β are the scale and shape parameter, respectively.

Probability density function: $f_x\left(x\right)=\frac{\lambda^{\beta}(x-\varepsilon)^{\beta-1}e^{-\lambda(x-\varepsilon)}}{\Gamma(\beta)}$ for $x\geq\varepsilon$

Descriptions: It is one of the seven different types of Pearson distribution. The Pearson type III distribution is a three-parameter distribution from the family of Pearson distributions. It is sometimes called the three-parameter Gamma distribution. The *pdf* is given by,

$$f_x\left(x\right)=\frac{\lambda^{\beta}\left(x-\varepsilon\right)^{\beta-1}e^{-\lambda(x-\varepsilon)}}{\Gamma\left(\beta\right)}\quad\text{for } x\geq\varepsilon\tag{4.76}$$

The lower bound is at $x=\varepsilon$.

If a random variable $Y=\log\left(X\right)$ follows Pearson type III distribution, then the random variable X follows the log-Pearson type III distribution. The *pdf* of log-Pearson type III distribution is given by,

$$f_x\left(x\right)=\frac{\lambda^{\beta}\left(y-\varepsilon\right)^{\beta-1}e^{-\lambda(y-\varepsilon)}}{\Gamma\left(\beta\right)}\quad\text{for } y\geq\varepsilon\tag{4.77}$$

where $y=\log(x)$.

Applications: Both Pearson and log-Pearson type III distributions are used in hydrology and hydroclimatology for frequency analysis. Detailed description is provided in Chap. 5. Pearson distribution can be utilized to evaluate the flood peaks or frequency analysis. Annual maximum flood peaks are generally described by Pearson type III distribution. If the observations present a very highly positively skewed data, then log-Pearson type III distribution is used for modeling. This log transformation reduces the skewness and can even change a positively skewed data to a negatively skewed one.

4.3 Mixed Distribution

Basics: When a random variable has discrete as well as continuous part, it is called a mixed random variable.

Descriptions: Data for some of the hydrologic and hydroclimatic variables may be continuous over a specific range but frequently come across a specific value. For example, daily rainfall data may contain significant number of zero values though it is continuous over nonnegative values. Such data is commonly known as zero-inflated data. Many a times, the nonzero values from such data are treated separately. However, a theoretically sound method of analysis would be to use the Theorem of Total Probability that is given by,

$$P\left(X \geq x\right) = P\left(X \geq x \middle| X = 0\right) P\left(X = 0\right) + P\left(X \geq x \middle| X \neq 0\right) P\left(X \neq 0\right) \tag{4.78}$$

Since $P\left(X \geq x \middle| X = 0\right) P\left(X = 0\right) = 0$, the above expression is reduced to

$$P\left(X \geq x\right) = P\left(X \geq x \middle| X \neq 0\right) P\left(X \neq 0\right) \tag{4.79}$$

In this relationship, $P(X \neq 0)$ would be estimated by the fraction of nonzero values and $P(X \geq x|X \neq 0)$ would be estimated by a standard analysis of the nonzero values with the sample size taken to be equal to the number of nonzero values.

Applications: Many hydrologic variables are bounded on the left by zero. For example, if we wish to find out the distribution of daily rainfall at a particular location, there will be a considerable percentage of zero values. The zero values will follow a discrete distribution, and the nonzero values will follow a continuous distribution. Thereby, overall it will be a mixed distribution. This theory is useful in frequency analysis if data contains significant number of zeros. This is explained in Chap. 5

Example 4.3.1

Consider the proportion of zero daily rainfall in the year 2012 is 0.4. If the nonzero values follow exponential distribution with mean 5 cm, find out the mean of the daily rainfall and the probability of rainfall less than 3 cm.

Solution The *pdf* will be of the form,

$$f(x) = \begin{cases} 0.4 & x = 0 \\ 0.6\lambda e^{-\lambda x} & x > 0 \end{cases}$$

Here, $\lambda = 1/5 = 0.2$. Thereby the *pdf* can be written as,

$$f(x) = \begin{cases} 0.4 & x = 0 \\ 0.12e^{-0.2x} & x > 0 \end{cases}$$

Mean of the daily rainfall can be calculated as,

$$E(x) = \int_0^\infty x f_x(x)\, dx = 0.4 \times 0 + \int_0^\infty x \times 0.12e^{-0.2x} = 3 \text{ cm}$$

Probability of rainfall less than 3 cm can be calculated as,

$$P(x < 3) = 0.4 + \int_0^3 0.12e^{-0.2x} = 0.4 + \frac{3}{5}(1 - e^{-0.2\times 3}) = 0.67$$

4.4 Some Important Distributions of Sample Statistics

4.4.1 Chi-Square Distribution

Basics: The chi-square distribution describes the distribution of a sum of the squares of ν independent standard normal random variables. It is a special case of the gamma distribution and is one of the most widely used probability distributions in inferential statistics like hypothesis testing and construction of confidence intervals. The interpretation could be as follows:

Random Variable: A random variable (X) which is the sum of squares of standard normal distribution and takes positive values always.
Parameters: ν, known as the degree of freedom.
Probability density function: $f_{\chi^2}(x) = \frac{x^{(\nu/2-1)}e^{(-x/2)}}{2^{(\nu/2)}\Gamma(\nu/2)} \qquad x, \nu > 0$

Descriptions: Let us consider the random variables $Z_1, Z_2, \ldots, Z_\nu$ follow standard normal distribution, then

$$Y = \sum_{i=1}^{\nu} Z_i^2 \tag{4.80}$$

follows chi-square distribution with ν degree of freedom. The *pdf* and *CDF* of the chi-square distribution are as follows:

$$f_{\chi^2}(x) = \frac{x^{(\nu/2-1)} e^{(-x/2)}}{2^{(\nu/2)} \Gamma(\nu/2)} \tag{4.81a}$$

$$F_{\chi^2}(x) = \frac{\gamma\left(\frac{\nu}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \tag{4.81b}$$

where $x, \nu > 0$ and $\gamma(p, q)$ is a lower incomplete gamma function that is defined as

$$\gamma(p, q) = \int_0^q t^{p-1} e^{-t} dt \tag{4.82}$$

The chi-square distribution may be linked to gamma distribution. In gamma distribution, if $\alpha = \nu/2$ and $\beta = 2$, it becomes a chi-square distribution with a single parameter ν, known as the degree of freedom. The mean, variance, and coefficient of skewness of chi-square distribution are,

$$E(X) = \nu \tag{4.83}$$

$$Var(X) = 2\nu \tag{4.84}$$

$$\gamma = \sqrt{\frac{8}{\nu}} \tag{4.85}$$

Application: The chi-square distribution is mostly used for statistical inference of variance of a small sample with certain conditions. It could be stated as follows:

If S^2 is the variance of a random sample of size n drawn from normally distributed population with some mean and variance σ^2, then the random variable $\left(\frac{(n-1)S^2}{\sigma^2}\right)$ follows a chi-square distribution with degree of freedom $\nu = n - 1$, where S is the sample standard deviation, computed as $S = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$. In Chap. 6 on hypothesis testing, further applications of chi-square distribution has been explained.

4.4.2 The t-Distribution

Basics: The t-distribution (also known as Students' t-distribution) is a continuous probability distribution utilized when estimating the mean of a normally distributed population in situations where the sample size is small and the variance of the population is unknown. The interpretation could be as follows:

Random Variable: A random variable (X) which is defined as the ratio of two random variables following standard normal distribution and chi-squared distribution, respectively.

Parameters: ν is the degree of freedom.
Probability density function:

$$f_T(t) = \frac{\Gamma\left[(\nu+1)/2\right]\left(1+t^2/\nu\right)^{-(\nu+1)/2}}{\left[\sqrt{\pi\nu}\,\Gamma(\nu/2)\right]} \qquad -\infty < t < \infty;\ \nu > 0$$

Description: Let us consider a random variable Z to follow standard normal distribution and a random variable U to follow chi-square distribution with ν degrees of freedom. Considering Z and U to be independent we may state that,

$$T = \frac{Z\sqrt{\nu}}{\sqrt{U}} \tag{4.86}$$

follows t-distribution with ν degrees of freedom. The expression of *pdf* and *CDF* of t-distribution is as follows:

$$f_T(t) = \frac{\Gamma\left[(\nu+1)/2\right]\left(1+t^2/\nu\right)^{-(\nu+1)/2}}{\left[\sqrt{\pi\nu}\,\Gamma(\nu/2)\right]} \tag{4.87a}$$

$$F_T(t) = \int_{-\infty}^{t} \frac{\Gamma\left[(\nu+1)/2\right]\left(1+t^2/\nu\right)^{-(\nu+1)/2}}{\left[\sqrt{\pi\nu}\,\Gamma(\nu//2)\right]}dt \tag{4.87b}$$

where $-\infty < t < \infty;\ \nu > 0$. The *pdf* of t-distribution is also symmetrical (bell-shaped) like normal distribution. Like standard normal distribution, it has zero mean but the variance depends on the degree of freedom (ν). The mean, variance, and coefficient of skewness of the t-distribution are,

$$E(T) = 0 \tag{4.88}$$

$$Var(T) = \frac{\nu}{\nu-2} \tag{4.89}$$

$$\gamma = 0 \tag{4.90}$$

Thus, as $\nu \to \infty$, variance approaches to 1 and t-distribution approaches to standard normal distribution. Approximately, t-distribution and standard normal distribution are essentially same for a sample size of 30 or more.

Application: The t-distribution is mostly used for statistical inference of mean of a small sample with certain conditions. It could be stated as follows:

If $\overline{X}$ is the mean of a random sample of size n drawn from normally distributed population with mean μ and variance σ^2, then the random variable $\left(\frac{\overline{X}-\mu}{S/\sqrt{n}}\right)$ follows a t-distribution with degree of freedom $\nu = n-1$, where S is the sample standard deviation, computed as $S = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$. In Chap. 6 on hypothesis testing, further applications of t-distribution have been explained.

4.4.3 The F Distribution

Basics: The F distribution is another continuous probability distribution that is asymmetric and takes only positive values. It is the ratio of two random variables following chi-square distribution. The interpretation could be as follows:

Random Variable: A random variable (X) which is defined as the ratio of two random variables following chi-squared distributions.
Parameters: ν_1 and ν_2 are the degrees of freedom.
Probability density function:

$$f_F(x) = \frac{\Gamma\left[(\nu_1+\nu_2)/2\right] \nu_1^{\nu_1/2} \nu_2^{\nu_2/2} x^{(\nu_1-2)/2} (\nu_2+\nu_1 x)^{-(\nu_1+\nu_2)/2}}{\left[\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)\right]} \qquad \nu_1,\ \nu_2,\ x > 0$$

Descriptions: Let us consider two independent random variables U and V to follow chi-square distribution with degree of freedom ν_1 and ν_2, respectively. Then,

$$X = \frac{(U/\nu_1)}{(V/\nu_2)} \tag{4.91}$$

follows F distribution with ν_1 and ν_2 degrees of freedom. The *pdf* and *CDF* of F distribution are as follows:

$$f_F(x) = \frac{\Gamma\left[(\nu_1+\nu_2)/2\right] \nu_1^{\nu_1/2} \nu_2^{\nu_2/2} x^{(\nu_1-2)/2} (\nu_2+\nu_1 x)^{-(\nu_1+\nu_2)/2}}{\left[\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)\right]} \tag{4.92a}$$

$$F_F(x) = \int_0^x \frac{\Gamma\left[(\nu_1+\nu_2)/2\right] \nu_1^{\nu_1/2} \nu_2^{\nu_2/2} x^{(\nu_1-2)/2} (\nu_2+\nu_1 x)^{-(\nu_1+\nu_2)/2}}{\left[\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)\right]} dx \tag{4.92b}$$

where $\nu_1, \nu_2, x > 0$. The mean, variance, and coefficient of skewness of the F distribution are

$$E(X) = \frac{\nu_2}{(\nu_2-2)} \tag{4.93}$$

$$Var(X) = \frac{2\nu_2^2(\nu_2+\nu_1-2)}{\left[\nu_1(\nu_2-2)^2(\nu_2-4)^2\right]} \tag{4.94}$$

$$\gamma = \frac{2(\nu_2+2\nu_1-2)}{\nu_2-6}\sqrt{\frac{2(\nu_2-4)}{\nu_1(\nu_2+\nu_1-2)}} \tag{4.95}$$

One nice property of F distribution is that

$$F_{1-\alpha}\left(\nu_1, \nu_2\right) = \frac{1}{F_\alpha\left(\nu_1, \nu_2\right)} \tag{4.96}$$

where $F_\alpha\left(\nu_1, \nu_2\right)$ and $F_{1-\alpha}\left(\nu_1, \nu_2\right)$ are the values of the random variable such that $P\left(F > F_\alpha\left(\nu_1, \nu_2\right)\right) = \alpha$ and $P\left(F > F_{1-\alpha}\left(\nu_1, \nu_2\right)\right) = 1 - \alpha$ respectively.

Applications: The F distribution is mostly used for statistical inference of variance of two small samples with certain conditions. It could be stated as follows:

If S_1 and S_2 are the standard deviations of two random samples of sizes n_1 and n_2, then the random variable $\left(\frac{S_1^2}{S_2^2}\right)$ follows a F distribution with degrees of freedom $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$. In Chap. 6 on hypothesis testing, further applications of F distribution have been explained.

As mentioned before, mathematical details of all the distributions are summarized in Table 4.2 to facilitate the readers.

4.5 MATLAB Examples

Examples solved in this chapter can also be solved using MATLAB scripts. Following MATLAB built-in functions can be used for fitting different distributions over the data:

- `pd = fitdist(x,dist_name)`
  `fitdist` function is used for fitting parametric distribution over data '`x`.' The argument '`dist_name`' is the name of distribution to be fitted. This function returns a probability distribution object '`pd`' having the details of fitted distribution and its parameters.
- `y = pdf('dist_name',x,A)` or `y = pdf(pd, x)`
  This function can be used for calculating the probability mass/density function. In form of `y = pdf('dist_name',x,A)`, *pdf* or *pmf* is calculated for single-parameter distribution. '`dist_name`' is the distribution name, `x` is the value for which *pdf* or *pmf* is calculated, and `A` is the distribution parameter. Commonly used distribution is supported by this function. In form of `y = pdf(pd, x)`, this function can be used for any probability distribution object '`pd`' (fitted using `fitdist` function). Hence, when `pd` is used, the scope of `pdf` is not limited for one-parameter distributions.
- `y = cdf('dist_name',x,A)` or `y = cdf(pd, x)`
  This function calculates cumulative probability function for `x`. Its arguments are same as `pdf` function.

Apart from these generic functions applicable to commonly used distributions, MATLAB also has many built-in functions for calculating *pdf*, *pmf* and *CDF* for specific distribution. Some of these functions are following:

Table 4.2 Properties of different distributions and the relationship between population parameters and sample statistics

Name of the distribution	Probability mass/distribution function (pmf/pdf)	Cumulative distribution function (CDF)	Range/support	Population parameters and sample statistics
Binomial distribution	$p_x(x) = {}^{n}C_x\, p^x\, (1-p)^{n-x}$	$F_x(x) = \sum_{i=0}^{x} p_x(i)$	$x = 0, 1, \ldots, n$	$\mu = np$ $\sigma^2 = np(1-p)$
Negative binomial distribution	$p_x(x) = {}^{x-1}C_{j-1}\, p^j\, (1-p)^{x-j}$	$F_x(x) = \sum_{i=j}^{x} p_x(i)$	$x = j, j+1, \ldots.$	$\mu = j/p$ $\sigma^2 = j\,(1-p)\big/p^2$
Multinomial distribution	$p_x\left(x_1, \ldots, x_k\right) = \frac{n!}{x_1!\ldots x_k!}\, p_1^{x_1} \ldots p_k^{x_k}$	$F_x(x) = \sum_{i=0}^{x} p_x(i)$	$x_i = 0, 1, \ldots, n$	$\mu = np_i$ $\sigma^2 = np_i\,(1-p_i)$
Hypergeometric distribution	$p_x(x) = \frac{{}^{k}C_x \times {}^{N-k}C_{n-x}}{{}^{N}C_n}$	$F_x(x) = \sum_{i=\max(0,n+k-N)}^{x} p_x(i)$	$x =$ $\max(0, n+k-N)$ $, \ldots, \min(n, k)$	$\mu = nk\big/N$ $\sigma^2 = \frac{nk(N-k)(N-n)}{N^2(N-1)}$
Geometric distribution	$p_x(x) = p\,(1-p)^{x-1}$	$F_x(x) = 1-(1-p)^x$	$x = 1, 2, \ldots, n$	$\mu = 1\big/p$ $\sigma^2 = \frac{(1-p)}{p^2}$
Poisson distribution	$p_x(x) = \lambda^x \frac{e^{-\lambda}}{x!}$	$F_x(x) = \sum_{i=0}^{x} p_x(i)$	$x = 0, 1, \ldots$	$\overline{x} = \lambda$ $S_x^2 = \lambda$
Uniform distribution	$f_x(x) = \frac{1}{\beta-\alpha}$	$F_x(x) = \frac{x-\alpha}{\beta-\alpha}$	$\alpha \le x \le \beta$	$\mu = \frac{(\alpha+\beta)}{2}$ $\sigma^2 = \frac{(\beta-\alpha)^2}{12}$
Exponential distribution	$f_x(x) = \lambda e^{-\lambda x}$	$F_x(x) = 1 - e^{-\lambda x}$	$x \ge 0$	$\overline{x} = \frac{1}{\lambda}$ $S_x^2 = \frac{1}{\lambda^2}$
Normal distribution	$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2\big/2\sigma^2}$	$F_x(x) = \int_{-\infty}^{x} f_x(x)\,dx$	$-\infty < x < \infty$	$\mu = \overline{x}$ $\sigma = S_x$
Lognormal distribution	$f_x(x) = \frac{1}{x\sqrt{2\pi\beta^2}} e^{-\frac{(\ln x-\alpha)^2}{2\beta^2}}$	$F_x(x) = \int_{0}^{x} f_x(x)\,dx$	$x > 0$	$\overline{y} = \alpha$ $S_y = \beta$ where $y = \ln x$

(continued)

Table 4.2 (continued)

Name of the distribution	Probability mass/distribution function (pmf/pdf)	Cumulative distribution function (CDF)	Range/support	Population parameters and sample statistics
Gamma distribution	$f_x\,(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta}$	$F_x\,(x) = \int_0^x f_x\,(x)\;dx$	$x \geq 0$	$\overline{x} = \alpha\beta$ $S_x^2 = \alpha\beta^2$
Extreme value type I (Gumbel) distribution	$f_x\,(x) = \frac{1}{\alpha}\exp\left[\mp\frac{x-\beta}{\alpha} - \exp\left(\mp\frac{x-\beta}{\alpha}\right)\right]$	$F_x\,(x) =$ $\exp\left(\mp\exp\left(\mp\frac{x-\beta}{\alpha}\right)\right)$	$-\infty < x < \infty$	$\alpha = \frac{\sqrt{6}S_x}{\pi}$ $\beta = \overline{x} - 0.5772\alpha$
Extreme value type III (Weibull) distribution	$f_x\,(x) = \alpha x^{\alpha-1}\beta^{-\alpha}\exp\left[-\left(x/\beta\right)^{\alpha}\right]$	$F_x(x) =$ $1 - \exp\left[-\left(x/\beta\right)^{\alpha}\right]$	$x \geq 0$	$\beta = \overline{x}/\left[\Gamma\left(1+1/\alpha\right)\right]$ $\frac{\left[\Gamma(1+2/\alpha)-\Gamma^2(1+1/\alpha)\right]}{[\Gamma(1+1/\alpha)]^2} =$ $\frac{S_x}{\overline{x}^2}$
Beta distribution	$f_x\,(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,.\,\Gamma(\beta)}x^{\alpha-1}\,(1-x)^{\beta-1}$	$F_x\,(x) = \int_0^x f_x\,(x)\;dx$	$0 < x < 1$	$\overline{x} = \frac{\alpha}{(\alpha+\beta)}$ $S_x^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Pearson type III distribution	$f_x\,(x) = \frac{\lambda^{\beta}(x-\varepsilon)^{\beta-1}e^{-\lambda(x-\varepsilon)}}{\Gamma(\beta)}$	$F_x\,(x) = \int_{\varepsilon}^x f_x\,(x)\;dx$	$x \geq \varepsilon$	$\lambda = \frac{s_x}{\sqrt{\beta}},\ \beta = \left(\frac{2}{C_S}\right)^2$ $\varepsilon = \overline{x} - S_x\sqrt{\beta}$
Log-Pearson type III distribution	$f_x\,(x) = \frac{\lambda^{\beta}(y-\varepsilon)^{\beta-1}e^{-\lambda(y-\varepsilon)}}{\Gamma(\beta)}$ where $y = \ln x$	$F_x\,(x) = \int_{e^{\varepsilon}}^x f_x\,(x)\;dx$	$\ln x \geq \varepsilon$	$\lambda = \frac{S_y}{\sqrt{\beta}},\ \beta = \left[\frac{2}{C_S(y)}\right]^2$ $\varepsilon = \overline{y} - S_y\sqrt{\beta}$
Chi-square distribution	$f_{\chi^2}\,(x) = \frac{x^{-(1-\nu/2)}e^{-x/2}}{2^{\nu/2}\,\Gamma(\nu/2)}$	$F_{\chi^2}\,(x) = \int_0^x f_{\chi^2}\,(x)\;dx$	$x > 0$	$\mu = \nu$ $\sigma^2 = \nu^2$
t-distribution	$f_T\,(t) = \frac{\Gamma[(\nu+1)/2]\left(1+t^2/\nu\right)^{-(\nu+1)/2}}{\left[\sqrt{\pi\nu}\,\Gamma(\nu/2)\right]}$	$F_t\,(x) = \int_{-\infty}^x f_t\,(x)\;dx$	$-\infty < t < \infty$	$\mu = 0$ $\sigma^2 = \frac{\nu}{\nu-2}$
F-distribution	$f_F(x) = \frac{\Gamma\left[(\nu_1+\nu_2)/2\right]\nu_1^{\nu_1/2}\,\nu_2^{\nu_2/2}x^{(\nu_1-2)/2}\,(\nu_2+\nu_1 x)^{-(\nu_1+\nu_2)/2}}{\left[\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)\right]}$	$F_F\,(x) = \int_0^x f_t\,(x)\;dx$	$x > 0$	$\mu = \frac{\nu_2}{(\nu_2-2)}$ $\sigma^2 = \frac{\nu_2^2(\nu_1+2)}{\left[\nu_1(\nu_2-2)(\nu_2-4)\right]}$

- `normcdf`, `norminv`, and `normpdf`
  These functions are used for calculating cumulative distribution function, inverse cumulative distribution function, and probability density function for normal or Gaussian distribution.
- `gamcdf`, `gaminv`, and `gampdf`
  These functions serve the similar purpose as above discussed function for gamma distribution.
- `expcdf`, `expinv`, and `exppdf`
  Similar functions for exponential distribution.

Further, many functions having suffix '`rnd`' exist for generating random number following different distributions in MATLAB. These functions are discussed in details in Sect. 8.8 in Chap. 8.

Sample MATLAB scripts for solving examples using the above-mentioned functions are presented here. For instance, the Example 4.1.12 can be solved using script that is shown in Box 4.1.

Box 4.1 Sample MATLAB script for Example 4.1.12

```
clear all
clc

% Inputs
n=4; % number of trials
p=0.1; % probability of success
x=1; %number of flood events

%% Calculation of required probabilities
%Evaluation of the required probability using % Binomial distribution
binomial_prob=binopdf(x,n,p);

%Evaluation of the required probability using Poisson %distribution
lamda=n*p; %shape parameter
poission_pdf = poisspdf(x,lamda);

%% Display Results
output_file=['output' filesep() 'code_1_result.txt'];
delete(output_file);diary(output_file);diary on;
disp('Probability that a flood with 10 years return period')
disp('will occur once in 4 years')
fprintf('\t Using the Binomial distribution is %2.3f.\n', binomial_prob)
fprintf('\t Using the Poisson distribution is %2.3f.\n', poission_pdf)
diary off;
```

The result of the code provided in Box 4.1 is provided in Box 4.2. The answers match with the solution of the Example 4.1.12.

Box 4.2 Results for Box 4.1

```
Probability that a flood with 10 years return
   period
will occur once in 4 years
   Using the Binomial distribution is 0.292.
   Using the Poisson distribution is 0.268.
```

Similarly, the Example 4.2.4 based on normal distribution can be solved using sample script provided in Box 4.3.

Box 4.3 Sample MATLAB script for Example 4.2.4

```
clear all;clc

%% Inputs
mean_temp=10; % mean value
std_temp=5; % variance value

%% Calculate and Display Result
output_file=['output' filesep() 'code_2_result.txt'
   ];
delete(output_file);diary(output_file);diary on;
disp('The probability of mean monthly tmeperature
   being');
fprintf('\t a) between 15C and 24C is %2.3f.\n',...
    normcdf(24,mean_temp,std_temp)-normcdf(15,
       mean_temp,std_temp))
fprintf('\t a) greater than 5C is %2.3f.\n',...
    1-normcdf(5,mean_temp,std_temp))
fprintf('\t a) less than 20C is %2.3f.\n',...
     normcdf(20,mean_temp,std_temp))
diary off;
```

The output of sample code provided in Box 4.3 is provided in Box 4.4. The results obtained using code provided in Box 4.3 are same as obtained in the solution of Example 4.2.4.

Box 4.4 Results for Box 4.3

```
The probability of mean monthly tmeperature being
   a) between 15C and 24C is 0.156.
   a) greater than 5C is 0.841.
   a) less than 20C is 0.977.
```

Exercise

4.1 A weir is designed for a flood with 20-year return period. The design life of the weir is 30 years. What is the probability of at least 1 exceedance during the life of the project? (Ans: 0.215)

4.2 What return period should be used to ensure an 80% chance that the design will not be exceeded in a period of 20 years? (Ans: 90 years)

4.3 In 90 years, the following number of flood was recorded at a specific location. Draw a relative frequency histogram of the data. Fit a Poisson distribution to the data and plot the relative frequencies according to the Poisson distribution on the histogram. Evaluate

(a) The probability of 6 successive years without a flood. (Ans: 0.012)
(b) The probability of exactly 4 years between floods. (Ans: 0.014)

No. of floods in a year	No. of years	No. of floods in a year	No. of years
0	49	4	1
1	25	5	1
2	10	6	0
3	4		

4.4 Two widely separated watersheds are considered to study the peak discharge at the outlet. Considering the peak discharge from the two watersheds to be independent, what is the probability of experiencing a total of 4-year, 10-year events in a 6-year period? (Ans: 0.021)

4.5 A spillway was built to a certain height above the mean water level in the river and has a probability of 0.15 of being overtopped. If the spillway is overtopped, then the probability of damage is 70%; what is the probability that the dam will be damaged within three years? (Ans: 0.283)

4.6 A dam is designed against a flood with 30-year return period. What is the probability that the first such flood will occur within 3 years after the structure is built? (Ans: 0.097)

4.7 The following table presents data for the mean number of days with rainfall more than 10 mm at a particular station.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
No of Days	3	3	3	2	1	3	8	8	4	3	3	3

If the occurrence of rainfall more than 10 mm in any month can be considered as an independent Poisson process, what is the probability of fewer than 40 days with more than 10 mm of rainfall in one year? (Ans: 0.253; *Note: Numerical solution may be required*)

4.8 The time span before first breakdown (T years) of a specialized pump used to extract groundwater follows exponential distribution with mean of 10 years. Find the time span L which a typical pump is 80% certain to exceed. If six pumps are used at a sight, find the probability that at least one of them will breakdown before L years. (Ans: 2.231 years, 0.738)

4.9 The time between occurrences of successive rainstorms at a particular station during monsoon is considered to be exponentially distributed with mean 5 days. What is the probability that the time elapsed for 3 such storms will exceed 40 days? (Ans: 0.014)

4.10 The following table provides the number of rainy days for the monsoon months (June, July, August, and September) over a period of 10 years.

Month	Year									
	1	2	3	4	5	6	7	8	9	10
June	5	9	7	2	1	9	13	1	7	8
July	10	15	17	8	9	10	17	14	20	4
Aug	4	9	8	3	0	10	12	2	8	6
Sep	3	10	6	2	0	11	11	3	10	9

Assuming the number of rainy days to follow normal distribution, what is the probability of 10 or more rainy days in the month of August and September? What is the probability of 30 or more rainy days in the monsoon season (combined for all four months)? (Ans: 0.162, 0.202, 0.536)

4.11 The annual rainfall at a location is considered to follow normal distribution with a mean of 1050 mm and standard deviation of 246 mm at a certain location. The runoff coefficient based on the physical characteristics of the area is considered to vary between 0.75 and 0.9. What is the probability that the annual runoff will exceed 800 mm based on the maximum value of runoff coefficient? (Ans: 0.767)

4.12 The peak discharge at a particular gauging station is considered to be 600 cumec. Discharge at the station is found to be lognormally distributed with mean 500 cumec and standard deviation 85 cumec. What is the probability that the discharge at the station will exceed the peak discharge? (Ans: 0.122)

4.13 At a rain gauge station last 30 years, data indicates the distribution of monthly rainfall as gamma distribution. The sample mean and standard deviation are 18.3 and 5.271 cm respectively. What is the probability that the rainfall will exceed 25 cm in a month? (Ans: 0.108; *Note: Numerical solution may be required*)

4.14 The interarrival time between two droughts follows an exponential distribution with a mean of 6 years. Assuming the interarrival time of drought to be independent events, find the probability that the maximum time between two droughts exceeds 20. (Ans: 0.036)

4.15 The lifetime (in years) of a rain gauge follows Weibull distribution with $\alpha = 1$ and $\beta = 2$. What is the probability that the rain gauge will be in working condition after 3 years. (Ans: 0.223)

Chapter 5
Frequency Analysis, Risk, and Uncertainty in Hydroclimatic Analysis

Analysis of extreme events like severe storms, floods, droughts is an essential component of hydrology and hydroclimatology. The extreme events have catastrophic impact on the entire agro-socioeconomic sector of a country as well as of the whole world. It has aggravated in a changing climate. Thus, it has become really important to predict their occurrences or their frequency of occurrences. This chapter focuses on different methods to analyze these extreme events and forecast their possible future occurrences. At the very beginning of the chapter, the concept of return period has been discussed elaborately which is the building block of any frequency analysis. However, identification of the best-fit probability distribution for a sample data is essential for any frequency analysis. Concept of probability paper is important in this regard, and its construction is discussed along with graphical concept of frequency factor. Next, the concept of frequency analysis is discussed using different parametric probability distributions, such as normal distribution, lognormal distribution, log-Pearson type III distribution, Gumbel's distribution. Basic concepts of risk, reliability, vulnerability, resiliency, and uncertainty are also explained which are inevitable in any kind of hydrologic design based on frequency analysis of extreme events.

5.1 Concept of Return Period

The concept of return period (also sometimes known as 'average recurrence interval' or 'repeat interval') of any hydrologic event (e.g., flood, rainfall, river discharge, landslide, wind storms, tornadoes) plays a key role in risk and uncertainty analysis in hydroclimatic studies. The return period can be defined as *the average length of time for an event of given magnitude to be equalled or exceeded in a statistical sense*. It is basically a statistical measurement typically based on historic data denoting the average recurrence interval of an event over an extended period of time.

R. Maity, *Statistical Methods in Hydrology and Hydroclimatology*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-10-8779-0_5

Table 5.1 Annual maximum discharge data at a gauging station in a river, 1950–1989

Year	Flood discharge (cumecs)	Year	Flood discharge (cumecs)	Year	Flood discharge (cumecs)	Year	Flood discharge (cumecs)
1950	7065	1960	3345	1970	1569	1980	1356
1951	3456	1961	1987	1971	1862	1981	2944
1952	4215	1962	1689	1972	2592	1982	1541
1953	2435	1963	3200	1973	3059	1983	2111
1954	3218	1964	5067	1974	1595	1984	774
1955	4767	1966	4369	1975	1768	1985	911
1956	5368	1966	2589	1976	2987	1986	1123
1957	3891	1967	1306	1977	3679	1987	2884
1958	2015	1968	3761	1978	4597	1988	3868
1959	2498	1969	2450	1979	5582	1989	1812

Let us take an example; consider Fig. 5.1 which show the time series of maximum annual discharge values at some river gauging station from 1950 to 1989, plotted using data given in Table 5.1. Suppose we want to find out the return period of annual maximum discharge of 4000 cumec or more. Now observing Table 5.1 or Fig. 5.1, we can clearly see annual maximum discharge exceeds 4000 cumec 8 times in this period of record. Years of exceedance are 1950, 1952, 1955, 1956, 1964, 1966, 1978, and 1979. Thus, the recurrence intervals (τ) are 2, 3, 1, 8, 2, 12, 1 years. Now according to the definition, return period is the average or expected value of these recurrence intervals, $E(\tau)$. Here, for 4000 cumec discharge magnitude, there are 8 time gaps covering a total period of 29 years, so the return period of '4000 cumec annual maximum discharge' will be $= 29/8 = 3.625$ years. However, this estimate is very rough and simple way of calculating return period directly from the data. This estimate may vary significantly unless the length of data is very large. However, several other procedures are available to compute the return period through some probabilistic assumptions. Such examples are shown later in Examples 5.2.1, 5.4.3, 5.4.4, and 5.4.6 with the same data set.

Let us consider another issue. If it is said that '*at a river gauging station, the stage height with 50* year *return period is* 2 m *above maximum flood level*,' that means the event, i.e., *stage height* of 2 m above *maximum flood level*, or greater, should occur only once in every 50 years on an average at that location. It does not mean that the event will definitely occur once in every 50 years; rather, it indicates average time gap between two such successive events is 50 years.

The definition of return period explained before may slightly be modified in case of lower extreme hydrologic events, e.g., low flows, drought, shortages. For such extreme events, the definition may read as the average time gap (τ) between events with a magnitude equal to or less than a certain value. But still the concept of

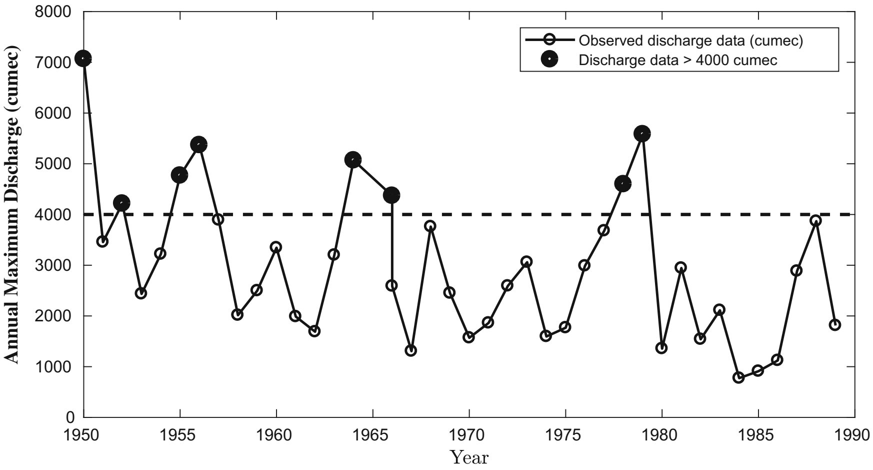


Fig. 5.1 Time series of annual maximum discharges at a river gauging station

'exceedance' can be used to indicate the severity of drought or low flow that exceeds some predefined level toward lower extreme side.

The return period of a hydrologic event can be related to probability of exceedance of that hydrologic event in the following way. Let us consider a hydrologic event as a random variable X, and suppose an extreme event is defined to have occurred if magnitude of X is greater than (or equal to) a level x_T. Now for each observation, there are two possible outcomes, either 'exceedance' (i.e., $X \geq x_T$) or 'non-exceedance' (i.e., $X < x_T$). Let us designate probability of exceedance as $P(X \geq x_T) = p$ and that of non-exceedance as $(1 - p)$. As all the observations are independent, probability mass function (*pmf*) of τ will be the product of probabilities of $\tau - 1$ times non-exceedance followed by one exceedance.

Hence,

$$p(\tau) = (1-p)^{\tau-1} p^1$$

And the expectation of τ, $E(\tau) = \sum_{\tau=1}^{\infty} \tau(1-p)^{\tau-1} p$

$$= p + 2(1-p)p + 3(1-p)^2 p + \ldots$$

$$= p\left[1 + 2(1-p) + 3(1-p)^2 + \ldots\right]$$

Expanding by power of expansion,[1] $E(\tau) = \dfrac{p}{\{1-(1-p)\}^2} = \dfrac{1}{p} = \dfrac{1}{P(X \geq x_T)}$

[1] By power series expansion, $(1+x)^n = 1 + nx + [n(n-1)/2]x^2 + [n(n-1)(n-2)/6]x^3 + \ldots$. So, here, $x = -(1-p)$ and $n = -2$.

Now by the definition of return period (T), it is the average or expected value of recurrence interval and hence

$$T = \frac{1}{P(X \geq x_T)} \tag{5.1}$$

Example 5.1.1
Determine the probability that the annual maximum discharge at the river gauging station (Table 5.1) will equal or exceed 4000 cumec at least once in the next five years. Assume the return period of 4000 cumec discharge is 3.625 years.

Solution Return period of 4000 cumec discharge (T) = 3.625 years. Hence, exceedance probability $P(X \geq 4000)$ can be evaluated as

$$p = \frac{1}{T} = \frac{1}{3.625} = 0.276$$

Thus, the probability that the annual maximum discharge will never exceed in 5 years $= (1 - p)^5$.

Thus, the probability of the same to exceed at least once in 5 years

$$= 1 - (1 - p)^5 = 1 - (1 - 0.276)^5 = 0.801.$$

Example 5.1.2
If the return period of a severe hurricane in North America is 243 years, then find out the probability that no such severe hurricane will occur in next 10 years. Consider occurrence of severe hurricane in North America follows Poisson distribution.

Solution Return period of the event $T = 243$ years.

Hence, exceedance probability $p = \frac{1}{T} = \frac{1}{243} = 0.0041$.

Number of years $n = 10$.

So, for Poisson distribution, $\lambda = np = 10 \times 0.0041 = 0.041$.

Probability of non-occurrence of severe hurricane in next 10 years is

$$p_X\,(0,\ 10) = \frac{0.041^0 \times e^{-0.041}}{0!} = 0.96.$$

Example 5.1.3
If the exceedance probability of a particular flood is 1/50th of its non-exceedance probability, then find out its return period. Also, find out the probability of such an event occurring exactly once in 10 successive years. Consider that the flood follows binomial distribution.

Solution Let us consider exceedance probability = p.

Non-exceedance probability = $q = 50 \times p$ (according to the example statement).

Again, we know

$$\begin{aligned} p + q &= 1 \\ \text{or, } p + 50p &= 1 \\ \text{or, } p &= 0.0196 \\ \text{so, } q &= 50 \times 0.0196 = 0.98 \end{aligned}$$

Now, return period $T = 1/p = 51$ years.

The probability of such an event occurring exactly once in 10 successive years is

$$p\,(1, 10, 0.0196) = {}^{10}C_1\,(0.0196)^1\,(0.98)^9 = 0.163.$$

5.2 Probability Plotting and Plotting Positions Formulae

A probability plot is a plot of magnitude of a particular event versus its probability of exceedance. This type of plot helps to check if a data set fits a particular distribution or not. The plot can be used for interpolation, extrapolation, and comparison purposes. It can be useful for estimating magnitudes with specified return periods. However, any kind of extrapolation must be attempted only when a reasonable fit is assured for the distribution.

The primary step to obtain a probability plot for a given set of data is to determine the probability of exceedance of each data point. Commonly, this technique of determining exceedance probability of a data point is referred to as plotting position. In case of population, the procedure is to determine the fraction of observations greater than or equal to the given magnitude. Thus, exceedance probability of zero is assigned to the largest observation and exceedance probability equal to one is assigned to the smallest observation in the population. However, in case of sample data the range of the population is unknown. So, we cannot assign exceedance probability equal to zero to the largest and exceedance probability equal to one to the smallest data in the sample. So for sample data, this can be analyzed either by empirical methods or by analytical methods. In Table 5.2, plotting position formulae for some of the common empirical methods are listed.

For application of these plotting position formulae, the first task is to arrange the sample data (of size = N) in descending order of magnitude and to assign an order number or rank (m). Thus, for the first member of the arranged data, i.e., for the largest data, $m = 1$ will be assigned, for the second largest data $m = 2$ and so on. Hence, for the smallest data in the sample, $m = N$ will be assigned. Then, using any of the

Table 5.2 Different available plotting position formulae m = rank, N = number of observations, p = probability of exceedance

Name of the method	Plotting position formula
California	$p = \frac{m}{N}$
Hazen	$p = \frac{m-0.5}{N}$
Weibull	$p = \frac{m}{N+1}$
Chegodayev	$p = \frac{m-0.3}{N+0.4}$
Blom	$p = \frac{m-0.44}{N+0.12}$
Gringoten	$p = \frac{m-0.375}{N+0.25}$

above-mentioned formulae (Table 5.2) probability of exceedance (p) is to be calculated for all data in the series. Here, it can be noted that Weibull formula is the most popular among the others and hence only this formula is used for further discussion in this book. After determining p (hence, T which is equal to $1/p$, see Sect. 5.1), we can obtain the probability plot for the given data by plotting its different magnitudes with corresponding probability of exceedance. Plotting different magnitudes of the events with their corresponding return period (T) in semilog or log–log graph (shown in Example 5.2.1) is another popular and useful way of presenting the probability plot. When such probability plot is prepared for flood events (with magnitude Q), then this kind of plot is known as *flood frequency plot*. Depending upon the range of parameters Q and T, the scales in Y- and X-axes can be arithmetic or logarithmic. A best- fit curve (trend line) is drawn through the plotted points, and then by suitable extrapolation of the line, we can either find out the return period of a certain magnitude of the event (which is absent from the sample data series) or find out the magnitude of the event for a particular return period.

This simple procedure yields reasonably good results for small extrapolation, but with increase in extrapolation error increases. Some analytical methods are available for more accurate analysis using Gumbel's extreme value distribution, log-Pearson type III distribution, etc., explained later in this chapter.

There is another common practice to designate a particular magnitude (M) of a hydrological event with percentage dependability. For example, '90% dependable annual precipitation' means on an average for 90% times the annual precipitation to be exceeded or be equalled to that particular magnitude M. In other words, we can expect to observe annual precipitation exceeded or equal to M for on average for 90 years out of 100-year time period.

Example 5.2.1

Consider the annual maximum flood discharge data at a river gauging station over a time period of 40 years as shown in Table 5.1. Construct the flood frequency plot, and estimate the following:

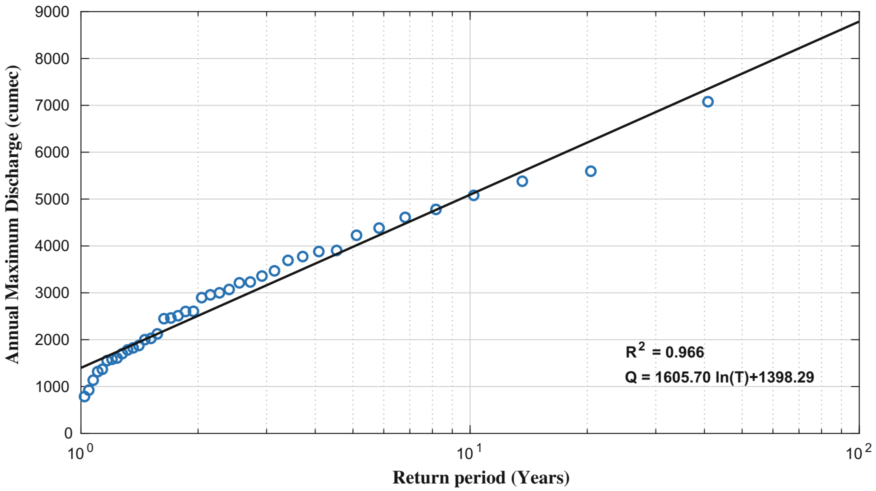


Fig. 5.2 Flood frequency plot for Example 5.2.1

(a) Flood magnitude with return period 10 years, 50 years, and 100 years;
(b) Return period of a flood with 4000 cumec magnitude.

Solution First, the given flood discharge series is arranged in descending order and a rank (m) is assigned to each data point. Here, data length (N) is 40. Exceedance probability of each flood data is calculated by Weibull formula,

$$P = m/(N + 1) = m/41$$

Similarly, return period is determined for each flood discharge magnitude. The rank, ordered flood magnitude, exceedance probability (P), and return period (T) are shown in Table 5.3. A graph is drawn by plotting flood discharge magnitudes (Q) in Y axis (in arithmetic scale) versus return period (T) in X axis (in logarithmic scale), shown in Fig. 5.2. A best-fit line is drawn for the plotted points, and equation of the line is obtained as $Q = 1605.70 \ln(T) + 1398.29$, with coefficient of determination R^2 as 0.966.

(a) From flood frequency plot in Fig. 5.2 or from the equation of trend line, we get
For return period 10 years, flood magnitude is 5095.6 cumec.
For return period 50 years, flood magnitude is 7679.8 cumec.
For return period 100 years, flood magnitude is 8792.8 cumec.
(b) Return period of annual flood of 4000 cumec magnitude is approximately 5 years.

Table 5.3 Calculation for flood frequency plot

Rank m	Flood discharge in descending order (cumecs)	Exceedance probability by Weibull formula $P = \frac{m}{N+1}$	Return period (years) $T = \frac{1}{P}$	Rank m	Flood discharge in descending order (cumecs)	Exceedance probability by Weibull formula $P = \frac{m}{N+1}$	Return period (years) $T = \frac{1}{P}$
1	7065	0.024	41.000	21	2592	0.512	1.952
2	5582	0.049	20.500	22	2589	0.537	1.864
3	5368	0.073	13.667	23	2498	0.561	1.783
4	5067	0.098	10.250	24	2450	0.585	1.708
5	4767	0.122	8.200	25	2435	0.610	1.640
6	4597	0.146	6.833	26	2111	0.634	1.577
7	4369	0.171	5.857	27	2015	0.659	1.519
8	4215	0.195	5.125	28	1987	0.683	1.464
9	3891	0.220	4.556	29	1862	0.707	1.414
10	3868	0.244	4.100	30	1812	0.732	1.367
11	3761	0.268	3.727	31	1768	0.756	1.323
12	3679	0.293	3.417	32	1689	0.780	1.281
13	3456	0.317	3.154	33	1595	0.805	1.242
14	3345	0.341	2.929	34	1569	0.829	1.206
15	3218	0.366	2.733	35	1541	0.854	1.171
16	3200	0.390	2.563	36	1356	0.878	1.139
17	3059	0.415	2.412	37	1306	0.902	1.108
18	2987	0.439	2.278	38	1123	0.927	1.079
19	2944	0.463	2.158	39	911	0.951	1.051
20	2884	0.488	2.050	40	774	0.976	1.025

5.3 Probability Paper

Several probability distribution functions are available in statistics, and some of them have already been discussed in the previous chapter. A set of hydrologic data (sample) is tested with different distribution functions to identify the most suitable probability distribution that best fits the data set.

Probability paper uses a graphical technique to assess whether a given data set follows a certain probability distribution or not. A probability paper is a graph paper representing data and exceedance probability represented in two orthogonal axes. Probability paper is different for different probability distribution functions. In general, the probability axis (generally the Y-axis) is transformed (linearly or non-linearly) in such a way that the resulting cumulative density function appears as a straight line if the data follows that particular distribution. Deviations from this straight line indicate deviations from the specified distribution. In this way, the best-fit distribution is selected for the data set. There are two techniques to transform the probability axis, viz. mathematical construction and graphical construction.

5.3.1 Mathematical Construction of Probability Paper

Probability paper can also be constructed analytically so that the cumulative distribution function $F(x)$ plots as a straight line, on the paper. This can be achieved by transforming the cumulative distribution function to the form

$$Y = mX + c \tag{5.2}$$

where Y is a function of parameter(s) of the distribution and $F(x)$. X is a function of parameter(s) of the distribution and x. m and c are functions of parameters.

For demonstration, let us consider exponential distribution, and the detailed procedure for construction of probability paper is explained. For exponential distribution, the *CDF* is given by

$$F_X(x) = 1 - e^{-\lambda x} \qquad x > 0, \lambda > 0 \tag{5.3}$$

$$\text{or, } 1 - F_X(x) = e^{-\lambda x} \tag{5.4}$$

Taking log on both sides

$$\ln(1 - F_X(x)) = -\lambda x \tag{5.5}$$

Now comparing Eqs. 5.5 and 5.2, we can write

$$\left.\begin{aligned} &Y = -\ln(1 - F_X(x)) \\ &m = \lambda \\ &X = x \\ &c = 0 \end{aligned}\right\} \tag{5.6}$$

Next, assuming a specific value of λ, a set of X and Y is generated to prepare a graph. However, the axes of the graph are labeled with the corresponding values of x and $F_x(x)$, respectively. This is the probability paper for exponential distribution. If any data follows exponential distribution and the corresponding value of cumulative distribution is plotted on this probability paper, it will appear as a straight line. As it is clear from Eq. 5.6, the slope of the line gives the λ value. The entire procedure is illustrated through Example 5.3.1.

Example 5.3.1
Construct probability paper for exponential distribution with $\lambda = \frac{1}{5}$.

Table 5.4 Coordinates of exponential probability paper

$F_x(x)$ (assumed)	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
Y (Eq. 5.6)	0.01	0.11	0.22	0.36	0.51	0.69	0.92	1.20	1.61	2.30	3.00	4.61
$X = x$ (Eq. 5.5)	0.05	0.53	1.12	1.78	2.55	3.47	4.58	6.02	8.05	11.51	14.98	23.03

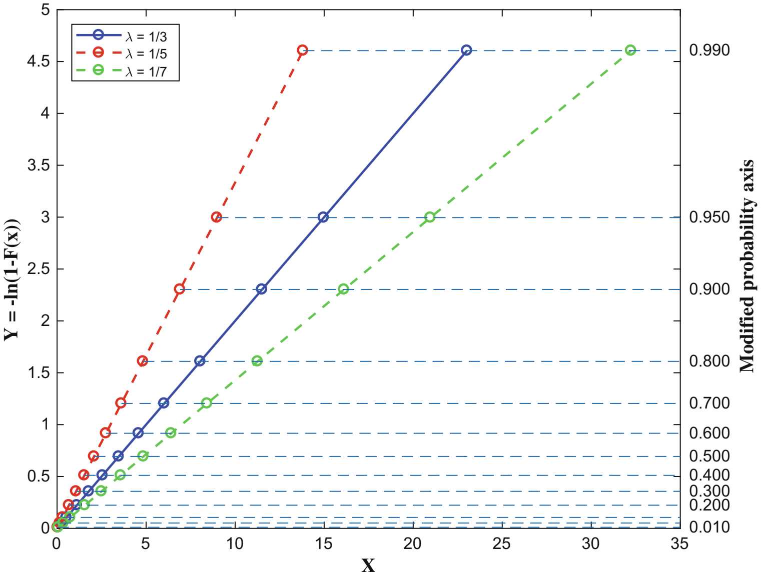


Fig. 5.3 Probability paper for exponential distribution. Straight line for different values of parameter (lambda, λ) is also shown

Solution To construct probability paper, first a table (shown in Table 5.4) is prepared containing $F_x(x)$, Y, and X. The values of $F(x)$ are assumed, and corresponding Y and X values are calculated considering $\lambda = 1/5$. Then, Y is plotted against X, the Y axis is labeled with the corresponding values of $F(x)$ and the X axis is labeled with corresponding values of x in Fig. 5.3. Plots are also shown for $\lambda = 1/3$ and $\lambda = 1/7$ for comparison purpose.

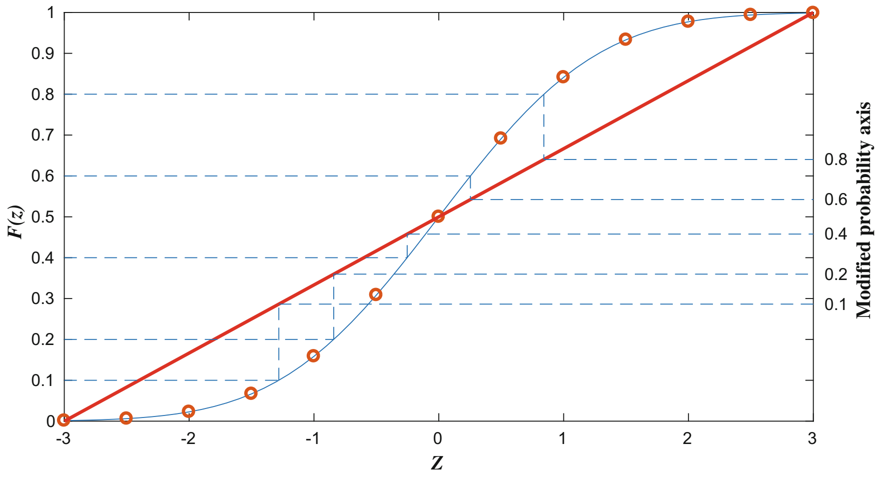


Fig. 5.4 Probability paper for normal distribution

5.3.2 Graphical Construction of Probability Paper

Construction of probability paper can be carried out graphically also. Let us take the example of normal probability paper, which is most widely used to test whether the sample data belongs to a normal population or not. The procedure of graphical construction of normal probability paper is as follows.

The normal probability paper is constructed on the basis of standard normal probability distribution function. Most often, the random variable X (or its standard normal variate Z) is represented on the horizontal arithmetic scale and the vertical axis represents the cumulative probability values $\phi(x)$ or $F(z)$ ranging from 0 to 1 (for a general description of normal distribution, refer to Chap. 4).

First, we consider some random numbers (x) ranging from $-\infty$ to $+\infty$ and calculate their respective z values. Now, from standard normal distribution table (Table B.1 p. 434), we can obtain corresponding *CDF* values, i.e., $F(z)$ values. Then on a simple arithmetic graph paper, these z values are plotted against their $F(z)$ values. For this particular example, we have considered -3 to 3 as the range of Z, as 99% probability is occupied within this limits (for further description, refer to Sect. 4.2.3 in Chap. 4). Then, cumulative distribution function is drawn by plotting z values against their $F(z)$ values, as shown in Fig. 5.4. This distribution generally takes a particular curvilinear shape, which is asymptotic to 0 at $-\infty$ and asymptotic to one at ∞.

Now if we want to test whether a given set of sample data (X) follows normal distribution or not, we can plot the *CDF* of standardized data, $(X-\mu)/\sigma$, and check if it follows approximately this curvilinear shape as shown in Fig. 5.4 (thin continuous line). However, as in general probability paper, the probability axis is transformed in such a way that the CDF appears as a straight line.

The concept to transform the probability axis is as follows. First a straight line from (−1.96, 0.025) to (1.96, 0.975) line (thick continuous line in Fig. 5.4) is drawn and extended to cover the entire probability axis. It can be noted that quantile −1.96 corresponds to cumulative probability of 0.025 for standard normal distribution. Similarly, 1.96 corresponds to cumulative probability of 0.975 (Table B.1, p. 434). Next, starting from a particular value of $F(z)$ (e.g., 0.2), a straight line (dashed line in Fig. 5.4), parallel to X-axis (Z), is drawn till it hits the *CDF* (thin continuous line). Then, a right-angle bend is taken toward the 45° line (thick continuous line). After hitting the 45° line, another right-angle bend is taken to make it parallel to X-axis (Z) again. Next, the line is extended to the secondary Y-axis, the transformed axis, and the point of intersection is labeled as the same value of $F(z)$ from where it was started (i.e., 0.2). In this way, the procedure is repeated for all possible values of $F(z)$ to locate the respective values on the transformed axis. If the transformed axis is noticed carefully, the central part of the axis may be found as more compact than both ends. The combination of X-axis (Z) and the obtained transformed probability axis provides the standard normal probability paper. If a normally distributed set of data is plotted on this probability paper, it will appear as a straight line. Generally, real-life data may not exactly fall on the straight line, and in such cases probabilistic decision is taken from some statistics based on these deviations from the straight line. This requires a hypothesis testing which is discussed in Chap. 6.

Aforementioned concept is general, and probability paper for any distributions can be prepared following the same steps.

Example 5.3.2

The following table shows 20-year annual rainfall data (mm) for a catchment. Check whether this rainfall data follows normal distribution, by using normal probability paper.

Year	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
Annual rainfall (mm)	515.5	257.2	277.3	498.6	806.5	346.1	574.3	454.9	723.5	282.2
Year	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
Annual rainfall (mm)	506.5	610.5	720.1	808.8	517.2	201.7	351.5	287.7	970.1	376.9

Solution First, the rainfall data is sorted in descending order and the exceedance probability is calculated using the Weibull formula. For probability paper, we need to plot the random variable against their cumulative probability which is actually their non-exceedance probability. All the calculations are listed in Table 5.5. Then on a normal probability paper, rainfall data is plotted against cumulative probability. From Fig. 5.5, we can see that the given data points (. sign) are found to follow an approximate straight line. Hence, we can conclude that the given annual rainfall data approximately follows a normal distribution. There are some statistical tests to check goodness-of-fit, i.e., how good the data fits the distribution. However, knowledge of

Table 5.5 Calculation for Example 5.3.2

Sorted annual rainfall data (mm)	Rank (m)	Exceedance probability by Weibull formula	Cumulative probability = non-exceedance probability	Sorted annual rainfall data (mm)	Rank (m)	Exceedance probability by Weibull formula	Cumulative probability = non-exceedance probability
970.1	1	0.0476	0.9524	498.6	11	0.5238	0.4762
808.8	2	0.0952	0.9048	454.9	12	0.5714	0.4286
806.5	3	0.1429	0.8571	376.9	13	0.6190	0.3810
723.5	4	0.1905	0.8095	351.5	14	0.6667	0.3333
720.1	5	0.2381	0.7619	346.1	15	0.7143	0.2857
610.5	6	0.2857	0.7143	287.7	16	0.7619	0.2381
574.3	7	0.3333	0.6667	282.2	17	0.8095	0.1905
517.2	8	0.3810	0.6190	277.3	18	0.8571	0.1429
515.5	9	0.4286	0.5714	257.2	19	0.9048	0.0952
506.5	10	0.4762	0.5238	201.7	20	0.9524	0.0476

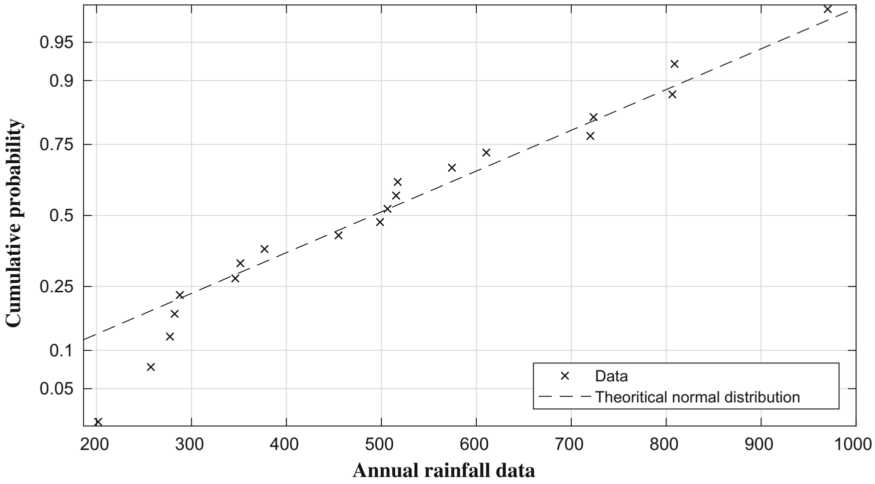


Fig. 5.5 Normal probability paper used for Example 5.3.2

hypothesis testing is required, which is discussed in Chap. 6. Thus, statistical tests to check the goodness-of-fit are explained in Chap. 6.

5.4 Frequency Analyses of Hydroclimatic Extremes

When the magnitude of a hydroclimatic event differs significantly from the average or usual range of magnitudes, then such events are termed as extreme events. This may take place over one day or a period of time, e.g., severe storms, flash floods, droughts. These types of hydroclimatic extreme events influence the system to a great extent. Frequency analysis is done to determine the frequency of occurrence (or probability of occurrence) of such extreme events.

Frequency analysis generally refers to stationary frequency analysis that assumes the data to be stationary. Most of the frequency distribution functions in hydroclimatic studies can be expressed in the form of the following equation, known as the *general equation of frequency analysis*, given by

$$x_T = \overline{x} + KS \tag{5.7}$$

where

x_T = magnitude of the hydrologic variable with a return period of T;
$\overline{x}$ = mean of the hydrologic variable;
S = standard deviation of the hydrologic variable; and

K = frequency factor, a function of the return period T and the assumed frequency distribution function.

Different probability distribution functions are available for the prediction of extreme events. Some of them are listed below

(i) Normal distribution
(ii) Lognormal distribution
(iii) Log-Pearson type III distribution
(iv) Extreme value type I distribution (or Gumbel's distribution)
(v) Mixed distribution.

Estimation of frequency factors using all these above-mentioned distributions is discussed in the following sections of this chapter.

5.4.1 Normal Distribution

General description of normal distribution is explained in Chap. 4. If a hydrologic variable (X) follows normal distribution, the frequency factor K equals its standard normal variate Z. From Eq. 5.7, we can express K as $K = (x_T - \overline{x})\big/S$, which is the standard normal variate Z. So, in order to determine an extreme event with a particular return period, we have to calculate its exceedance probability (hence non-exceedance probability) and corresponding Z value using a standard normal table. Now using this Z value as frequency factor K, the extreme event can be determined from Eq. 5.7 (shown in Example 5.4.1).

Example 5.4.1
Consider a 50-year data of annual maximum 24 h rainfall depth at a particular place follows normal distribution with mean 92.5 mm and standard deviation 34 mm. Determine the magnitude of annual maximum rainfall with return period of 20 years.

Solution For the given 50-year data, mean $\overline{x} = 92.5$ mm and standard deviation $S =$ 34 mm.

Now, for 20-year return period, $T = 20$; $P(X > x_{20}) = \frac{1}{20} = 0.05$

$$P(X \leq x_{20}) = 1 - 0.05 = 0.95$$

From a standard normal table (Table B.1), for $\phi(Z) = 0.95$, $Z = 1.645$

Thus, the frequency factor $K = 1.645$.

From Eq. 5.7, $x_{20} = \overline{x} + KS = 92.5 + (1.645 \times 34) = 148.43$ mm.

Example 5.4.2
Consider the data used in Example 5.2.1, and determine the 10-, 50-, and 100-year floods using normal distribution.

Solution For the given maximum flood data series (X), the mean flood magnitude $\overline{X}$ is 2932.6 and standard distribution (S_x) of 1427.2. Now, for a 10-year flood, $T = 10$; $P(X > x_{10}) = \frac{1}{10} = 0.1$

$$P(X \leq x_{10}) = 1 - 0.1 = 0.90 = \phi(Z)$$

From a standard normal table (Table B.1), we got $Z = 1.282 =$ the frequency factor K.

Hence, $x_{10} = 2932.6 + 1.282 \times 1472.2 = 4762.3$ cumec

Similarly, we can obtain $x_{50} = 5863.7$ cumec and $x_{100} = 6252.8$ cumec.

5.4.2 *Lognormal Distribution*

General description of lognormal distribution is explained in Chap. 4. If a hydrologic variable (X) follows lognormal distribution, we have to transform the X values into a series of Y values where $Y = \ln(X)$. As X follows lognormal distribution, Y will follow normal distribution. Then, we have to follow the same procedure explained in Sect. 5.4.1 to determine the frequency factor for Y series. Then using Eq. 5.8 (same as Eq. 5.7 but for variable Y), we can determine the magnitude y_T for a particular return period T and from y_T we can compute x_T by taking antilog(y_T)

$$y_T = \overline{y} + K_y S_y \tag{5.8}$$

Just like the previous case, here y_T= magnitude of the variable Y with a return period of T, $\overline{y}$= mean of the magnitudes of Y, S_y= standard deviation of the magnitudes of Y, and K_y= frequency factor for Y.

The values of $\overline{y}$ and S_y can also be computed from the mean ($\overline{x}$) and standard deviation (S_x) of the original data. The equations are as follows:

$$\begin{aligned} \overline{y} &= \frac{1}{2} \ln\left[\frac{\overline{x}^2}{C_v^2 + 1}\right] \\ S_y &= \sqrt{\ln\left(C_v^2 + 1\right)} \end{aligned} \tag{5.9}$$

where $C_v = \frac{S_x}{\overline{x}}$.

Example 5.4.3

Consider the Example 5.2.1, and determine the 10-, 50-, and 100-year floods using lognormal distribution.

Solution For the given maximum flood data series (X), convert the X values into a series of Y values where $y = \ln(x)$. Now, mean and standard deviation are calculated for this Y series and obtained as mean $(\overline{y}) = 7.862$ and standard deviation $(S_y) = 0.514$.

Now, for a 10-year flood, $T = 10$; $P(Y > y_{10}) = \frac{1}{10} = 0.1$

$$P(Y \leq y_{10}) = 1 - 0.1 = 0.90 = \phi(Z_{0.1})$$

From a standard normal table (Table B.1), we got $Z_{10} = 1.282 =$ the frequency factor K.

From Eq. 5.8, $y_{10} = \overline{y} + K_y S_y = 7.862 + (1.282 \times 0.514) = 8.52$.

Thus, $y_{10} = \ln\left(x_{10}\right) = 8.52$, hence, $x_{10} = 5014$ cumec.

Similarly, we can obtain $x_{50} = 7445$ cumecs and $x_{100} = 8566$ cumec.

5.4.3 Log-Pearson Type III Distribution

Pearson and log-Pearson type III distribution are discussed in Chap. 4. As mentioned, these distributions are popularly used for flood frequency analysis. The idea to estimate frequency factor by these distributions is very much similar to that of lognormal distribution, discussed in previous Sect. 5.4.2. First, we have to convert the X values into a series of Y values where $y = \log_{10}(x)$. Then, three statistical parameters are calculated for this transformed data series Y, namely mean $(\overline{y})$, standard deviation (S_y), and coefficient of skewness (C_s). Now, frequency factors are obtained from Table 5.6 for a particular return period or exceedance probability. When C_s takes the value zero, log-Pearson type III distribution becomes lognormal distribution. Next, the magnitude y_T for a particular return period T can be computed using Eq. 5.8. The value of x_T can be computed from y_T, using antilog(y_T). The formula to calculate C_s from the sample data is shown in Table 3.1 (p. 65) and reproduced here as follows:

$$C_s = \frac{n}{(n-1)(n-2)} \frac{\sum (y - \overline{y})^3}{S_y^3} \tag{5.10}$$

where n is the total number of data.

Example 5.4.4
Consider the Example 5.2.1, and determine the 10-, 50-, 100-year floods using log-Pearson type III distribution.

Solution For the given flood data series (X), convert the X values into a series of Y values where $y = \log_{10}(x)$. Now, three parameters are calculated for this Y series and obtained as mean $(\overline{y}) = 3.415$, std. deviation $(s_y) = 0.224$, and coefficient of skewness $(C_s) = -0.33$. Now, for a 10-year flood, $T = 10$; $P(Y > y_{10}) = \frac{1}{10} = 0.1$.

From Table 5.6, we got K_{10}= 1.245 for C_s= −0.3 and K_{10}= 1.231 for C_s= −0.4

Table 5.6 Frequency factors for log-Pearson type III distribution

Coefficient of skewness C_s	Return period in years							
	1.0101	2	5	10	25	50	100	200
	Exceedance probability							
	0.99	0.5	0.2	0.1	0.04	0.02	0.01	0.005
3.0	−0.667	−0.396	0.42	1.18	2.278	3.152	4.051	4.97
2.9	−0.69	−0.39	0.44	1.195	2.277	3.134	4.013	4.904
2.8	−0.714	−0.384	0.46	1.21	2.275	3.114	3.973	4.847
2.7	−0.74	−0.376	0.479	1.224	2.272	3.093	3.932	4.783
2.6	−0.769	−0.368	0.499	1.238	2.267	3.071	3.889	4.718
2.5	−0.799	−0.36	0.518	1.25	2.262	3.048	3.845	4.652
2.4	−0.832	−0.351	0.537	1.262	2.256	3.023	3.8	4.584
2.3	−0.867	−0.341	0.555	1.274	2.248	2.997	3.753	4.515
2.2	−0.905	−0.33	0.574	1.284	2.24	2.97	3.705	4.444
2.1	−0.946	−0.319	0.592	1.294	2.23	2.942	3.656	4.372
2.0	−0.99	−0.307	0.609	1.302	2.219	2.912	3.605	4.298
1.9	−1.037	−0.294	0.627	1.31	2.207	2.881	3.553	4.223
1.8	−1.087	−0.282	0.643	1.318	2.193	2.848	3.499	4.147
1.7	−1.14	−0.268	0.66	1.324	2.179	2.815	3.444	4.069
1.6	−1.197	−0.254	0.675	1.329	2.163	2.78	3.388	3.99
1.5	−1.256	−0.24	0.69	1.333	2.146	2.743	3.33	3.91
1.4	−1.318	−0.225	0.705	1.337	2.128	2.706	3.271	3.828
1.3	−1.383	−0.21	0.719	1.339	2.108	2.666	3.211	3.745
1.2	−1.449	−0.195	0.732	1.34	2.087	2.626	3.149	3.661
1.1	−1.518	−0.18	0.745	1.341	2.066	2.585	3.087	3.575
1.0	−1.588	−0.164	0.758	1.34	2.043	2.542	3.022	3.489
0.9	−1.66	−0.148	0.769	1.339	2.018	2.498	2.957	3.401
0.8	−1.733	−0.132	0.78	1.336	1.993	2.453	2.891	3.312
0.7	−1.806	−0.116	0.79	1.333	1.967	2.407	2.824	3.223
0.6	−1.88	−0.099	0.8	1.328	1.939	2.359	2.755	3.132
0.5	−1.955	−0.083	0.808	1.323	1.91	2.311	2.686	3.041
0.4	−2.029	−0.066	0.816	1.317	1.88	2.261	2.615	2.949
0.3	−2.104	−0.05	0.824	1.309	1.849	2.211	2.544	2.856
0.2	−2.178	−0.033	0.83	1.301	1.818	2.159	2.472	2.763
0.1	−2.252	−0.017	0.836	1.292	1.785	2.107	2.4	2.67
0.0	−2.326	0	0.842	1.282	1.751	2.054	2.326	2.576
−0.1	−2.4	0.017	0.846	1.27	1.716	2	2.252	2.482
−0.2	−2.472	0.033	0.85	1.258	1.68	1.945	2.178	2.388
−0.3	−2.544	0.05	0.853	1.245	1.643	1.89	2.104	2.294

(continued)

Table 5.6 (continued)

Coefficient of skewness C_s	Return period in years							
	1.0101	2	5	10	25	50	100	200
	Exceedance probability							
	0.99	0.5	0.2	0.1	0.04	0.02	0.01	0.005
−0.4	−2.615	0.066	0.855	1.231	1.606	1.834	2.029	2.201
−0.5	−2.686	0.083	0.856	1.216	1.567	1.777	1.955	2.108
−0.6	−2.755	0.099	0.857	1.2	1.528	1.72	1.88	2.016
−0.7	−2.824	0.116	0.857	1.183	1.488	1.663	1.806	1.926
−0.8	−2.891	0.132	0.856	1.166	1.448	1.606	1.733	1.837
−0.9	−2.957	0.148	0.854	1.147	1.407	1.549	1.66	1.749
−1.0	−3.022	0.164	0.852	1.128	1.366	1.492	1.588	1.664
−1.1	−3.087	0.18	0.848	1.107	1.324	1.435	1.518	1.581
−1.2	−3.149	0.195	0.844	1.086	1.282	1.379	1.449	1.501
−1.3	−3.211	0.21	0.838	1.064	1.24	1.324	1.383	1.424
−1.4	−3.271	0.225	0.832	1.041	1.198	1.27	1.318	1.351
−1.5	−3.33	0.24	0.825	1.018	1.157	1.217	1.256	1.282
−1.6	−3.88	0.254	0.817	0.994	1.116	1.166	1.197	1.216
−1.7	−3.444	0.268	0.808	0.97	1.075	1.116	1.14	1.155
−1.8	−3.499	0.282	0.799	0.945	1.035	1.069	1.087	1.097
−1.9	−3.553	0.294	0.788	0.92	0.996	1.023	1.037	1.044
−2.0	−3.605	0.307	0.777	0.895	0.959	0.98	0.99	0.995
−2.1	−3.656	0.319	0.765	0.869	0.923	0.939	0.946	0.949
−2.2	−3.705	0.33	0.752	0.844	0.888	0.9	0.905	0.907
−2.3	−3.753	0.341	0.739	0.819	0.855	0.864	0.867	0.869
−2.4	−3.8	0.351	0.725	0.795	0.823	0.83	0.832	0.833
−2.5	−3.845	0.36	0.711	0.711	0.793	0.798	0.799	0.8
−2.6	−3.899	0.368	0.696	0.747	0.764	0.768	0.769	0.769
−2.7	−3.932	0.376	0.681	0.724	0.738	0.74	0.74	0.741
−2.8	−3.973	0.384	0.666	0.702	0.712	0.714	0.714	0.714
−2.9	−4.013	0.39	0.651	0.681	0.683	0.689	0.69	0.69
−3.0	−4.051	0.396	0.636	0.66	0.666	0.666	0.667	0.667

So, for $C_s = -0.33$, by linear interpolation, K_{10}=$1.245 - \frac{1.245-1.231}{0.4-0.3} \times (0.33 - 0.3) = 1.241$.

From Eq. 5.10, $y_{10} = \overline{y} + K_{10}s_y = 3.415 + (1.241 \times 0.224) = 3.690$

$y_{10} = \log_{10}\left(x_{10}\right) = 3.690$. Hence, $x_{10} = 4898$ cumec.

Similarly, we can obtain $x_{50} = 6832$ cumec and $x_{100} = 7613$ cumec.

5.4.4 Extreme Value Type I Distribution

Details of extreme value type I distribution, also known as Gumbel distribution, is discussed in Chap. 4. In hydrology and hydroclimatology, extreme value type I distribution is mostly used to analyze extreme events like flood peaks, maximum rainfall. For a hydrologic extreme event X, following Gumbel's distribution, exceedance probability of $X = x_0$ is given by

$$P(X \geq x_0) = 1 - e^{-e^{-(x-\beta)/\alpha}} \tag{5.11}$$

Let us simplify the equation by introducing a dimensionless variable y known as Gumbel's reduced variate, given by $y = (x - \beta)\big/\alpha$, where α and β are scale and location parameter of Gumbel's distribution, respectively. So, Eq. 5.11 can be rearranged as

$$\begin{aligned} P(X \geq x_0) &= 1 - e^{-e^{-(x-\beta)/\alpha}} \\ \Rightarrow P(X \geq x_0) &= 1 - e^{-e^{-y}} = \frac{1}{T} \\ \text{or, } y &= -\left[\ln\left(\ln\frac{T}{T-1}\right)\right] \end{aligned} \tag{5.12}$$

For Gumbel's distribution, standard deviation and mean are given by

$$\begin{aligned} &S_x = 1.2825\alpha && \Rightarrow \alpha = S_x\big/1.2825 \\ &\overline{x} = \beta + 0.5772\alpha && \Rightarrow \beta = \overline{x} - 0.5772\alpha && \Rightarrow \beta = \overline{x} - 0.4501 S_x \end{aligned}$$

Using above equations of α and β, we can express y as

$$y = \frac{(x-\beta)}{\alpha} = \frac{1.2825(x-\overline{x})}{S_x} + 0.5772$$

Now for a particular return period T, let us designate y as y_T and x as x_T, then

$$y_T = \frac{1.2825(x_T - \overline{x})}{S_x} + 0.5772$$

$$\text{or, } x_T = \overline{x} + \frac{(y_T - 0.5772)}{1.2825} S_x \qquad \text{or, } x_T = \overline{x} + K S_x \tag{5.13}$$

Equation 5.13 is the general equation for hydrologic frequency analysis (Eq. 5.7), where the frequency factor $K = \frac{(y_T - 0.5772)}{1.2825}$.

Table 5.7 Frequency factors of Gumbel's extreme value for different return periods and finite sample sizes

Samplesize (N)	Return periods (T years)					
	2.33	5	10	25	50	100
15	0.06	0.97	1.70	2.63	3.32	4.01
20	0.05	0.91	1.63	2.52	3.18	3.84
25	0.04	0.89	1.58	2.44	3.09	3.73
30	0.04	0.87	1.54	2.39	3.03	3.65
40	0.03	0.84	1.50	2.33	2.94	3.55
50	0.03	0.82	1.47	2.28	2.89	3.49
60	0.02	0.81	1.45	2.25	2.85	3.45
70	0.02	0.80	1.43	2.23	2.82	3.41
80	0.02	0.79	1.42	2.21	2.80	3.39
100	0.02	0.77	1.40	2.19	2.77	3.35
200	0.01	0.74	1.33	2.08	2.63	3.18
400	0.00	0.70	1.27	1.99	2.52	3.05

Equation 5.13 constitutes basic Gumbel's equations and is only applicable to a sample of infinite size (i.e., sample size $N \rightarrow \infty$). But in practice, annual data series of extreme hydrological events like maximum flood, maximum rainfall are of finite sample size. Hence, Eq. 5.13 is modified to take care of the finite sample size N as shown below.

$$x_T = \overline{x} + K\,S_x \tag{5.14}$$

$$K = \frac{(y_T - \overline{y}_n)}{S_n} \tag{5.15}$$

where $y_T = -\left[\ln\left(\ln \frac{T}{T-1}\right)\right]$ is reduced variate for return period T and $\overline{y}_n$ is reduced mean, a function of T and sample size$N$; as $N \rightarrow \infty, \overline{y}_n \rightarrow 0.5772$. S_n is the reduced standard deviation, a function of T and sample size N; as $N \rightarrow \infty$, $S_n \rightarrow 1.2825$.

Tables are available for determining $\overline{y}_n$ and S_n for a certain sample size (N) and return period (T). Reduced variate y_T can be directly calculated from Eq. 5.12. Then, K can be calculated by Eq. 5.15. All these steps of determining frequency factors (K) have been summed up and listed in Table 5.7, which directly gives the values of frequency factors (K) for different sample size (N) and return period (T). Instead of doing all the steps shown above, readers can directly estimate the values of K from this table and use to predict some T years extreme event, for a given finite sample size. If the given sample size is said to be infinite (practically very large), then we do not need to use this table, and we can directly use Eq. 5.13 to calculate K. Examples 5.4.5 and 5.4.6 illustrate the process further.

Example 5.4.5

The mean annual maximum daily rainfall in a city is 105 mm and standard deviation is 45 mm. Determine the depth of daily rainfall with 5-year return period in that city. Use Gumbel's method and assume sample size to be very large.

Solution If X is a random variable which describes mean annual maximum daily rainfall in a city, then mean $(\overline{x}) = 105$ and standard deviation $(S_x) = 45$. For 5-year return period, reduced variate will be

$$y_T = -\left[\ln\left(\ln\frac{T}{T-1}\right)\right] = -\left[\ln\left(\ln\frac{5}{5-1}\right)\right] = 1.5$$

As the sample size (N) is very large, we can directly use Eq. 5.13 to evaluate frequency factor K as shown below

$$K = \frac{(y_T - 0.5772)}{1.2825} = \frac{1.5 - 0.5772}{1.2825} = 0.72$$

Now, the depth of annual maximum daily rainfall with 5-year return period in the city $x_T = \overline{x} + K S_x = 105 + (0.72 \times 45) = 137.40$ mm.

Gumbel or Extreme Value Probability Paper

The Gumbel or extreme value probability paper helps to verify whether the given data follows the Gumbel distribution or not. In this probability paper, the X-axis is used to represent return period (T). First, y_T values are plotted on an arithmetic scale parallel to X-axis, say from -2 to 5, as shown in Fig. 5.6. Then, some values of T (e.g., 2, 10, 50, 100 years) are chosen and corresponding y_T values are marked on X-axis. Thus, the X-axis is prepared. The Y-axis of the probability paper is used to represent the value of the variate x_T either in arithmetic scale or in logarithmic scale. From Eqs. 5.14 and 5.15, we can see y_T varies linearly with x_T. So, a Gumbel distribution will plot as straight line on a Gumbel paper and linear interpolation/extrapolation is carried out to evaluate any other value including extreme values with certain return period.

In order to check if a given set of data follows Gumbel's distribution, value of x_T for some particular T (maybe 2–3 values of T, where $T < N$) is calculated using Eqs. 5.14 and 5.15 and those 2–3 computed data points are plotted on the Gumbel probability paper. As per the linear property explained in previous paragraph, these points will lie on a straight line.

So, for the theoretical Gumbel distributions curve, only two points are enough to draw the straight line. In case of unavailability of the Gumbel paper, a semilog plot with logarithmic scale of T can be used but then a large set of (x_T, T) values are required to identify the theoretical curve. Next, it is checked whether the theoretical Gumbel distribution curve fits the observed data points or not. Gumbel's distribution

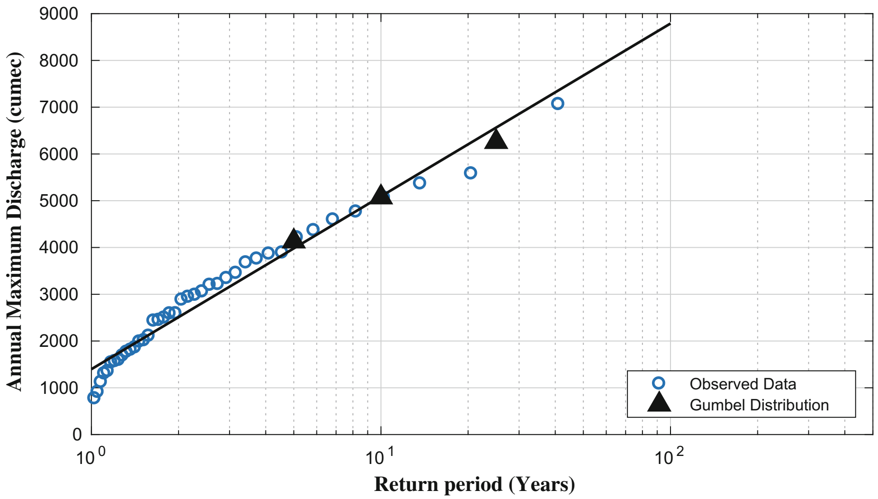


Fig. 5.6 Flood frequency analysis by the Gumbel distribution for Example 5.4.6

has one important property; i.e., the value of x_T at $T = 2.33$ years gives the average value of the data series if N is very large. Hence, the theoretical plot of x_T versus T must pass through this point.

Example 5.4.6
Consider the Example 5.2.1, and verify whether the Gumbel extreme value distribution fits this data series. Then, determine 50-year flood and 100-year flood using linear extrapolation.

Solution For the given maximum flood data series (X), mean and standard deviation are calculated as mean, $\overline{x} = 2932.625$ cumec and standard deviation, $S_x = 1427.193$ cumec.

As in this case, we are using a semilog plot for verifying the given data, and we need to estimate x_T values for different T values. The sample size (N) is 40 here, and hence consider $T < 40$. Let us take $T = 5$, 10 and 25 years. The values of K from Table 5.7 for $N = 40$ are as follows: $K_5 = 0.84$, $K_{10} = 1.5$, and $K_{25} = 2.33$

Now, we can calculate $x_{10} = \overline{x} + K_{10}\, S_x = 2932.6 + (1.5 \times 1427.2) = 5073.4$ cumec.

Similarly, we get $x_5 = 4131$ cumec and x_{25} =6258 cumec.

Figure 5.6 shows a semilogarithmic graph paper, where all the observed maximum annual flood data (as per Weibull formula) and above computed seven points are plotted along with their return periods. A best-fit line is drawn through these seven points. This straight line basically indicates the theoretical Gumbel distribution. Observing this figure, we can see how well the given data series fits the theoretical Gumbel distribution.

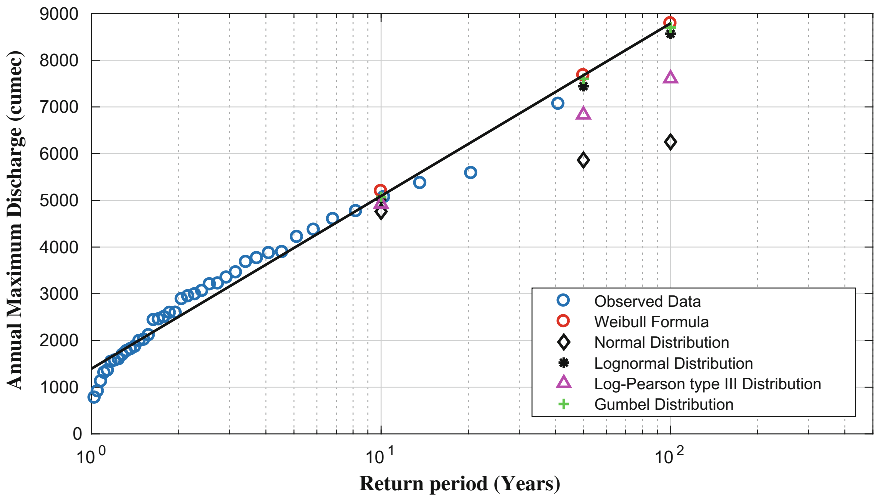


Fig. 5.7 Graphical comparison of results obtained from different methods

Table 5.8 Comparison of results obtained in different methods

Annual maximum flood (cumec)	Return period		
	10 years	50 years	100 years
Plotting position (Weibull formula)	5197	7677	8789
Normal distribution	4761	5863	6252
Lognormal distribution	5014	7445	8566
Log-Pearson distribution	4920	6832	7607
Gumbel's distribution	5073	7600	8700

Now, by linear extrapolation of the line we get

50-year annual maximum flood = 7600 cumec;
100-year annual maximum flood = 8700 cumec.

Note: In this chapter, we have seen that the same set of annual maximum flood data at a particular river gauging station (given in Table 5.1) is analyzed in different methods and considering different distributions, Table 5.8 shows a comparison of results obtained by different methods (Fig. 5.7).

Confidence Limits of the Gumbel Distribution

The estimation of the magnitude of a random variable x_T for a particular return period T for Gumbel's extreme value distribution is shown in the previous section. However, the x_T value obtained in this way is uncertain due to the limited sample size. Hence, it is useful to compute a range of x_T, say x_1 and x_2, which is termed as confidence limit or confidence interval (CI). The CI is always associated with a probability measure, known as level of confidence (Chap. 6 for more discussion). Thus, the confidence interval can be defined as the limits of the estimated value of the variable x_T between which the actual value will lie with a probability of c, known as confidence level.

If x_1 and x_2 be the upper and lower bounds of the confidence interval, then

$$x_{1,2} = x_T \pm f(c) S_e \tag{5.16}$$

where x_T = estimated extreme value of the variable with the return period T; $f(c)$ = a function of confidence probability/level c, which is the standard normal variate Z value for the non-exceedance probability c; S_e = standard error = $b\frac{S_x}{\sqrt{N}}$; $b = \sqrt{1 + 1.3K + 1.1K^2}$ N = sample size; K = frequency factor for the Gumbel distribution; and S_x = Standard deviation of the sample.

The values of $f(c)$ can be read from a standard normal table.

Example 5.4.7

Consider the annual maximum flood data series given in Table 5.1, and estimate the 95 and 99% confidence interval for 100-year maximum annual flood. Use Gumbel's extreme value distribution.

Solution Using Gumbel's extreme value distribution, the 100-year maximum annual flood is already calculated in Example 5.4.6 and obtained as x_{100} = 8700 cumec.

The frequency factor K_{100} can be read from Table 5.7 for sample size $N = 40$ and obtained as $K_{100} = 3.55$

$$b = \sqrt{1 + 1.3K_{100} + 1.1K_{100}^2} = \sqrt{1 + (1.3 \times 3.55) + (1.1 \times 3.55^2)} = 4.413.$$

From Example 5.4.6, $S_x = 1427.193$, and hence the standard error can be calculated as

$$S_e = b\frac{S_x}{\sqrt{N}} = 4.413\frac{1427.193}{\sqrt{40}} = 995.8.$$

For 95% confidence interval, $f(c) = 1.96$ and hence,

$$x_{1,2} = x_T \pm f(c)S_e = 8700 \pm (1.96 \times 995.8)$$

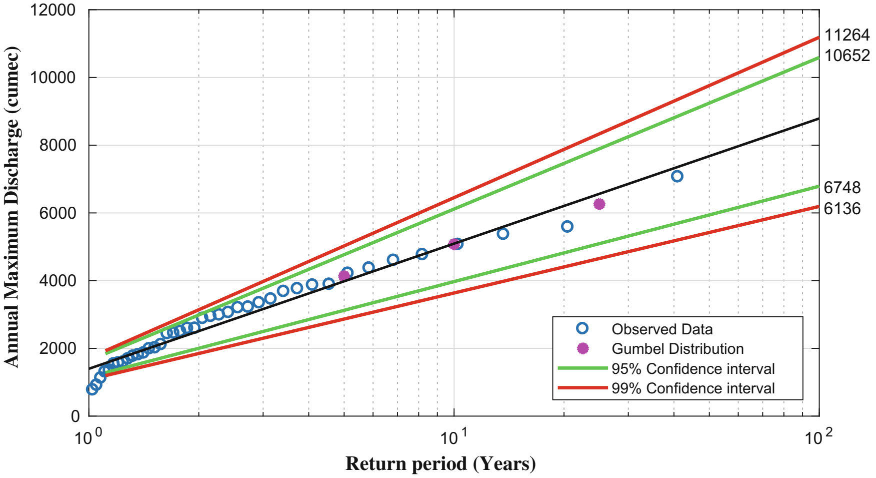


Fig. 5.8 Confidence intervals (95 and 99%) for annual maximum discharge obtained using Gumbel's extreme value distribution

x_1= 10652 cumec and x_2 = 6748 cumec. So, 95% confidence interval of the estimated value of 100-year maximum annual flood is 6748 cumec and 10652 cumec.

For 99% confidence interval, $f(c) = 2.575$ and hence,

$$x_{1,2} = x_T \pm f(c)S_e \; = \; 8700 \pm \; (2.575 \times 995.8)$$

$x_1 \; = 11264$ cumec and $x_2 \; = 6136$ cumec. So, the calculated value of 100-year maximum annual flood 8700 cumec has a 99% confidence probability of lying between 11264 cumec and 6136 cumec.

In Fig. 5.8, the black points indicate the values of annual maximum flood magnitudes for different return periods and a straight line is fitted to these points, shown by black thick line. The 95% and 99% confidence intervals for various values of return period are also shown. It can be observed that the range of confidence interval increases with the increase in confidence level. It can also be noted that the range of confidence interval increases as T increases.

Frequency Analysis for Zero-Inflated Data

Zero-inflated data contains many zero values apart from other values over a continuous range. For example, daily rainfall data or peak flow values from an ephemeral river may contain significant number of zero values and positive values over the range of 0 to ∞. Presence of significant number of zeros in a set of data needs some special treatment, especially if logarithmic transformation is required. Treatment of zeros can be done by any of the following ways

(i) Addition of small constant to all of the observations. Hence, logarithmic transformation becomes feasible.
(ii) Analysis of the nonzero values only and conditioned probabilistic assessment is carried out. Among the nonzero values is the condition.
(iii) Using the total probability theorem to handle zero values along with nonzero values. This method is more accurate and discussed here.

The distribution of zero-inflated data will have a probability mass at $x = 0$ and continuous density function over $x > 0$. Such a distribution is known as mixed distribution (refer Chap. 4, Sect. 4.3). Application of total probability theorem is as follows.
Range of the random variable is grouped into two parts, $x = 0$ and $x \neq 0$. Next, by Theorem of Total Probability, we can write

$$P(X \geq x) = P(X \geq x|\ x = 0)\, P(x = 0) + P(X \geq x|\ x \neq 0)\, P(x \neq 0) \quad (5.17)$$

Now, $P(X \geq x|\ x = 0) = 0$, and hence,

$$P(X \geq x) = P(X \geq x|\ x \neq 0)\, P(x \neq 0) \quad (5.18)$$

In Eq. 5.18, $P(x \neq 0)$ can be determined based on fraction of nonzero values (k) in the data. Estimation of $P(X \geq x|\ x \neq 0)$ needs analysis of nonzero values only with sample size equal to number of nonzero values.

Suppose *pdf* and *CDF* of X are given by $f_x(x)$ and $F_x(x)$. Also, consider a random variable X_{nz} which takes all nonzero values of X with *pdf* and *CDF* as $g_X(x)$ and $G_X(x)$. So, Eq. 5.18 can be rewritten as

$$\begin{aligned} 1 - F_X(x) &= k(1 - G_X(x)) \\ \Rightarrow F_X(x) &= (1 - k) + kG_X(x) \end{aligned} \quad (5.19)$$

Example 5.4.8
In a set of 100 records of daily rainfall data, 30 values are found to be zero. The rest of the data have a mean of 50 mm and standard deviation of 12.5 mm. Consider the nonzero daily rainfall values to follow a lognormal distribution.

(a) Estimate the probability of daily rainfall exceeding 60 mm.
(b) Estimate the magnitude of daily rainfall with an exceedance probability of 0.01.

Solution

(a) Here, we have to find out $P(X > 60) = 1 - P(X \leq 60) = 1 - F_x(60)$
From Eq. 5.19, we get $F_x(60) = (1 - k) + kG_X(60)$

Here, fraction of nonzero values $k = 70/100 = 0.7$
Now for the nonzero values, $\overline{x}_{nz} = 50$, $S_{nz} = 12.5$, and hence, $CV_{nz} = 12.5/50 = 0.25$
$\overline{x}_{nz}$, S_{nz}, and CV_{nz} indicate mean, standard deviation, and coefficient of variation of nonzero values, respectively.
As X_{nz} follows lognormal distribution, let us consider another variable $Y = \log(X_{nz})$, which follows normal distribution with mean μ_y and σ_Y^2

$$\mu_Y = 0.5 \ln \left[\frac{\overline{x}_{nz}^2}{1 + CV_{nz}^2} \right] = 0.5 \ln \left[\frac{50^2}{1 + 0.25^2} \right] = 3.882$$

$$\sigma_Y^2 = \ln \left[1 + CV^2 \right] = \ln \left[1 + 0.25^2 \right] = 0.0606 \Rightarrow \sigma_Y = 0.246$$

Now, $G_X(60) = P(X_{nz} \leq 60) = P(\ln X_{nz} \leq \ln 60) = P(Y \leq 4.094)$

$$\Rightarrow P\left(\frac{Y - \mu_Y}{\sigma_Y} \leq \frac{4.094 - 3.882}{0.246} \right) = P(Z \leq 0.862) = 0.806$$

Hence,

$$P(X > 60) = 1 - P(X \leq 60) = 1 - F_X(60) = k(1 - G_X(60)) = 0.7(1 - 0.806) = 0.136$$

So, the probability of daily rainfall exceeding 60 mm is 0.136.

(b) For daily rainfall with exceedance probability 0.01, $P(X > x) = 0.01$

$$F_X(x) = 1 - P(X > x) = 0.99$$

From Eq. 5.19,

$$G_X(x) = (F_X(x) - 1 + k)/k = (0.99 - 1 + 0.7)/0.7 = 0.9857$$
$$\text{Further, } G_X(x) = P(X_{nz} \leq x) = P(\ln X_{nz} \leq \ln x) = P(Y \leq \ln x)$$

Hence,

$$\Rightarrow P\left(\frac{Y - \mu_Y}{\sigma_Y} \leq \frac{\ln x - 3.882}{0.246} \right) = 0.9857$$
$$\Rightarrow P\left(Z \leq \frac{\ln x - 3.882}{0.246} \right) = 0.9857$$
$$\Rightarrow \frac{\ln x - 3.882}{0.246} = 2.189$$
$$\Rightarrow \ln x = 4.42$$

$$\Rightarrow x = 83.14$$

Hence, the magnitude of daily rainfall with exceedance probability of 0.01 is **83.14 mm**.

5.5 Risk and Reliability in Hydrologic Design

Hydrologic design is always subject to risk due to uncertainty present in the available record. Risk (R) is defined as the probability of occurrence of an event $[P\left(X > x_T\right)]$ at least once over a period of n successive years. Thus,

$$\begin{aligned} R &= P(\text{occurrence of the event } X > x_T \text{ at least once over a period of } n \text{ successive years}) \\ &= 1 - P(\text{non-occurrence of the event } X > x_T \text{ in } n \text{ successive years}) \\ &= 1 - (1 - p)^n = 1 - \left(1 - \frac{1}{T}\right)^n \end{aligned} \tag{5.20}$$

where $P = P(X > x_T)$, return period $T = \frac{1}{p}$, and n is the design life of the structure.

On the other hand, reliability (R_e) is opposite of risk. It may be defined as the probability that no extreme event $X > x_T$ will occur during the lifetime of the structure. So, it is given by

$$R_e = P(\text{non-occurrence of the event } X > x_T \text{ in } n \text{ successive years})$$

$$R_e = \left(1 - \frac{1}{T}\right)^n = 1 - R \tag{5.21}$$

In design practice, a *factor of safety* (F_s) is also used to take care of uncertainties arising from various sources. F_s is expressed as

$$F_s = \frac{P_a}{P_e} \tag{5.22}$$

where P_a is the actual value of the parameter adopted in design and P_e is the estimated value of the parameter obtained from hydrological analysis. Sometimes, the difference $(P_a - P_e)$ is termed as *Safety margin*.

Example 5.5.1
A flood embankment has an expected life of 20 years. (a) For an acceptable risk of 5% against the design flood, what design return period should be adopted? (b) If the above return period is adopted and the life of the structure is revised to be 50 years, what is the new risk value?

Solution Expected life of the flood embankment, $n = 20$ years

(a) Acceptable risk

$$R = 1 - \left(1 - \frac{1}{T}\right)^n = 0.05$$
$$\Rightarrow 1 - \left(1 - \frac{1}{T}\right)^{20} = 0.05$$
$$\Rightarrow T = 390.4 \approx 400 \text{ years}$$

(b) If the life of the embankment (n) becomes 50 years and adopted return period is 400 years, new value of risk is given by

$$R = 1 - \left(1 - \frac{1}{T}\right)^n = 1 - \left(1 - \frac{1}{400}\right)^{50} = 0.118$$

So, the new value of risk is 11.8%.

Example 5.5.2
A barrage is constructed for 75-year design life on a river with a 10% risk. Analysis of annual peak flow in the river gives a sample mean of 1000 cumec and standard deviation of 300 cumec. Estimate design flood of the barrage assuming peak flows follow a Gumbel's extreme value distribution. If *factor of safety* $F_s = 2$, then what will be the design flood?

Solution For the river,

Mean annual peak flow ($\overline{x}$) = 1000 cumec.
Standard deviation (S_x) = 300 cumec.

The design considers 10% risk for a design life of 75 years.
Hence,

$$R = 1 - \left(1 - \frac{1}{T}\right)^{75} = 0.1$$
$$\Rightarrow T = 712.34 \approx 720 \text{ years}$$

Using Gumbel's extreme value distribution method,

$$y_T = -\ln\left(\ln\frac{T}{T-1}\right) = -\ln\left(\ln\frac{720}{719}\right) = 6.578$$

$$K = \frac{y_T - 0.5772}{1.2825} = \frac{6.578 - 0.5772}{1.2825} = 4.679$$

So, the design flood $x_T = \overline{x} + KS_x = 1000 + (4.679 \times 300) = 2404$ cumec. Considering a factor of safety of 2, the design flood $= F_s \times x_T = (2 \times 2404) = 4808 \approx 5000$ cumec.

5.6 Concept of Uncertainty

Uncertainty in hydrology can be defined as a situation which involves imperfect and/or lack of information about any hydrological variable. Uncertainty is to be dealt with in various aspects of hydrologic design and analysis. There are several factors that cause uncertainty in hydrologic system, as stated below.

(i) **Uncertainty due to inherent randomness of any hydrological event**: Intrinsic dynamics of hydrologic processes are not known and perhaps could never be known with certainty. Inherent variation of different hydrologic variables is influenced by several physical, chemical, biological, and socioeconomic processes. As a consequence, uncertainty due to inherent randomness is very complex, unavoidable, and can never be eliminated. Spatio-temporal variation of hydrological events like flood, rainfall is significantly caused by the natural inherent uncertainty of the system.

(ii) **Uncertainty due to the model**: Assumptions are always necessary to model or design any complex system. Hydrologic phenomena are very complex, and often some simple assumptions are made to develop any model. These simplifications bring uncertainty into the developed model due to the lack of complete representation of physical processes in the real system. Model uncertainty can be reduced to some extent by improving such representations closest to the reality. For example, model uncertainty may be more in a simple linear rainfall–runoff model as compared to physically based rainfall–runoff model.

(iii) **Uncertainty due to model parameters**: Hydrological models consist of a few to several model parameters that are estimated during model calibration. Inability to accurately estimate model parameters due to lack of data and knowledge leads to parameter uncertainty. Apart from estimation, if some changes occur in operational conditions of a hydrologic system or hydraulic structure, it can also cause parameter uncertainty. This kind of uncertainty is also reducible to some extent.

(iv) **Uncertainty due to data**: Generally, a hydrologist has to work under unavoidable situation of data scarcity. Not only that, data uncertainty may arise due to measurement errors, data handling errors, non-homogeneous and inconsistent data. All these factors result in data uncertainty. Uncertainty due to data can be avoided by improving the data quality and quantity through improved data collection and data handling.

(v) **Operational uncertainty**: This kind of uncertainty is due to human errors during the execution phase of a design. It incorporates randomness in manufacturing, construction, and maintenance. All these factors lead to operational uncertainty. Good workmanship and quality control can be adopted to reduce such uncertainty.

5.6.1 Analysis of Uncertainty

Hydrologic models are always based on simpler approximations of the complex real system. These models accept different hydrological inputs, operate internally using some model parameters, and produce output. Both of these inputs and model parameters are stochastic in nature, i.e., associated with randomness. The focus of uncertainty analysis is to quantify uncertainty in the model outputs. Uncertainty analysis may provide two important results: firstly quantification of uncertainty associated with output and secondly relative contribution of each stochastic input variable to the overall uncertainty of the system output. The former result helps to quantify the confidence in the overall output of the model. The latter helps the investigator to identify the most sensitive input variable.

Uncertainty analysis has three components, namely qualitative uncertainty analysis, quantitative uncertainty analysis, and communication of uncertainty. Qualitative analysis identifies different uncertainties associated, and quantitative analysis measures effect of uncertainties of different variables on the system in quantitative terms. Finally, communication of uncertainty analysis, i.e., how the uncertainty from input variables and model parameters transfers to model outputs.

5.6.2 Measures of Uncertainty

Quantitative analysis of uncertainty needs to quantify the uncertainty associated with a random variable. Several methods are available to measure uncertainty, and some of them are listed below.

(i) In statistical analysis, uncertainty of a random variable can be expressed through the statistical parameters of the distribution, which describes the stochastic nature of that random variable. One common way to measure the uncertainty is to use different orders of statistical moments of the distribution. In particular, variance is the most commonly used measures of uncertainty. Since the variance is a measure of dispersion of a random variable (refer Chap. 2), increase in variance of data implies the increase in the associated uncertainty (Fig. 5.9).

(ii) Another measure of uncertainty of a random variable is to quantify it in terms of confidence interval. A confidence interval is a numerical range that would enclose the quantity of the variable with a specific level of confidence. Estimation of confidence interval is discussed in Chap. 6.

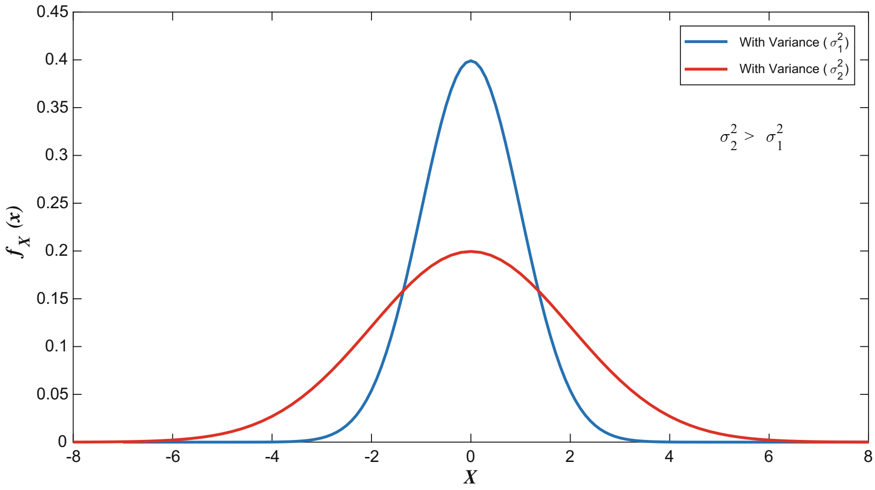


Fig. 5.9 Variance and uncertainty

(iii) Uncertainty is also represented non-parametrically in terms of different quartile values. When an ordered data set is divided into quarters, the division points are called sample quartiles. The different quartiles in an ordered data set are:

First Quartile (Q_1): It is a value of the data set such that one-fourth of the observations are less than this value.
Second Quartile (Q_2): It is a value of the data set such that half of the observations are less than this value. It is equivalent to the median.
Third Quartile (Q_3): It is a value of the data set such that three-fourth of the observations are less than this value.

Difference between first and third quartile is known as inter-quartile range (IQR). Often the quantiles are represented through a boxplot.

Boxplot: The information regarding the quartiles and the inter-quartile range in an ordered data set can be represented by a boxplot. The significant information depicted in a boxplot is:

- Upper whisker ($Q_3 + 1.5\,\mathrm{IQR}$)
- Third quartile (Q_3)
- Median or second quartile (Q_2)
- First quartile (Q_1)
- Lower whisker ($Q_1 - 1.5\,\mathrm{IQR}$).

During the construction of a boxplot, first, the range between Q_1 and Q_3 is represented by a rectangle with a line at Q_2. Then, the range between Q_1 and lower whisker, and Q_3 and upper whisker are connected by lines. Sometimes, 5th and 95th quartile values may also be used as lower and upper whiskers

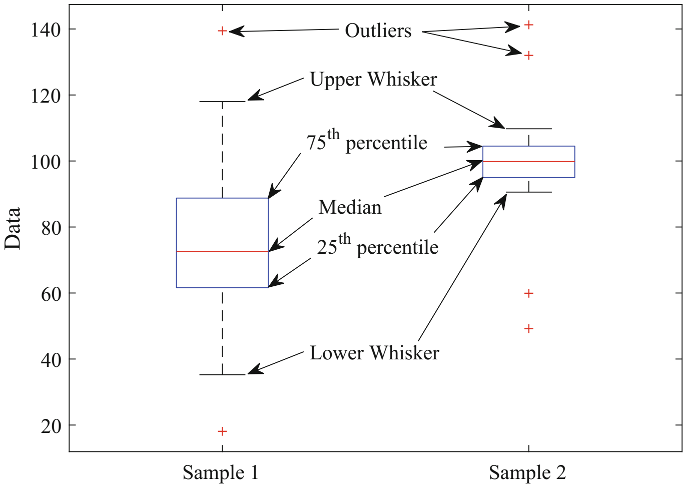


Fig. 5.10 A typical boxplot

respectively, for large data sets. A typical example of boxplot is shown in Fig. 5.10.

5.7 Reliability, Resilience, and Vulnerability of Hydrologic Time Series

A chronological sequence of values of a hydrologic variable, collected over a period of time, is termed as a hydrologic time series. Details of time series analysis are discussed elaborately in Chap. 9. Three properties of a typical hydrologic time series are discussed in this section that help to characterize the variable with respect to lower extreme events. Considering a threshold to delineate satisfactory and unsatisfactory states, these measures describe how likely a system remains in satisfactory state (reliability), how quickly it recovers from unsatisfactory state (resilience), and how severe the consequences of satisfactory state may be (vulnerability). A typical example could be the series of soil moisture and permanent wilting point (PWP) as the threshold to determine the satisfactory state, since plants cannot extract water from soil if the moisture falls below PWP. Let X_t $(t = 1, 2, \ldots, n)$ be the time series of a hydrologic variable having a data length n.

5.7.1 Reliability

Reliability (α) is defined by the probability that a system remains in a satisfactory state. It is expressed as:

$$\alpha = P(X_t \in S) \tag{5.23}$$

where S is the set of all satisfactory states. For a time series, α can be computed as follows

$$\alpha = \underset{n\to\infty}{Lt} \frac{1}{n} \sum_{t=1}^{n} Z_t \tag{5.24}$$

where $Z_t = 1$, if $X_t \in S$; $Z_t = 0$, if $X_t \in F$, and F is the set of all unsatisfactory states.

5.7.2 Resilience

Resilience (γ) is a measure that indicates how quickly the system can return to a satisfactory state after it has fallen in unsatisfactory state (below the threshold). This can be defined as the ratio between the probability of transition from the unsatisfactory to the satisfactory state to the probability of failure. Thus,

$$\gamma = \frac{P(X_t \in F,\ X_{t+1} \in S)}{P(X_t \in F)} \tag{5.25}$$

where the numerator $P(X_t \in F,\ X_{t+1} \in S)$ is probability of transition from the unsatisfactory to the satisfactory state (denoted as ρ). In the long run, the number of times the system transforms from the satisfactory to the unsatisfactory state and from the unsatisfactory to the satisfactory state will be same. Thus, it can be eventually expressed as $P(X_t \in F,\ X_{t+1} \in S) = P(X_t \in S,\ X_{t+1} \in F)$. From a time series, ρ can be computed as

$$\rho = \underset{n\to\infty}{Lt} \frac{1}{n} \sum_{t=1}^{n} W_t \tag{5.26}$$

where W_t is the event of transformation from the satisfactory to the unsatisfactory state (or vice versa) and $W_t = 1$, if $X_t \in S$, $X_{t+1} \in F$, and $W_t = 0$ otherwise. The denominator of Eq. 5.25 can be expressed as $P(X_t \in F) = 1 - P(X_t \in S)$. Again, $P(X_t \in S)$ is expressed as reliability α as explained before. Thus, Eq. 5.25 can be expressed as

$$\gamma = \frac{\rho}{1-\alpha} \tag{5.27}$$

5.7.3 Vulnerability

Vulnerability is a measure of severity of an event in unsatisfactory state, once it has occurred. It can be estimated in different ways. In the context of hydrologic time series analysis dealing with lower side extreme, one of the estimate could be

$$\upsilon = \frac{1}{k} \sum_{j \in F} x_j \tag{5.28}$$

where x_j is an observation that belongs to the unsatisfactory state and k is the number of times the unsatisfactory state occurs.

Example 5.7.1

For a particular location, daily soil moisture was recorded since January 1, 2017. From the data set, first 100 daily soil moisture data is provided in Table A.5 of Appendix A. If the permanent wilting point is 0.1, then estimate the reliability, resilience, and vulnerability of this time series of soil moisture data.

Solution Figure 5.11 shows the time series soil moisture data (θ) given in Table A.5. As PWP is given as 0.1, the daily soil moisture values falling below PWP ($\theta = 0.1$) are considered to be falling in 'unsatisfactory zone.'

For calculation of reliability of this data, we have to use Eq. 5.24, i.e., $\alpha = \underset{n\to\infty}{Lt} \frac{1}{n} \sum_{t=1}^{n} Z_t$

$$Z_t = \begin{cases} 1 & \text{if } X_t \in S (\text{Non-filled points above the PWP line in Fig. 5.11}) \\ 0 & \text{if } X_t \in F (\text{Filled points below the PWP line in Fig. 5.11}) \end{cases}$$

So, for each data points we can determine their Z_t values and hence $\sum_{t=1}^{n} Z_t$. Here, $n = 100$. We are considering $n = 100$ is large enough to assume $n \to \infty$. Thus, we obtain $\sum_{t=1}^{n} Z_t = 85$.

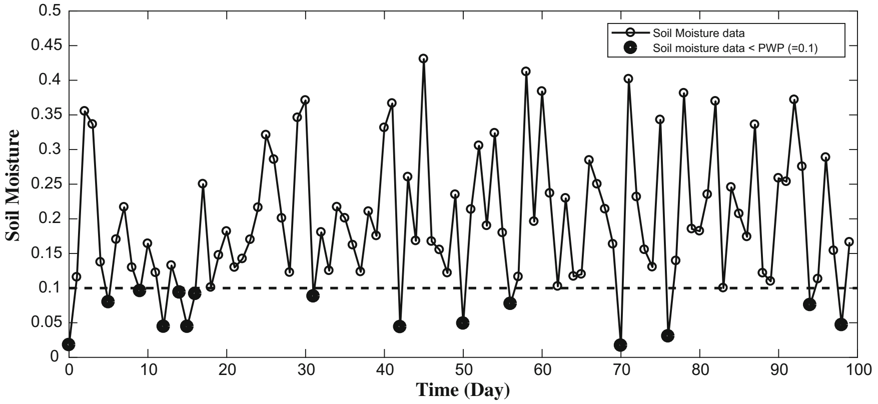


Fig. 5.11 Time series of soil moisture data (θ) for Example 5.7.1

Table 5.9 Calculation of soil moisture deficit from PWP

Day	Soil moisture data	Deficit from PWP	Day	Soil moisture data	Deficit from PWP
0	0.0179	0.0821	42	0.0439	0.0561
5	0.0798	0.0202	50	0.049	0.051
9	0.0959	0.0041	56	0.0774	0.0226
12	0.0444	0.0556	70	0.0171	0.0829
14	0.0938	0.0062	76	0.0305	0.0695
15	0.0443	0.0557	94	0.0757	0.0243
16	0.0917	0.0083	98	0.0468	0.0532
31	0.0882	0.0118	Average deficit		0.04024

So, reliability $\alpha = \frac{1}{n}\sum_{t=1}^{n} Z_t = \frac{85}{100} = 0.85$.

For calculation of resiliency of this data, we have to use Eq. 5.27, i.e., $\gamma = \frac{\rho}{1-\alpha}$

Where $\rho = \underset{n\to\infty}{Lt} \ \frac{1}{n}\sum_{t=1}^{n} W_t$

$$W_t = \begin{cases} 1 & \text{if } X_t \in S, X_{t+1} \in F \text{ or } X_t \in F, X_{t+1} \in S \\ W_t = 0 & \text{otherwise} \end{cases}$$

Similarly, for each data point we can determine their W_t values and hence $\sum_{t=1}^{n} W_t$. Here, also $n = 100$. Thus, we obtain $\sum_{t=1}^{n} Z_t = 24$ and hence, $\rho = \frac{1}{n}\sum_{t=1}^{n} W_t = \frac{24}{100} = 0.24$.

So, resilience $\gamma = \frac{\rho}{1-\alpha} = \frac{0.24}{1-0.85} = 1.6$.

As discussed earlier in Sect. 5.7.3, vulnerability is measured in terms of the mean soil moisture deficit caused during the failure events. In this case, there are total 15 failure events. The deficit of soil moisture from PWP is calculated for all of those 15 data points (Table 5.9). Average of these deficits, i.e., vulnerability, is obtained as 0.04.

5.8 MATLAB Examples

The frequency analysis of hydrological variable/events can be done in MATLAB using a number of built-in functions. Some of the function related to distribution of data is also discussed in Sect. 4.5. Apart from earlier discussed function, following function is useful for this chapter:

- `probplot(dist_name,y)`: This function can be used for plotting any data (`y` argument) over probability paper of distribution specified by its name (`dist_name` argument).

This section will provide examples for solving examples using MATLAB. A brief description of each command line is provided at the end of each line after % symbol.

Following sample script can be used for solving Examples 5.2.1, 5.4.2, 5.4.3, and 5.4.4.

Box 5.1 Sample MATLAB code for Example 5.2.1 and associated examples

```
clear all;close all;clc

%% Input
obs_flood=[7065, 3456, 4215, 2435, 3218, 4767, 5368, 3891, 2015,
    2498,...
  3345, 1987, 1689, 3200, 5067, 4369, 2589, 1306, 3761, 2450,...
  1569, 1862, 2592, 3059, 1595, 1768, 2987, 3679, 4597, 5582,...
  1356, 2944, 1541, 2111, 774, 911, 1123, 2884, 3868, 1812];


%% Flood Magnitude for given return period
required_return_periods=[10;50;100];
n=length(obs_flood);
% start logging output in a file
output_file=['output' filesep() 'code_1_result.txt'];
delete(output_file);diary(output_file);diary on;

%% Evaluation return period flood using Weibull formula
% Example 5.2.1
sorted_obs_flood=sort(obs_flood,'descend');
rank=1:n;
exceedence_prob=rank/(1+n);
return_period_sorted=1./exceedence_prob;

regress_coeff = [ones(n,1) log(return_period_sorted)']\
    sorted_obs_flood';
intercept = regress_coeff(1);
slope = regress_coeff(2);

%Evaluation of flood magnitude with given return periods
flood_return_period_10=slope*log(required_return_periods(1))+
    intercept;
flood_return_period_50=slope*log(required_return_periods(2))+
    intercept;
flood_return_period_100=slope*log(required_return_periods(3))+
    intercept;

%Evaluation of return period for given flood magnitude
flood_threshold=4000;
return_period_flood_4000=exp((flood_threshold-intercept)/slope);

% Display Results
disp('Using Weibull Formula')
disp(' The flood magnitude with return periods of')
fprintf('\t 10 years is %3.1f cumec.\n',flood_return_period_10);
fprintf('\t 50 years is %3.1f cumec.\n',flood_return_period_50);
fprintf('\t 100 years is %3.1f cumec.\n',flood_return_period_100);
fprintf(' The return period for flood magnitude of 4000 cumec\n')
fprintf(' is %1.0f years.\n\n',return_period_flood_4000);


%% Evaluation return period flood using normal Distribution
% Example 5.4.2
mean_obs_flood=mean(obs_flood);
```

```
std_obs_flood=std(obs_flood);

%Evaluation of flood magnitude with given return periods
K_10=norminv(1-1/required_return_periods(1));
flood_return_period_10=mean_obs_flood+K_10*std_obs_flood;
K_50=norminv(1-1/required_return_periods(2));
flood_return_period_50=mean_obs_flood+K_50*std_obs_flood;
K_100=norminv(1-1/required_return_periods(3));
flood_return_period_100=mean_obs_flood+K_100*std_obs_flood;


%Evaluation of return period for given flood magnitude
flood_threshold=4000;
Z_threshold=(flood_threshold-mean_obs_flood)/std_obs_flood;
return_period_flood_4000=1/(1-normcdf(Z_threshold));

% Display Results
disp('Using normal distribution')
disp(' The flood magnitude with return periods of')
fprintf('\t 10 years is %3.1f cumec.\n',flood_return_period_10);
fprintf('\t 50 years is %3.1f cumec.\n',flood_return_period_50);
fprintf('\t 100 years is %3.1f cumec.\n',flood_return_period_100);
fprintf(' The return period for flood magnitude of 4000 cumec\n')
fprintf(' is %1.0f years.\n\n',return_period_flood_4000);

%% Evaluation return period flood using lognormal Distribution
% Example 5.4.3
Y=log(obs_flood);
mean_Y=mean(Y);
std_Y=std(Y);

%Evaluation of flood magnitude with given return periods
K_10=norminv(1-1/required_return_periods(1));
flood_return_period_10=exp(mean_Y+K_10*std_Y);
K_50=norminv(1-1/required_return_periods(2));
flood_return_period_50=exp(mean_Y+K_50*std_Y);
K_100=norminv(1-1/required_return_periods(3));
flood_return_period_100=exp(mean_Y+K_100*std_Y);


%Evaluation of return period for given flood magnitude
flood_threshold=4000;
Z_threshold=(log(flood_threshold)-mean_Y)/std_Y;
return_period_flood_4000=1/(1-normcdf(Z_threshold));

% Display Results
disp('Using lognormal distribution')
disp(' The flood magnitude with return periods of')
fprintf('\t 10 years is %3.1f cumec.\n',flood_return_period_10);
fprintf('\t 50 years is %3.1f cumec.\n',flood_return_period_50);
fprintf('\t 100 years is %3.1f cumec.\n',flood_return_period_100);
fprintf(' The return period for flood magnitude of 4000 cumec\n')
fprintf(' is %1.0f years.\n\n',return_period_flood_4000);

%% Evaluation return period flood using log-Pearson Distribution
% Example 5.4.4
Y=log10(obs_flood);
mean_Y=mean(Y);
std_Y=std(Y);
coeff_skewness_Y=skewness(Y);

%Evaluation of flood magnitude with given return periods
```

```
K_10=1.245-(1.245-1.231)/(-0.1)*(skewness(Y)-round(skewness(Y),1));
flood_return_period_10=10^(mean_Y+K_10*std_Y);
K_50=1.89-(1.89-1.834)/(-0.1)*(skewness(Y)-round(skewness(Y),1));
flood_return_period_50=10^(mean_Y+K_50*std_Y);
K_100=2.104-(2.104-2.029)/(-0.1)*(skewness(Y)-round(skewness(Y),1));
flood_return_period_100=10^(mean_Y+K_100*std_Y);


%Evaluation of return period for given flood magnitude
flood_threshold=4000;
Z_threshold=(log10(flood_threshold)-mean_Y)/std_Y;
return_period_flood_4000=1/(1-normcdf(Z_threshold));

% Display Results
disp('Using log-Pearson distribution')
disp(' The flood magnitude with return periods of')
fprintf('\t 10 years is %3.1f cumec.\n',flood_return_period_10);
fprintf('\t 50 years is %3.1f cumec.\n',flood_return_period_50);
fprintf('\t 100 years is %3.1f cumec.\n',flood_return_period_100);
fprintf(' The return period for flood magnitude of 4000 cumec\n')
fprintf(' is %1.0f years.\n\n',return_period_flood_4000);
diary off;
```

The output of sample code provided in Box 5.1 is provided in Box 5.2. Barring inconsistency due to rounding off, the results match with the solution obtained in respective examples.

Box 5.2 Results for Box 5.1

```
Using Weibull Formula
 The flood magnitude with return periods of
   10 years is 5095.5 cumec.
   50 years is 7679.8 cumec.
   100 years is 8792.8 cumec.
 The return period for flood magnitude of 4000 cumec
 is 5 years.

Using normal distribution
 The flood magnitude with return periods of
   10 years is 4761.6 cumec.
   50 years is 5863.7 cumec.
   100 years is 6252.8 cumec.
 The return period for flood magnitude of 4000 cumec
 is 4 years.

Using lognormal distribution
 The flood magnitude with return periods of
   10 years is 5024.8 cumec.
   50 years is 7477.6 cumec.
   100 years is 8604.1 cumec.
 The return period for flood magnitude of 4000 cumec
 is 5 years.

Using log-Pearson distribution
 The flood magnitude with return periods of
   10 years is 4923.8 cumec.
   50 years is 6832.6 cumec.
```

```
  100 years is 7613.1 cumec.
The return period for flood magnitude of 4000 cumec
is 5 years.
```

Similarly, Example 5.4.8 can be solved by using sample code produced in Box 5.3.

Box 5.3 Sample MATLAB code for Example 5.4.8

```
clear all;close all;clc;

%% Inputs
zero_rainfall_mass=30/100;
mean_non_zero_rainfall=50;
std_non_zero_rainfall=12.5;

%% Probability of daily rainfall exceeding 60 mm
x=60;
k=1-zero_rainfall_mass;
CV=std_non_zero_rainfall/mean_non_zero_rainfall;
mean_Y=0.5*log(mean_non_zero_rainfall^2/(1+CV^2));
std_Y=sqrt(log(1+CV^2));

% if G_X is the probability of nonzero rainfall being less than 60
% see example 5.4.8
G_X=normcdf(log(x),mean_Y,std_Y);

prob_rainfall_more_than_60=k*(1-G_X);

%% Magnitude of daily rainfall with an exceedence probability of 0.01
exceedence_prob=0.01;
rainfall_cdf=1-exceedence_prob;

% if G_X is the probability of nonzero rainfall being less than x
% see example 5.4.8
G_X=(rainfall_cdf-1+k)/k;
rainfall_with_exceedence_prob=exp(norminv(G_X,mean_Y,std_Y));

%% Display Result
output_file=['output' filesep() 'code_2_result.txt'];
delete(output_file);diary(output_file);diary on;
fprintf('The probability of daily rainfall exceeding 60 mm is %1.2f.\n',...
    prob_rainfall_more_than_60)
fprintf('Magnitude of daily rainfall with an exceedence probability of 0.01 is %1.2f.\n',...
    rainfall_with_exceedence_prob)
diary off
```

The output of sample code provided in Box 5.3 is provided in Box 5.4. The result matches with the solution obtained in Example 5.4.8.

Box 5.4 Results for Box 5.3

```
The probability of daily rainfall exceeding 60 mm is 0.14.
Magnitude of daily rainfall with an exceedence probability of 0.01 is 83.16.
```

The Example 5.5.2 can be solved using the sample script given in Box 5.5.

Box 5.5 Sample MATLAB code for Example 5.5.2

```
clear all;close all;clc;

%% Inputs
mean_peak_flow=1000;
std_peak_flow=300;
design_life=75;
FOS=2; % Factor of safety

T=eval(solve(['1-(1-1/x)^' num2str(design_life) '=0.1']));
y_T=-log(log(T/(T-1)));
K=(Y_T-0.5772)/1.2825;
design_flood=mean_peak_flow+K*std_peak_flow;
design_flood_with_FOS=design_flood*FOS;

%% Display Result
output_file=['output' filesep() 'code_3_result.txt'];
delete(output_file);diary(output_file);diary on;
fprintf('The required design flood is %3.2f.\n',...
    design_flood_with_FOS)
diary off
```

The sample code presented in Box 5.5 calculated the design flood of barrage to be 4808 cumec which matches with the Example 5.5.2.

Exercise

5.1 If the return period of a hurricane is 500 years, find out the probability that no such hurricane will occur in next 10 years. Consider occurrence of such hurricanes follows Poisson distribution. **Ans:0.98**

5.2 The annual rainfall magnitudes at a rain gauge station for a period of 20 years are given below in the table

Year	Annual rainfall (cm)	Year	Annual rainfall (cm)
1975	120	1985	100
1976	85	1986	108
1977	67	1987	105
1978	95	1988	113
1979	108	1989	98
1980	92	1990	93
1981	98	1991	76
1982	87	1992	83
1983	79	1993	91
1984	86	1994	87

Determine the following

(a) The probability of occurrence of an annual rainfall more then 100 cm. **Ans:0.286**
(b) Dependable (80%) rainfall at this rain gauge station. **Ans:79.84 cm**.

5.3 The records of peak annual flow in a river are available for 25 years. Plot the graph of return period versus annual peak flow, and estimate the magnitude of peak flow for (a) 50 year and (b) 100 year return period. Use Weibull plotting position formula. **Ans: (a) 6991 cumec, (b) 7912 cumec**.

Year	Annual peak flow (cumec)	Year	Annual peak flow (cumec)
1960	4780	1973	989
1961	2674	1974	1238
1962	4432	1975	1984
1963	1267	1976	2879
1964	3268	1977	2276
1965	3789	1978	3256
1966	2348	1979	3674
1967	2879	1980	4126
1968	3459	1981	4329
1969	4423	1982	2345
1970	5123	1983	1678
1971	4213	1984	1198
1972	3367		

5.4 Use the annual peak flow data in Exercise 5.3, and find out the best-fits distribution for the data using probability paper among (a) normal distribution, (b) lognormal distribution, and (c) Gumbel's distribution.

5.5 From analysis of flood peaks in a river, the following information is obtained

(a) The flood peak data follows lognormal distribution.
(b) Flood peak of 450 cumec has a return period of 50 year.
(c) Flood peak of 600 cumec has a return period of 100 year.

Estimate the flood peak in the river with 1000-year return period. **Ans:1347 cumec**.

5.6 Repeat the Exercise 5.5 if the flood peak data follows Gumbel's extreme value distribution. **Ans:1096 cumec**

5.7 Maximum annual flood at a river gauging station is used for frequency analysis using 30-year historical data. The frequency analysis performed by Gumbel's method provides the following information.

Return period (years)	Max. annual flood (cumec)
50	1060
100	1200

(a) Determine the mean and standard deviation of sample data used for frequency analysis. **(Ans: mean = 385 cumec, std. deviation = 223 cumec)**

Sl. No.	Station	Sample size (years)	Mean annual flood (cumec)	Std. deviation of annual flood (cumec)
1	A	92	6437	2951
2	B	54	5627	3360

(b) Estimate the magnitude of flood with return period 500- year **(Ans: 1525 cumec)**.

5.8 Consider the following annual flood data at two river gauging stations

(a) Estimate the 100- and 1000-year floods for both the stations. Use the Gumbel method.
(b) Determine the 95% confidence interval for the predicted value.

Ans: (a) $Q_{100} = 16359 \pm 2554$ **cumec and** $Q_{1000} = 22023 \pm 3744$ **cumec and (b)** $Q_{100} = 17298 \pm 3885$ **cumec and** $Q_{1000} = 23935 \pm 5721$ **cumec**.

5.9 A structure is proposed to be built within the 50-year flood plain of the river. If the life of the industry is 25 years, what is the reliability that the structure will never face flood. **(Ans: 0.603)**

5.10 A bridge with 25 years expected life is designed for a flood magnitude of 100 years. (a) What is the risk involved in the design? (b) If only 10% risk is acceptable in the design, what return period should be adopted in the design? **(Ans: (a) 0.222 (b) 240 years)**.

5.11 Frequency analysis of flood data at a river gauging station is performed by log-Pearson type III distribution which yields the following information

Coefficient of skewness = 0.4

Return period (years)	Max. annual flood (cumec)
50	10600
100	13000

Estimate the magnitude of flood with return period of 1000 years *(Ans:23875 cumec)*.

5.12 The following table gives annual peak flood magnitudes in a river. Estimate the flood peaks with return period 10, 100, and 500 years using (a) Gumbel's extreme value distribution, (b) log-Pearson type III distribution, and (c) lognormal distribution

Year	Q (cumec)	Year	Q (cumec)	Year	Q (cumec)	Year	Q (cumec)
1950	1982	1965	1246	1980	2291	1995	1252
1951	1705	1966	2469	1981	3143	1996	983
1952	2277	1967	3256	1982	2619	1997	1339
1953	1331	1968	1860	1983	2268	1998	2721
1954	915	1969	1945	1984	2064	1999	2653
1955	1557	1970	2078	1985	1877	2000	2407
1956	1430	1971	2243	1986	1303	2001	2591
1957	583	1972	3171	1987	1141	2002	2347
1958	1325	1973	2381	1988	1642	2003	2512
1959	2200	1974	2670	1989	2016	2004	2005
1960	1736	1975	1894	1990	2265	2005	1920
1961	804	1976	1518	1991	2806	2006	1773
1962	2180	1977	1218	1992	2532	2007	1274
1963	1515	1978	966	1993	1996	2008	2466
1964	1903	1979	1484	1994	1540	2009	2387

(Ans: (a) Q_{10} = 2829 **cumec,** Q_{100} = 4066 **cumec,** Q_{500} = 4672 **cumec,** (b) Q_{10} = 2762 **cumec,** Q_{100} = 3351 **cumec,** Q_{500} = 3553 **cumec,** (c) Q_{10} = 2851 **cumec,** Q_{100} = 4142 **cumec,** Q_{500} = 5045 **cumec)**

5.13 The following table gives soil moisture (SM) data at a particular location. Consider PWP as 0.12 and evaluate reliability, resilience, and vulnerability of the data.

Day	SM	Day	SM	Day	SM	Day	SM
1	0.0816	11	0.4080	21	0.0717	31	0.2834
2	0.2253	12	0.3745	22	0.2253	32	0.2953
3	0.1944	13	0.1647	23	0.4149	33	0.1647
4	0.3370	14	0.2654	24	0.3370	34	0.1190
5	0.1208	15	0.1300	25	0.2500	35	0.0655
6	0.0954	16	0.2703	26	0.1423	36	0.0532
7	0.0562	17	0.3837	27	0.1258	37	0.0296
8	0.2382	18	0.3152	28	0.1228	38	0.2145
9	0.1949	19	0.1448	29	0.2948	39	0.1526
10	0.3500	20	0.1152	30	0.4024	40	0.1210

(Ans: Reliability = 0.775, resilience = 0.889, vulnerability = 0.044).

Chapter 6
Hypothesis Testing and Nonparametric Test

It is often required to make some inferences about some parameter of the population on the basis of available data. Such inferences are very important in hydrology and hydroclimatology where the available data is generally limited. This is done through hypothesis testing. However, hypothesis testing requires the knowledge of sampling distribution of different statistics and parameter estimation. Sampling distribution of mean and variance and two types of parameter estimation – point estimation and interval estimation – are discussed at the starting of this chapter. Next, the hypothesis testing is taken up. Different cases are discussed elaborately with illustrative examples. Later, a few statistical tests are discussed that deal with the goodness-of-fit of a probability distribution to the data using the knowledge of hypothesis testing. Some of the commonly used nonparametric tests are also explained along with appropriate examples in the field of hydrology and hydroclimatology.

6.1 Populations and Samples

The concept of sample and population is very important. A population is a complete set of items that share at least one attribute in common that is the subject of a statistical analysis, for example mean soil moisture content (SMC) over a field. As we can imagine, we can collect countably infinite soil samples to measure the SMC. The entire set of such measurements (data), which is infinite (over a range), forms the population. Practically, we may collect some samples and have as many as possible but finite number of SMC data. This finite number of data forms the sample. It may also be noted that the population need not be always infinite. Number of rainy days over some span of periods is an example of finite population.

A population is characterized by the probability distribution function of the associated random variable X. If a population is infinite, it is impossible to observe all the values, and even if the population is finite, it is impractical to observe all the values. Thereby it is necessary to use a sample, which is a part of a population. To obtain a reliable assessment of the population, it is very important for the sample to be representative of the entire population that is called random samples.

R. Maity, *Statistical Methods in Hydrology and Hydroclimatology*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-10-8779-0_6

6.2 Random Samples

Random samples can be defined as a set of observations $X_1, X_2, \ldots, X_n$ drawn from finite or infinite population in such a way that *each element has equal probability of being selected* or *there is no biasness to any subset*. This ensures that the sample represents the same statistical properties of the population. The reliability of conclusions drawn from a sample depends on whether the sample is properly chosen following the aforementioned criteria so as to properly represent the population.

Different sample statistics are computed from the samples (Chap. 3). However, variation of sample statistics from sample-to-sample is inevitable. This is true for any application field including hydrology and hydroclimatology. This sample-to-sample variation gives rise to sampling distribution of different statistics.

6.3 Sampling Distribution

Consider the same example of a random sample of n soil samples that has been collected for soil moisture estimation, and let $\overline{x}$ and S^2 be the calculated mean and variance from the sample. Now if we consider another random sample of same size n, it is almost unlikely that the $\overline{x}$ or S^2 will have same values as the first sample. The difference among these sample statistics may be attributed to many issues including chance of selecting the samples and experimental procedure. The variation is a very important aspect, and a sample statistic is computed from a random sample $(X_1, X_2, \ldots, X_n)$. Refer to Chap. 3 for details. A sample statistic itself is a random variable since it varies from sample to sample. The sample statistics summarizes the characteristics of the sample and estimate population parameters through statistical inference. In other words, the properties of the population are inferred from the properties of the sample. It requires the knowledge of probability distribution of a sample statistic.

The probability distribution of a sample statistic is called the *sampling distribution* of the statistic. Sampling distributions of two most commonly used statistics—mean and variance are discussed in this section.

Note: A sample statistic is a *random variable*, while a population parameter is a *fixed value*.

6.3.1 *Sampling Distribution of the Mean*

Let $f_x(x)$ be the probability distribution of the population from which we have drawn the samples of size n each. Then, it is natural to look for the probability distribution of the mean $(\overline{x})$, which is called the sampling distribution of the mean. The following theorems are important in this connection:

Theorem 1 *The mean of the sampling distribution of means, denoted by $\mu_{\overline{x}}$, is given by:*

$$E\left(\overline{X}\right) = \mu_{\overline{x}} = \mu \tag{6.1}$$

where μ is the mean of the population.

Theorem 2 *If the population is infinite, then the variance of the distribution, denoted by $\sigma_{\overline{x}}^2$, is given by:*

$$\sigma_{\overline{x}}^2 = \frac{\sigma^2}{n} \tag{6.2}$$

where σ^2 is the variance of the population and n is the sample size.

Theorem 3 *If the population is finite, then the variance of the distribution, denoted by $\sigma_{\overline{x}}^2$, is given by:*

$$\sigma_{\overline{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \tag{6.3}$$

where σ^2 is the variance of the population, N is the size of the population, and n is size of the sample. The factor $(N-n)\big/(N-1)$ is called the finite population correction factor, close to 1 (and can be omitted for most practical cases) unless the sample constitutes a substantial portion of the population.

Theorem 4 *If the population from which samples are taken is normally distributed with mean μ and variance σ^2, then the sampling distribution of mean is normally distributed with mean μ and variance $\sigma^2\big/n$. In this case, larger the sample size, closer we can expect $\overline{x}$ to be to the mean of the population. In this sense, we can say that the mean becomes more and more reliable as an estimate of μ as the sample size is increased.*

Theorem 5 *Suppose that the population from which the samples are taken has a probability distribution with mean μ and variance σ^2 that is not necessarily a normal distribution. Then, the standardized sample mean is given by:*

$$Z = \frac{\overline{x} - \mu}{\sigma\big/\sqrt{n}} \tag{6.4}$$

where Z is a random variable whose distribution function approaches that of the standard normal distribution as $n \to \infty$.

Theorem 6 *Aforementioned theorems assume that the variance is known. However, in case of unknown variance, the variance is to be evaluated from a sample of the population (S^2). Then, the standardized sample mean is given by:*

$$Z = \frac{\overline{x} - \mu}{S\big/\sqrt{n}} \tag{6.5}$$

6.3.2 Sampling Distribution of the Variance

Similar to mean, the variance will also vary from sample to sample that can be estimated through its sampling distribution. Following theorems can be considered for the sampling distribution of variance:

Theorem 7 *If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then $\left((n-1)\,S^2/\sigma^2\right)$ is a random variable that follows a chi-square distribution $\left(\chi^2\right)$ with degree of freedom $\upsilon = n - 1$.*

Theorem 8 *If S_1^2 and S_2^2 are the variances of two independent random samples of size n_1 and n_2, respectively, taken from two populations that follow normal distribution having the same variance, then $\left(S_1^2/S_2^2\right)$ follows F distribution with degrees of freedom $\upsilon_1 = n_1 - 1$ and $\upsilon_2 = n_2 - 1$.*

Caution: The procedures for making inferences on variance are not robust. It must be ensured that the underlying population follows normal distribution. For non-normal populations sampling distribution of variance (S^2) not only depends on the population variance (σ^2) but also on higher-order moments (e.g., μ_3, μ_4). Thus, for the samples drawn from non-normal population, aforementioned procedure of making inference on variance is not applicable.

6.4 Statistical Inference

Recall that a sample statistic is a *random variable*, while a population parameter is a *fixed value*. Statistical inference is the method of quantitative assessment of population parameter in a statistical sense based on the sample data set which can be considered to represent the entire population. For example, following questions need statistical inference to answer based on the sample data:

(i) Is the mean seasonal rainfall lies between 750 mm and 900 mm?
(ii) Is the mean streamflow at a gauging site is greater than another gauging site?
(iii) Is the wind speed more in season A than in season B?
(iv) Is the variation of soil moisture over a region lies within the limit of a specific range?

Thus, propositions about a population, using sample data (drawn from the population), are made through the statistical inference. In other words, the characteristics of the population are learnt through a statistical inference from a sample. Statistical inference mainly deals with parameter estimation and hypothesis testing. Parameter estimation is generally of two types—point estimation and interval estimation. All these methods are explained in the following sections.

Table 6.1 Examples of point estimates most commonly used in statistics

Serial No.	Statistical parameter ($\boldsymbol{\theta}$)	Estimator ($\hat{\theta}$)
1	Population mean (μ)	$\overline{x} = \frac{1}{n}\sum_i x_i$
2	Population variance (σ^2)	$S^2 = \frac{1}{n-1}\sum_i (x_i - \overline{x})^2$
3	Coefficient of skewness (γ)	$C_s = \frac{n}{(n-1)(n-2)}\frac{1}{S^3}\sum_i (x_i - \overline{x})^3$
4	Coefficient of kurtosis (K)	$k = \frac{n^2}{(n-1)(n-2)(n-3)}\frac{1}{S^4}\sum_i (x_i - \overline{x})^4$

6.4.1 *Point Estimation*

Point estimation can be defined as a statistic, which is a single value evaluated from the sample data. The statistic can be considered to be reasonably close to the population parameter (e.g., mean and variance) it is supposed to estimate. Let us consider a random variable X, such that $X \sim f_X(x;\theta)$ with θ as its parameter. Also, let $x_1, x_2, \ldots, x_n$ is a random sample drawn from this population. Then, we can estimate a statistic $\hat{\theta}$ (where hat signifies a sample-based estimate, or *sample estimate*) such that $\hat{\theta} = h\left(x_1, x_2 \ldots . x_n\right)$ which can be considered as an estimate of θ (population parameter). A statistic $\hat{\theta}$ can be an unbiased (if on an average, the value of sample estimate is equal to the parameter, i.e., $E\left(\hat{\theta}\right) = \theta$) or biased estimator ($E\left(\hat{\theta}\right) \neq \theta$). Some very common examples of point estimation are given in Table 6.1 (or Table 3.1 p. 65 in Chap. 3). There are different methods for parameter estimation like mean square error, method of moments, and maximum likelihood method as explained elaborately in Chap. 3.

6.4.2 *Interval Estimation*

Interval estimation provides a range for a statistic evaluated from the sample data. It is estimated such that the corresponding parameter of the population will lie in this interval with certain statistical confidence. Let us consider a random variable X, such that $X \sim f_x(x;\theta)$ and $x_1, x_{2,}, \ldots, x_n$ is a random sample. The probability that θ lies within an interval (L, U), also referred to as confidence interval, is given as,

$$P\left(L < \theta < U\right) = 1 - \alpha \tag{6.6}$$

where α is the probability that θ will **not** lie in the given interval, also known as the significance level. The statistical confidence level of the estimated interval is $100\left(1-\alpha\right)\%$. The interpretation is shown in Fig. 6.1. For $\alpha = 0.05$, i.e., 95%

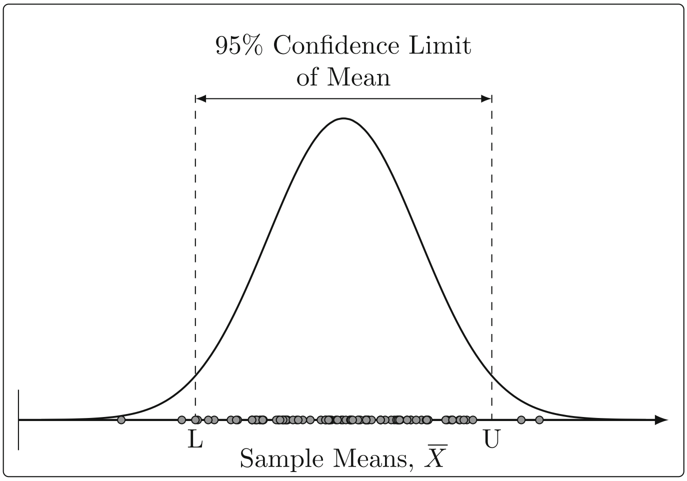


Fig. 6.1 Representation of 95% confidence interval of mean. The curve line indicates a probability density function for which 95% probability is within the upper and lower limits (dotted lines)

confidence interval of mean will capture 95% of means (considering many samples) within it in a statistical sense.

Single Sample Confidence Interval Estimations

In this section, we will discuss the confidence interval (CI) for statistical parameters of a population. Let us consider a random variable X, such that $X \sim N\left(\mu, \sigma^2\right)$ and $x_1, x_2, \ldots, x_n$ is a random sample (n is the sample size). The following cases provide the expression for *two-sided* $100\,(1-\alpha)\,\%$ CI of different statistical parameters.

- **Case 1**: $100\,(1-\alpha)\,\%$ CI of mean when variance (σ^2) is known,

$$\left(\overline{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

  where $\overline{x} = \frac{1}{n}\sum_i x_i$ and $P\left(x > Z_{\alpha/2}\right) = P\left(x < -Z_{\alpha/2}\right) = \alpha/2$.
- **Case 2**: $100\,(1-\alpha)\,\%$ CI of mean when variance is unknown,

$$\left(\overline{x} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \overline{x} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right)$$

  where $\overline{x} = \frac{1}{n}\sum_i x_i$, $S^2 = \frac{1}{n-1}\sum_i (x_i - \overline{x})^2$ and $P\left(x > t_{\alpha/2,n-1}\right) = P\left(x < -t_{\alpha/2,n-1}\right) = \alpha/2$ at $(n-1)$ degrees of freedom.

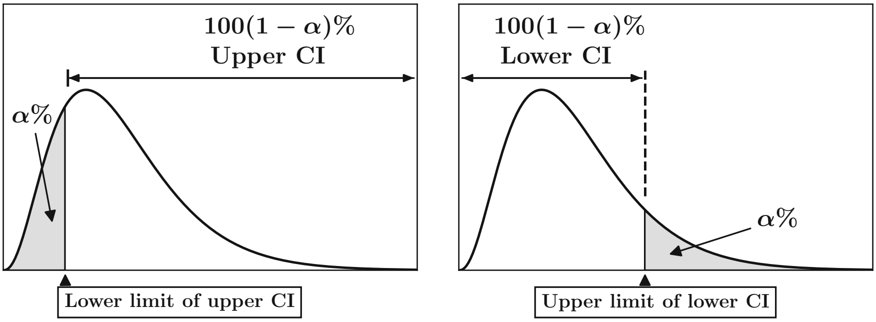


Fig. 6.2 Different *one-sided* confidence interval

- **Case 3**: $100\,(1-\alpha)\,\%$ CI of variance,

  $$\left(\frac{(n-1)\,S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)\,S^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

  where $S^2 = \frac{1}{n-1}\sum_i (x_i - \overline{x})^2$ and $P\left(x > \chi^2_{\alpha/2,n-1}\right) = P\left(x < \chi^2_{1-\alpha/2,n-1}\right) = \alpha/2$ at $(n-1)$ degrees of freedom.

  Note that in case of *one-sided* CI (contrasted against *two-sided* CI, mentioned before) as shown in Fig. 6.2, the upper (or lower) limit of $100\,(1-\alpha)\,\%$ lower (or upper) confidence interval for each of the above cases can be evaluated from the respective distributions, e.g., normal, t or chi-square distribution.
- **Case 4**: Upper limit of $100\,(1-\alpha)\,\%$ lower CI of mean is $\overline{x}+Z_\alpha\frac{\sigma}{\sqrt{n}}$, when variance (σ^2) is known. The corresponding *one-sided* lower CI is $\left(-\infty,\ \overline{x}+Z_\alpha\frac{\sigma}{\sqrt{n}}\right)$. Note that the lower limit of this CI is $-\infty$ since the lower bound of the sampling distribution of mean (Normal distribution) is $-\infty$.
- **Case 5**: Lower limit of $100\,(1-\alpha)\,\%$ upper CI of mean is $\overline{x}-Z_\alpha\frac{\sigma}{\sqrt{n}}$, when variance (σ^2) is known. The corresponding *one-sided* upper CI is $\left(\overline{x}-Z_\alpha\frac{\sigma}{\sqrt{n}},\ \infty\right)$. Note that the upper limit of this CI is ∞ (same reason as in case 4).
- **Case 6**: Similarly, in case of variance, upper limit of the $100\,(1-\alpha)\,\%$ *one-sided* CI is $\frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}$, and the lower bound of sampling distribution of variance (χ^2 distribution) is zero. Thus, the *one-sided* lower CI of variance is $\left(0,\ \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}\right)$.
- **Case 7**: Conversely, lower limit of $100\,(1-\alpha)\,\%$ *one-sided* upper CI of variance is $\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}}$, and the upper bound of sampling distribution of variance (χ^2 distribution) is ∞. Thus, the *one-sided* upper CI of variance is $\left(\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}},\ \infty\right)$.

Example 6.4.1

The rainfall data for the summer monsoon rainfall at a gauging station in the recent years is as follows,

Year	2010	2011	2012	2013	2014	2015	2016
Rainfall (mm)	443	456	503	480	536	600	545

Evaluate the 90% confidence interval of variance and upper limit of 95% confidence interval of mean.

Solution The sample mean and variance as calculated from the data are,

$$\overline{x} = 509 \text{ and } S^2 = 3058$$

The CI of variance can be evaluated using the chi-squared distribution as follows:

$$\left(\frac{(n-1)\,S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)\,S^2}{\chi^2_{1-\alpha/2,n-1}}\right) = \left(\frac{(7-1)\,3058}{\chi^2_{0.05,6}}, \frac{(7-1)\,3058}{\chi^2_{0.95,6}}\right) = (1457, 11222)$$

Therefore, the 90% confidence interval of variance for the given data is (1457, 11222).

The upper limit of 95% CI of mean can be evaluated using the t-distribution as follows,

$$\overline{x} + t_{\alpha,n-1}\frac{S}{\sqrt{n}} = 509 + t_{0.05,6}\frac{\sqrt{3058}}{\sqrt{7}} = 509 + (1.943) \times \frac{\sqrt{3058}}{\sqrt{7}} = 549.61$$

Therefore, the 95% upper confidence interval of mean for the given data is $(-\infty, 549.61)$.

Two-Sample Confidence Interval Estimations

In this section, we will discuss the confidence interval (CI) for statistical parameters involving two independent normal distributions. Let us consider two random variables X_1 and X_2, such that $X_1 \sim N\left(\mu_1, \sigma_1^2\right)$ and $X_2 \sim N\left(\mu_2, \sigma_2^2\right)$. Also, consider $x_{11}, x_{12}, \ldots, x_{1n_1}$ is a random sample of size n_1 and $x_{21}, x_{22}, \ldots, x_{2n_2}$ is a random sample of size n_2.

Case 1: $100\,(1-\alpha)\,\%$ CI of difference in mean when variances (σ_1^2 and σ_2^2) are known,

$$\left((\overline{x}_1 - \overline{x}_2) - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\overline{x}_1 - \overline{x}_2) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Case 2: $100\,(1-\alpha)\,\%$ CI of difference in mean when variances are unknown,

$$\left((\overline{x_1}-\overline{x_2})-t_{\alpha/2}\,S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\,(\overline{x_1}-\overline{x_2})+t_{\alpha/2}\,S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

where $S_p=\frac{S_1^2(n_1-1)+S_2^2(n_2-1)}{n_1+n_2-1}$

Case 3: $100\,(1-\alpha)\,\%$ CI of ratio of variance,

$$\left(\frac{1}{F_{\alpha/2}(n_1-1,n_2-1)}\frac{S_1^2}{S_2^2},\,\frac{S_1^2}{S_2^2}\,F_{\alpha/2}(n_2-1,n_1-1)\right)$$

where $P\left(x<F_{1-\alpha/2}(n_2-1,n_1-1)\right)=P\left(x>F_{\alpha/2}(n_2-1,n_1-1)\right)=\alpha/2$

Case 4: Upper limit of $100\,(1-\alpha)\,\%$ lower CI of difference in mean is $(\overline{x_1}-\overline{x_2})+Z_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$, when variances are known. The corresponding *one-sided* lower CI is $\left(-\infty,\,(\overline{x_1}-\overline{x_2})+Z_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$. Note that the lower limit of this CI is $-\infty$ since the lower bound of the sampling distribution of difference in mean (Normal distribution) is $-\infty$.

Case 5: Lower limit of $100\,(1-\alpha)\,\%$ upper CI of difference in mean is $(\overline{x_1}-\overline{x_2})-Z_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$, when variances are known. The corresponding *one-sided* upper CI is $\left((\overline{x_1}-\overline{x_2})-Z_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\,\infty\right)$. Note that the upper limit of this CI is ∞ (same reason as in case 4).

Case 6: Similarly, in case of ratio of variance, upper limit of the $100\,(1-\alpha)\,\%$ *one-sided* CI is $\frac{S_1^2}{S_2^2}\,F_\alpha(n_2-1,n_1-1)$, and the lower bound of sampling distribution of ratio of variance (F distribution) is zero. Thus, the *one-sided* lower CI of ratio of variance is $\left(0,\,\frac{S_1^2}{S_2^2}\,F_\alpha(n_2-1,n_1-2)\right)$.

Case 7: Conversely, lower limit of $100\,(1-\alpha)\,\%$ *one-sided* upper CI of ratio of variance is $\frac{1}{F_\alpha(n_1-1,n_2-2)}\frac{S_1^2}{S_2^2}$, and the upper bound of sampling distribution of ratio of variance (χ^2 distribution) is ∞. Thus, the *one-sided* upper CI of ratio of variance is $\left(\frac{1}{F_\alpha(n_1-1,n_2-2)}\frac{S_1^2}{S_2^2},\,\infty\right)$.

Example 6.4.2

The mean of maximum temperature (in °C) at locations A and B are observed to be 10 and 12, respectively. Evaluate the 95% confidence interval of difference in mean considering the data size at each location is 40. Variances are known to be 420 and 560 for location A and B, respectively.

Solution Given, $\overline{x_1} = 10$, $\sigma_1^2 = 420$ and $\overline{x_2} = 12$, $\sigma_2^2 = 560$

Sampling distribution of $\overline{x_1}$ follows $N\left(\overline{x_1}, \sigma_1 \big/ \sqrt{n_1}\right)$ and $\overline{x_2}$ follows $N\left(\overline{x_2}, \sigma_2 \big/ \sqrt{n_2}\right)$

Hence, the sampling distribution of the difference in mean $(\overline{x_1} - \overline{x_2})$ is a random variable that follows a normal distribution with mean $(\overline{x_1} - \overline{x_2})$ and standard deviation $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Thus, the CI of difference in mean can be evaluated using the standard normal distribution as follows,

$$\begin{aligned}
&\left((\overline{x_1} - \overline{x_2}) - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\overline{x_1} - \overline{x_2}) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \\
&= \left((10 - 12) - z_{0.025/2}\sqrt{\frac{420}{40} + \frac{560}{40}}, (10 - 12) + z_{0.025/2}\sqrt{\frac{420}{40} + \frac{560}{40}}\right) \\
&= (-11.85,\ 7.85)
\end{aligned}$$

Therefore, the 95% confidence interval of difference in mean is $(-11.85,\ 7.85)$.

6.4.3 *Hypothesis Testing*

Hypothesis testing is the process of accepting or rejecting an assumption regarding a population parameter, which may or may not be true. Often, we need to make decisions about population parameters on the basis of a sample. Such decisions are called statistical inferences. There are many problems in which we must decide whether a statement concerning a parameter is true or false; that is, we must test the hypothesis about a parameter. A procedure that enables us to accept or reject a hypothesis or to determine whether observed sample statistics differ significantly from population parameters are called test of hypothesis.

Null and Alternative Hypothesis

In attempting to reach a decision, it is always useful to make assumptions about the population involved. Initially, it is needed to decide the neutral or *by default* assumptions. This is denoted as null hypothesis (H_o). The opposite of it, that we want to test is assigned as alternative hypothesis (H_a). For instance, if we want to show that one irrigation technique is better than the other (H_a); initially, we hypothesize that both the techniques are equally effective (H_o). Similarly, if we want to decide whether rainfall at a location is greater than another location (H_a), we formulate the hypothesis that there is no difference in rainfall at two locations (H_o). Such hypothesis is often called null hypothesis denoted as H_o.

Table 6.2 Types of error in hypothesis testing

True fact	Decision making	
	Accept	Reject
Hypothesis is true	Correct	Type I error
Hypothesis is false	Type II error	Correct

Type I and Type II Errors

If null hypothesis H is true and not rejected or the hypothesis is false and rejected, the decisions in either case are correct. If hypothesis H is true but rejected, it is called Type I error . If Hypothesis H is false but not rejected, this is also an error and known as Type II error. These are shown in Table 6.2.

The probability of committing Type I error when the hypothesis is true is designated by α also known as level of significance. The probability of committing Type II error when the hypothesis is false is designated by β. Our major aim is to minimize the error which is generally achieved by fixing the value of α and minimizing β as far as possible.

Tests of Hypotheses

To approach the problem of hypotheses testing systematically, it will help to proceed as outlined in the following steps:

(i) Formulation of null hypotheses and appropriate alternative hypotheses which can be accepted when the null hypotheses are rejected.
(ii) Specification of the probability of Type I error or significance level, designated by α.
(iii) Based on the sampling distribution of an appropriate statistic, a criterion for testing the null hypothesis against the alternative is constructed.
(iv) From the data, the value of the test statistic on which the decision is to be based is evaluated.
(v) The final decision is whether to *reject* the null hypotheses or whether to *fail to reject* it.

Note that it is generally not concluded that the null hypothesis is accepted, instead it states whether one can or cannot reject the null hypothesis. This decision is based on the value of the test statistic and the significance level. The significance level decides the critical zone. If the test statistics fall within the critical zone, the null hypothesis is rejected; otherwise, it cannot be rejected. Explanation of critical zone is as follows.

There are two types of test—namely *one-sided* and *two-sided* test. Considering α as the significance level, the critical zone or rejection zone for one-sided test is either $\left[lb, x_{1-\alpha}\right]$ or $[x_\alpha, ub]$ based on the hypothesis to be tested. The zone $\left[lb, x_{1-\alpha}\right]$ is for '*greater than*' and $[x_\alpha, ub]$ is for '*less than*'. In these limits, lb and ub are the

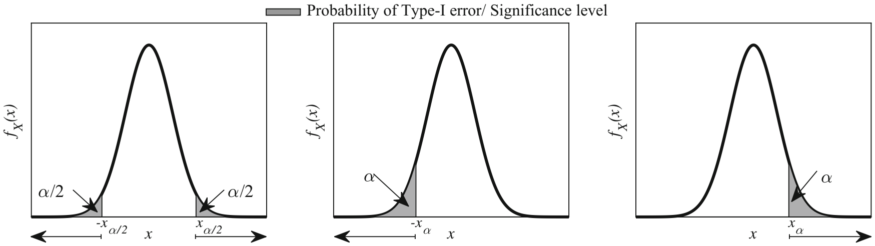


Fig. 6.3 Pictorial representation of rejection zone or critical zone for the two-sided (left panel) and one-sided (left sided–middle panel and right sided–right panel) test in hypothesis testing

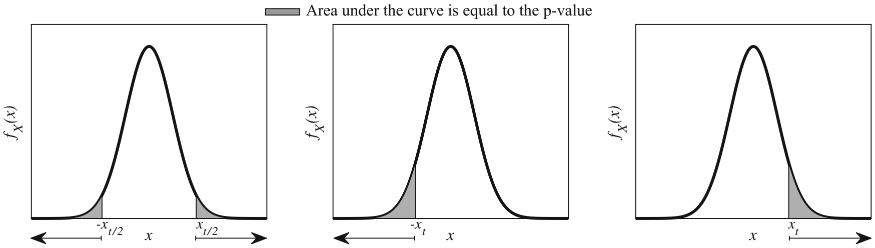


Fig. 6.4 Pictorial representation of p-value with respect to the test statistic (x_t) for two-sided (left panel) or one-sided (left sided—middle panel and right sided—right panel) tests. In case of two-sided, summation of the two regions is the p-value. Note: If the distribution is asymmetric (e.g., χ^2 or F distribution), the values of two limits of the shaded zone will not be same with opposite sign

lower and upper bound of the sampling distribution, respectively, $x_{1-\alpha}$ and x_α are the values such that $P(X \geq x_{1-\alpha}) = (1-\alpha)$ and $P(X \geq x_\alpha) = \alpha$. For a symmetrical sampling distribution, such as normal distribution, critical zone or rejection zone for *one-sided* test is either $(-\infty, -x_\alpha]$ or $[x_\alpha, \infty)$ based on the hypothesis to be tested, i.e., $(-\infty, -x_\alpha]$ is for '*greater than*' and $[x_\alpha, \infty)$ is for '*less than*'.

For the *two-sided* test, the critical zones are $\left[lb, x_{1-\alpha/2}\right]$ and $\left[x_{\alpha/2}, ub\right]$ with the same notations explained before. For a symmetrical sampling distribution, such as normal distribution, critical zones for *two-sided* test are $\left(-\infty, -x_{\frac{\alpha}{2}}\right]$ and $\left[x_{\frac{\alpha}{2}}, \infty\right)$. The representations of critical zones are shown in Fig. 6.3. In all the cases, the null hypothesis is rejected if the test statistic lies in the critical zone.

***p*-value**:

In each of these cases while carrying out the hypothesis test, a p-value can be evaluated. The p-value is the probability of obtaining a value of the test statistic that is as extreme as or more extreme than the value actually observed. Figure 6.4 explains the concept of p-value with respect to the test statistic (x) for *one-sided* or *two-sided* tests.

The following section explains the approach for selection of an appropriate test statistics and the rejection region for different cases.

Single Sample Test

In this section, we will discuss the criterions for hypothesis testing of statistical parameters for a normally distributed population. Let us consider a random variable X, such that $X \sim N\left(\mu, \sigma^2\right)$ and $x_1, x_2, \ldots x_n$ are a random sample of size n.

Case 1: Hypothesis concerning one mean when variance is known

Testing of the null hypothesis where a population mean equals/is greater than/is lesser than a specified constant value with suitable *one-sided* or *two-sided* test when the variance of the population is known.

$$\text{Test Statistic: } z = \frac{\overline{X} - \mu_0}{\sigma\big/\sqrt{n}} \sim N\,(0, 1) \tag{6.7}$$

where $\overline{X}$ is the sample mean and μ_o is a particular value of mean for which the hypothesis is to be tested.

The rejection criterion for three different cases is shown as follows,

H_o	H_a	Rejection Region
$\mu = \mu_o$	$\mu \neq \mu_o$	$\lvert z\rvert > Z_{\alpha/2}$
$\mu \geq \mu_o$	$\mu < \mu_o$	$z < -Z_\alpha$
$\mu \leq \mu_o$	$\mu > \mu_o$	$z > Z_\alpha$

Example 6.4.3

50 years of annual record is used to compute the mean annual rainfall at a gauging station. The mean is equal to 1460 mm. Is the population mean (μ) significantly different from 1500 mm at a level of significance of 0.05? Assume the population standard deviation as 150 mm.

Solution *Null hypothesis* $H_o : \mu = 1500\,\text{mm}$

Alternative hypothesis $H_a : \mu \neq 1500\,\text{mm}$

Level of significance: $\alpha = 0.05$ (given)

As the standard deviation of the population is given, the z statistics can be used.

$$z = \frac{\overline{x} - \mu}{\sigma\big/\sqrt{n}} = \frac{1460 - 1500}{150\Big/\sqrt{50}} = -1.88$$

Based on the alternative hypothesis, it is a two-sided test, thereby at 0.05 significance level $Z_{\alpha/2} = \pm 1.96\,(P\left(z > z_{\alpha/2}\right) = 0.025)$. Thus, the critical zone is $(-\infty, -1.96]$ and $[1.96, \infty)$.

Since the value of the test statistic does not lie in the critical zone, the null hypothesis cannot be rejected at a level of significance 0.05.

Therefore, the mean annual rainfall at the gauging station may be considered to be equal to 1500 mm at a level of significance of 0.05.

Example 6.4.4
It is found from the long-term historical data that the mean wind speed of a region is 51.35 km/h and standard deviation is 11 km/h. It is required to test whether the mean has increased or not. To test this, a sample of 80 stations in that region is tested and it is found that the mean wind speed is 54.47 km/h.

(a) Can we support the claim at a 0.01 level of significance?
(b) What is the p-value of the test?

Solution According to the example, the null and alternative hypothesis can be formulated as follows:

Null hypothesis $H_o : \mu \leq 51.35$ km/h
Alternative hypothesis $H_a : \mu > 51.35\ \text{km/h}$
Level of significance: $\alpha = 0.01(\text{given})$

As the standard deviation of the population is same as that obtained from historical data, the z statistics can be used.

$$z = \frac{\overline{x} - \mu}{\sigma \big/ \sqrt{n}} = \frac{54.47 - 51.35}{11 \Big/ \sqrt{80}} = 2.537$$

(a) Based on the alternative hypothesis, it is a one-sided test, thereby at 0.01 significance level $Z_\alpha = 2.325$. The critical zone is $[2.325, \infty)$.
Since the value of the test statistic lies in the critical zone, the null hypothesis must be rejected at a level of significance 0.01.
Therefore, it can be concluded that the wind speed is increased at a significance level of 0.01.
(b) The p-value of the test is $P(Z \geq 2.537) = 0.0056$, which is the probability that the mean wind speed equal to or more than 54.47 km/h would occur by chance if H_o is true.

Case 2: Hypothesis concerning one mean when variance is unknown

This case is same as case 1, but the variance of the population is unknown. In such case, the variance is calculated from the sample and the test statistic is modified as:

$$\text{Test statistic: } t = \frac{\overline{x} - \mu_0}{S \big/ \sqrt{n}} \tag{6.8}$$

where $\overline{x}$ is the sample mean, S is the sample variance, and μ_o is a particular value of mean for which the hypothesis is to be tested. Note that the test statistic follows student's t distribution with degrees of freedom $n - 1$ instead of standard normal distribution as in case 1.

The rejection criterion for three different cases is shown as follows:

H_o	H_a	Rejection Region
$\mu = \mu_o$	$\mu \neq \mu_o$	$\lvert t \rvert > t_{\alpha/2}(n-1)$
$\mu \geq \mu_o$	$\mu < \mu_o$	$t < -t_{\alpha/2}(n-1)$
$\mu \leq \mu_o$	$\mu > \mu_o$	$t > t_{\alpha/2}(n-1)$

Note: When the variance is unknown but the sample size is large enough ($n > 30$), then the test statistic approaches z. Thus, z statistic (Eq. 6.7) may also be used in such cases.

Example 6.4.5
The rainfall data for monsoon period at a gauging station is as follows:

Year	2000	2001	2002	2003	2004	2005	2006
Rainfall (mm)	543	496	523	450	576	590	505

Test the null hypothesis that the mean is greater than 570 mm at a confidence level of 95%. Also, evaluate the p-value.

Solution *Null hypothesis* $H_0 : \mu > 570$ mm
Alternative hypothesis $H_a : \mu \leq 570$ mm
The confidence level is 95%. Thus, *level of significance* $\alpha = 1 - 95/100 = 0.05$ (given)

As the standard deviation is to be calculated from the sample data, the t statistics is used.

Mean of the sample is $\overline{x} = 526.14$ mm and standard deviation of the sample is $S = 48.32$ mm

$$t = \frac{\overline{x} - \mu}{S/\sqrt{n}} = \frac{526.14 - 570}{48.32\big/\sqrt{7}} = -2.401$$

Based on the alternative hypothesis, it is a one-sided test, thereby at 0.05 significance level $t_\alpha(n-1) = -1.943$. The critical zone is $(-\infty, -1.943]$.

Since the value of the test statistic lies in the critical zone, the null hypothesis must be rejected at significance level 0.05.

The p-value of the test is $P\,(t < -2.401) = 0.027$.

Case 3: Hypothesis concerning one variance

This case considers the test of the hypothesis where a population variance equals/is greater than/is lesser than a specified constant value with suitable *one-sided* or *two-sided* test.

$$\text{Test statistic: } \chi^2 = \frac{(n-1)\,S^2}{\sigma_o^2} \tag{6.9}$$

where S is the sample variance and σ_o is a particular value of variance for which the hypothesis is to be tested. The test statistic follows chi-square distribution with $n-1$ degrees of freedom. The rejection criterion for three different cases is shown as follows:

H_o	H_a	**Rejection Region**
$\sigma^2 = \sigma_o^2$	$\sigma^2 \neq \sigma_o^2$	$\chi^2 > \chi^2_{\alpha/2}(n-1)$ or $\chi^2 < \chi^2_{1-\alpha/2,(n-1)}$
$\sigma^2 \geq \sigma_o^2$	$\sigma^2 < \sigma_o^2$	$\chi^2 < \chi^2_{1-\alpha}(n-1)$
$\sigma^2 \leq \sigma_o^2$	$\sigma^2 > \sigma_o^2$	$\chi^2 > \chi^2_{\alpha}(n-1)$

Example 6.4.6

Test the claim that the standard deviation of the streamflow at a gauging station is 220 cumec at the significance level of 0.01. The mean and standard deviation, calculated from a sample of size 16, are 8652 cumec and 200 cumec, respectively.

Solution *Null hypothesis* $H_o : \sigma^2 = 220^2$

Alternative hypothesis $H_a : \sigma^2 \neq 220^2$

Level of significance: $\alpha = 0.01(\text{given})$

In this case, the χ^2 statistics can be used.

$$\chi^2 = \frac{(n-1)\,S^2}{\sigma_o^2} = \frac{(16-1)\,200^2}{220^2} = 12.396$$

Based on the alternative hypothesis, it is a two-sided test, thereby at 0.01 significance level, $\chi^2_{\alpha/2}(n-1) = \chi^2_{0.005}(15) = 32.801$ and $\chi^2_{1-\alpha/2}(n-1) = \chi^2_{0.995}(15) = 4.601$. Thus, the critical zone is $(0, 4.601]$ and $[32.801, \infty)$.

Since the value of the test statistic does not lie in the critical zone, the null hypothesis cannot be rejected at a level of significance 0.01.

Therefore, it can be concluded that at a significance level of 0.01, the claim cannot be supported.

Two Sample Test

In this section, we will discuss about hypothesis testing involving two independent random samples that are drawn from normally distributed population. Let us consider two random variables X_1 and X_2, such that $X_1 \sim N\left(\mu_1, \sigma_1^2\right)$ and $X_2 \sim N\left(\mu_2, \sigma_2^2\right)$. Also, consider $x_{11}, x_{12}, \cdots, x_{1n_1}$ is a random sample of size n_1 for first random variable and $x_{21}, x_{22}, \cdots, x_{2n_2}$ is a random sample of size n_2 for second random variable.

Case 1: Hypothesis concerning two means when the variances are known

There are many statistical problems in which a decision is to be made about the comparison between the means of two or more samples when the population variance is known. In such cases, the test statistics are defined as

$$Z = \frac{(\overline{x_1} - \overline{x_2}) - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{6.10}$$

where $\overline{x_1}$ and $\overline{x_2}$ are the sample means and δ is the difference between the means for which the hypothesis is to be tested. The Z statistic is considered to follow standard normal distribution.

The rejection criterion for three different cases is shown as follows:

H_o	H_a	Rejection Region
$\mu_1 - \mu_2 = \delta$	$\mu_1 - \mu_2 \neq \delta$	$\lvert z\rvert > Z_{\alpha/2}$
$\mu_1 - \mu_2 \geq \delta$	$\mu_1 - \mu_2 < \delta$	$z < -Z_\alpha$
$\mu_1 - \mu_2 \leq \delta$	$\mu_1 - \mu_2 > \delta$	$z > Z_\alpha$

Example 6.4.7

Test the claim that the mean rate of evapotranspiration at station 1 is greater than that of station 2 by a magnitude of 0.5 mm/day. If the mean and standard deviation at the two stations are given as $\overline{x_1} = 4.59\,\text{mm/day}$, $\sigma_1 = 2.2\,\text{mm/day}$, $\overline{x_2} = 4.23\,\text{mm/day}$, and $\sigma_2 = 2.56\,\text{mm/day}$. The sample size for both the stations is 50; consider a significance level of 0.05.

Solution Let μ_1 and μ_2 are the mean values of evapotranspiration at stations 1 and 2, respectively.

Null hypothesis $H_o : \mu_1 - \mu_2 \leq 0.5$
Alternative hypothesis $H_a : \mu_1 - \mu_2 > 0.5$
Level of significance: $\alpha = 0.05$ (given)

As the standard deviation of the population is known, we can use the z statistics.

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(4.59 - 4.23) - 0.5}{\sqrt{\frac{2.2^2}{50} + \frac{2.56^2}{50}}} = -0.293$$

Based on the alternative hypothesis, it is a one-sided test, thereby at 0.05 significance level, $z_{0.05} = 1.645$. The critical zone is $[1.645, \infty)$.

Since the value of the test statistic does not lie in the critical zone, the null hypothesis cannot be rejected at significance level of 0.05.

Therefore, it can be concluded that at a significance level of 0.05, the claim cannot be supported.

Example 6.4.8

The maximum daily temperature values are recorded at a weather station since last 100 years. The data is divided into two epochs (50 years each). The following calculations are made,

Time period	Mean (°C)	Standard deviation (°C)
Epoch 1	35.21	3.48
Epoch 2	35.94	3.20

(a) Test the hypothesis that the mean of the maximum temperature is increasing from epoch 1 to epoch 2 at a significance level of 0.05.

(b) Calculate the p-value of the test.

Solution According to the example, the null and alternative hypothesis can be formulated as follows:

Let μ_1 and μ_2 are the mean temperature during epoch 1 and 2, respectively.

Null hypothesis $H_o : \mu_1 - \mu_2 \geq 0$
Alternative hypothesis $H_a : \mu_1 - \mu_2 < 0$
Level of significance: $\alpha = 0.05$ (given)

As the standard deviation of the population is given, the z statistics can be used.

$$z = \frac{(\overline{x}_2 - \overline{x}_1) - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(35.21 - 35.94) - 0}{\sqrt{\frac{3.48^2}{50} + \frac{3.20^2}{50}}} = -1.09$$

(a) Based on the alternative hypothesis, it is a one-sided test, thereby at 0.05 significance level $z_{0.05} = 1.645$. The critical zone is $(-\infty, -1.645]$.

Since the value of the test statistic does not lie in the critical zone, the null hypothesis cannot be rejected at a level of significance 0.5.

Therefore, the claim that the mean maximum temperature is increasing for epoch 1 to epoch 2 cannot be supported at a significance level of 0.05.

(b) The p-value of the test is $P(z) < -1.09 = 0.138$, the probability that the temperature during epoch 2 is not more than epoch 1.

Case 2: Hypothesis concerning two means when the variances are unknown

This is same as case 1, but the population variances are unknown. In such cases, a pooled variance $\left(S_p^2\right)$ is computed using the sample statistics. The test statistic is as follows:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - \delta}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{6.11}$$

where S_p is the pooled standard deviation and it is expressed as $S_p = \sqrt{\frac{S_1^2(n_1-1)+S_2^2(n_{21}-1)}{n_1+n_2-2}}$, $\overline{x_1}$ and $\overline{x_2}$ are the sample means and δ is the difference between the means for which the hypothesis is to be tested. The t statistic follows t-distribution with $n_1 + n_2 - 2$ degrees of freedom. The rejection criterion for three different cases is shown as follows:

H_o	H_a	**Rejection Region**
$\mu_1 - \mu_2 = \delta$	$\mu_1 - \mu_2 \neq \delta$	$\lvert t \rvert > t_{\alpha/2,(n-1)}$
$\mu_1 - \mu_2 \geq \delta$	$\mu_1 - \mu_2 < \delta$	$t < -t_{\alpha,(n-1)}$
$\mu_1 - \mu_2 \leq \delta$	$\mu_1 - \mu_2 > \delta$	$t > t_{\alpha,(n-1)}$

Example 6.4.9

Annual rainfall received on the leeward side (A) and windward side (B) of a mountain is as follows. Test the claim that the mean rainfall received on the windward side is higher than that on the leeward side. Consider the confidence level of 95%.

Year	1998	1999	2000	2001	2002	2003	2004	2005
Rainfall A (mm)	1225	1075	1260	1100	1125	1275	1300	1205
Rainfall B (mm)	1276	1135	1288	1255	–	1365	1345	1310

Solution Considering μ_A and μ_{B} as the mean rainfall on leeward side (A) and windward side (B), respectively.

Null hypothesis $H_o : \mu_B - \mu_{\mathrm{A}} \leq 0$
Alternative hypothesis $H_a : \mu_B - \mu_{\mathrm{A}} > 0$
Level of significance: $\alpha = 0.05$ (given)

As the standard deviation is to be calculated from the sample data, we will use the t statistic.

Mean rainfall for case A, $\overline{x}_A = 1195.62$ mm
Mean rainfall for case B, $\overline{x}_B = 1282.00$ mm
Standard deviation for case A, $S_A = 85.33$ mm
Standard deviation for case B, $S_B = 75.33$ mm
The pooled standard deviation,

$$S_p = \sqrt{\frac{S_A^2\,(n_1 - 1) + S_B^2\,(n_2 - 1)}{n_1 + n_2 - 2}} = \sqrt{\frac{85.33^2\,(8 - 1) + 75.33^2\,(7 - 1)}{8 + 7 - 2}} = 80.87 \text{ mm}$$

$$t = \frac{(\overline{x}_B - \overline{x}_A) - \delta}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(1282.00 - 1195.62) - 0}{86.81\sqrt{\frac{1}{8} + \frac{1}{7}}} = 1.922$$

Based on the alternative hypothesis, it is one-sided test, thereby at 0.05 significance level $t_\alpha(n_1 + n_2 - 2) = 1.771$. The critical zone is $[1.771, \infty)$.

Since the value of the test statistic lies in the critical zone, the null hypothesis must be rejected at a level of significance 0.05. Therefore, it can be concluded that at a significance level of 0.05, the claim can be supported.

Case 3: Hypothesis concerning two variances

This case deals with the testing of the null hypothesis if a population variance equals/is greater than/is less than that of another population variance with suitable one-sided or two-sided test.

$$F = \frac{S_1^2}{S_2^2} \tag{6.12}$$

where S_1 and S_2 are the sample variances. The F statistics follows F distribution with $n_1 - 1, n_2 - 1$ degrees of freedom.

The rejection criterion for three different cases is shown as follows:

H_o	H_a	Test Statistic	Rejection Region
$\sigma_1^2 \le \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$F > F_\alpha\,(n_1 - 1,\ n_2 - 1)$
$\sigma_1^2 \ge \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$F = \frac{S_2^2}{S_1^2}$	$F > F_\alpha\,(n_2 - 1,\ n_1 - 1)$
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \ne \sigma_2^2$	$F = \frac{S_M^2}{S_m^2}$	$F > F_\alpha\,(n_M - 1,\ n_m - 1)$

For the last case ($H_0 : \sigma_1^2 = \sigma_2^2$ and $H_a : \sigma_1^2 \ne \sigma_2^2$), the sample having higher standard deviation is identified and its standard deviation (S_M) is placed in the numerator. Other one (S_m) is in the denominator. This is to ensure that the rejection region is $F > F_\alpha\,(n_M - 1,\ n_m - 1)$. Relaxation of this criterion is also mathematically possible,but rejection region will have to be modified accordingly. It is recommended to stick to this rule to avoid confusion.

Example 6.4.10

Determine whether the variance of rainfall at a gauging station A is more than that at another gauging station B. Use the data given in Example 6.4.9. Consider a significance level of 0.01.

Solution *Null hypothesis* $H_o : \sigma_A^2 \leq \sigma_B^2$

Alternative hypothesis $H_a : \sigma_A^2 > \sigma_B^2$

Level of significance: $\alpha = 0.01$(given)

Standard deviation for the station A, $S_A = 85.33$

Standard deviation for the station B, $S_B = 75.33$

The test statistic

$$F = \frac{S_A^2}{S_B^2} = \frac{85.33^2}{75.33^2} = 1.283$$

Since the alternative hypothesis is one-sided test, thereby at 0.01 significance level $F_\alpha(n_1 - 1, n_2 - 1) = F_{0.01}(7, 6) = 8.26$. The critical zone is $[8.26, \infty)$.

Since the value of the test statistic does not lie in the critical zone, we cannot reject the null hypothesis at a significance level of 0.01. Thereby, the variance of rainfall at gauging station A may not be concluded to be more than that at rainfall gauging station B at a significance level of 0.01.

Test Concerning Proportion

Some hydrologic or hydroclimatic problems deal with the proportion or percentage of certain attributes. In such cases, it is often required to verify the null hypothesis that a proportion/percentage equals some specific value either for a single sample or among multiple samples.

Case 1: Hypothesis concerning one proportion

This case deals with testing the null hypothesis if a proportion/percentage based on a population is equal to some specific value with suitable one-sided or two-sided test. The test statistic

$$Z = \frac{X - np_o}{\sqrt{np_o\,(1 - p_o)}} \tag{6.13}$$

where n is the size of the sample, X is a subset of the sample which satisfies a given condition, and p_o is the constant value for which we have to test the hypothesis. Assuming that the sample size is sufficiently large, the statistic Z is a random variable that approximately follows standard normal distribution. The rejection criterion for three different cases is shown as follows:

H_o	H_a	Rejection Region
$p = p_o$	$p \neq p_o$	$\lvert z \rvert > Z_{\alpha/2}$
$p \geq p_o$	$p < p_o$	$z < -Z_\alpha$
$p \leq p_o$	$p > p_o$	$z > Z_\alpha$

Example 6.4.11

The probability of failure of the dam due to quick sand condition is 10%. A study is carried out on dams built under similar conditions following the same design details shows that 2 out of 50 dams fail. On the basis of this study, test the claim that the probability of not failing due to quick sand condition is greater than 90% at a significance level of 0.05.

Solution *Null hypothesis* $H_o : p \leq 0.9$

Alternative hypothesis $H_a : p > 0.9$

Level of significance: $\alpha = 0.05$ (given)

In this case, the z statistics can be used as follows.

$$Z = \frac{X - np_o}{\sqrt{np_o\left(1 - p_o\right)}} = \frac{48 - 50 \times 0.9}{\sqrt{50 \times 0.9\left(1 - 0.9\right)}} = 1.414$$

Based on the alternative hypothesis, it is one-sided test, thereby at 0.05 significance level $Z_\alpha = 1.645$. The critical zone is $[1.645, \infty)$.

Since the value of the test statistic does not lie in the critical zone, we cannot reject the null hypothesis at a significance level of 0.05. Thereby, the claim cannot be supported at a significance level of 0.05; i.e., the probability of success is not more than 90%.

Case 2: Hypothesis concerning multiple proportions

This case deals with the several proportions $(p_1, p_2, \ldots, p_k)$. The null hypothesis considers several proportions/percentages based on multiple populations. It is tested whether all the proportions are equal to one another with suitable one-sided or two-sided test. The null hypothesis considered for the test is that $p_1 = p_2 = \cdots = p_k$ against the alternative hypothesis which states that the proportions are not equal.

Two or more proportions from multiple populations can be compared using the test statistic

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} \tag{6.14}$$

is a random variable that follows an approximate chi-square distribution with (k-1) degrees of freedom. The null hypothesis should be rejected if $\chi^2 > \chi^2_\alpha$, where α is the significance level. The magnitudes of o_{ij} and e_{ij} can be computed by arranging the available data as follows:

Description	Sample#1	Sample#2	$\cdots$	Sample#k	Total
Total sample size	n_1	n_2	$\cdots$	n_k	n
Number of *Success* (o_{1j})	x_1	x_2	$\cdots$	x_k	x
Number of *Failures* (o_{2j})	$n_1 - x_1$	$n_2 - x_2$	$\cdots$	$n_k - x_k$	$n - x$
Expected cell frequency for *success* (e_{1j})	$n_1 x \big/ n$	$n_2 x \big/ n$	...	$n_k x \big/ n$	
Expected cell frequency for *failure* (e_{2j})	$n_1(n-x) \big/ n$	$n_2(n-x) \big/ n$	...	$n_k(n-x) \big/ n$	

Here, x is the total number of successes and n is the total number of trials for all the samples. The number of successes or failure (category) is known as observed cell frequency $\left(o_{ij}\right)$ where $i = 1,\ 2$ and $j = 1,\ 2,\ \ldots,\ k$. The values of e_{ij} ($i = 1,\ 2$ and $j = 1,\ 2,\ \ldots,\ k$) are the expected cell frequencies that are evaluated by multiplying the total of the column to the total of the row to which it belongs and then dividing by the grand total n.

Example 6.4.12

Number of rainy days in the year 2016 at three stations (A, B, and C) can be categorized as follows:

	Station A	Station B	Station C	Total
High	70	60	60	190
Low	180	170	190	540
Total	250	230	250	730

Use the 0.05 level of significance to test whether the probability of high rainfall days is the same for the three stations.

Solution *Null hypothesis* $H_o : p_1 = p_2 = p_3$

Alternative hypothesis $H_a : p_1, p_2$ and p_3 are not all equal.

Level of significance: $\alpha = 0.05$ (given)

The expected frequencies for each cell can be evaluated as follows:

Description	Station A	Station B	Station C	Total
Total number of rainy days	250	230	250	730
Number of *high rainfall days* (O_{1j})	70	60	60	190
Number of *low rainfall days* (O_{2j})	180	170	190	540
Expected cell frequency for *high rainfall days* (e_{1j})	$\frac{250\times 190}{730} = 65.07$	$\frac{230\times 190}{730} = 59.86$	$\frac{250\times 190}{730} = 65.07$	
Expected cell frequency for *low rainfall days* (e_{2j})	$\frac{250\times 540}{730} = 184.93$	$\frac{230\times 540}{730} = 170.14$	$\frac{250\times 540}{730} = 184.93$	

In this case, the χ^2 statistics can be evaluated as follows.

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{\left(o_{ij}-e_{ij}\right)^2}{e_{ij}}\\ &= \frac{(70-65.07)^2}{65.07}+\frac{(60-59.86)^2}{59.86}+\frac{(60-65.07)^2}{65.07}\\ &+\frac{(180-184.93)^2}{184.93}+\frac{(170-170.14)^2}{170.14}+\frac{(190-184.93)^2}{184.93}\\ &= 1.04\end{aligned}$$

The value of $\chi^2_{0.05}$ for degrees of freedom of $3-1=2$ is 5.991. As the value of test statistic is less than 5.991, the null hypothesis cannot be rejected. Therefore, the probability of high rainfall days is the same for the three stations.

6.4.4 *Goodness-of-Fit Test*

The goodness-of-fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. Such measures can be used in statistical hypothesis testing, e.g., to test for normality of residuals, to test whether two samples are drawn from population with identical distributions or whether outcome frequencies follow a specified distribution.

Chi-Square Goodness-of-Fit Test

Chi-square goodness-of-fit test is used to test how well a theoretical distribution fits the empirical distribution. It is used to compare the observed sample distribution with expected probability distribution. The sample data is divided into intervals, and the number of points in the interval is compared with expected number of points in each interval using hypothesized distribution. In the case of relative frequency function, the χ^2 *test* is used. The sample value of the relative frequency of ith interval is

$$f_S\left(x_i\right) = n_i \big/ n \tag{6.15}$$

where n_i is the observed number of occurrences in the ith interval and n is total number of observations. The theoretical value of relative frequency is

$$P\left(x_i\right) = F\left(x_i\right) - F\left(x_i - 1\right) \tag{6.16}$$

The χ^2 test statistics is given by,

$$\chi_c^2 = \sum_{i=1}^{m} \frac{n\left[f_s\left(x_i\right) - P\left(x_i\right)\right]^2}{P\left(x_i\right)} \tag{6.17}$$

where m is the number of intervals and the degree of freedom $\nu = m - p - 1$, where p is the number of parameters used in fitting the distribution.

It may be noted that $nf_s\left(x_i\right) = n_i$ is the observed number of occurrences in the interval i and $nP\left(x_i\right)$ is the corresponding expected number of occurrences in the interval i. A confidence level is chosen for the test which is often expressed as $(1 - \alpha)$ where α is termed as significance level. The null hypothesis for the test is that the proposed probability distribution fits the data adequately, and alternative hypothesis states that the data does not follow the proposed probability distribution. The null hypothesis should be rejected if $\chi_c^2 > \chi_\alpha^2$.

Example 6.4.13

The following table provides the range of rainfall during Indian summer monsoon months (total rainfall for four months) at a gauging station with the frequency of occurrence. The mean and standard deviation are given as 397 mm and 92 mm, respectively. Use the χ^2 test to determine whether the normal distribution adequately fits the data at a significance level of 0.05.

Range (mm)	<200	200–250	250–300	300–350	350–400	400–450	450–500	>500
Frequency	1	2	6	14	11	10	5	3

Solution The range of rainfall is divided into eight intervals, starting from less than 200 to more than 500, with intermediate intervals each covering a range of 50. The total number of observations/sample size evaluated as the sum of each frequency is 52.

Null hypothesis H_0: The data fits the normal distribution
Alternative hypothesis, H_a: The data does not fit the normal distribution

The χ^2 test statistics is as follows:

$$\chi_c^2 = \sum_{i=1}^{m} \frac{n\left[f_s\left(x_i\right) - P\left(x_i\right)\right]^2}{P\left(x_i\right)}$$

It follows an approximate chi-square distribution with degrees of freedom $\nu = m - p - 1$. For evaluation of the test statistic, the following table is to be formulated. As an example, each expression for the 5th interval is solved.

The relative frequency function,

$$f_s\left(x_5\right) = \frac{n_5}{n} = \frac{11}{52} = 0.211$$

The cumulative frequency function,

$$F_S\left(x_5\right) = \sum_{i=1}^{5} f_S\left(x_i\right) = 0.654$$

The standard normal variate,

$$z_5 = \frac{x_5 - \mu}{\sigma} = \frac{400 - 397}{92} = 0.033$$

The cumulative normal probability function,

$$\begin{aligned} P\left(x_5\right) &= P\left(350 \leq X \leq 400\right) \\ &= F\left(400\right) - F\left(350\right) \\ &= 0.5130 - 0.3047 \\ &= 0.2083 \end{aligned}$$

The χ^2 test statistic,

$$\chi_c^2 = \frac{n\left[f_s\left(x_i\right) - P\left(x_i\right)\right]^2}{P\left(x_i\right)} = \frac{52 \times (0.2115 - 0.2083)^2}{0.2083}$$

The final test statistic can be evaluated by evaluating the sum of the last column of the table.

Interval	Range (mm)	n_i	$f_s(x_i)$	$F_s(x_i)$	z_i	$F(x_i)$	$P(x_i)$	χ_c^2
1	<200	1	0.0192	0.0192	-2.1413	0.0161	0.0161	0.0311
2	200-250	2	0.0385	0.0577	-1.5978	0.0550	0.0389	0.0003
3	250-300	6	0.1154	0.1731	-1.0543	0.1459	0.0908	0.3455
4	300-350	14	0.2692	0.4423	-0.5109	0.3047	0.1589	3.9875
5	350-400	11	0.2115	0.6538	0.0326	0.5130	0.2083	0.0026
6	400-450	10	0.1923	0.8462	0.5761	0.7177	0.2047	0.0391
7	450-500	5	0.0962	0.9423	1.1196	0.8686	0.1508	1.0306
8	>500	3	0.0577	1.0000	1.6630	1.0000	0.1314	2.1521
							Sum =	7.588

In this case, the last column sums up to be 7.588, i.e., $\chi_c^2 = 7.588$.

The degree of freedom $\nu = m - p - 1$ is equal to five. The value of α is given as 0.05. Thereby, $\chi^2_{0.05}(5) = 11.07$, so the rejection zone is $[11.07, \infty)$. As χ_c^2 does not lie in the rejection zone, the null hypothesis cannot be rejected at a significance level of 0.05. Hence, it can be concluded that the data fits normal distribution.

6.4.5 Nonparametric Test

So far, some parametric form of distribution is assumed for data to perform the hypothesis tests. However, in many cases of hydrology and hydroclimatology, the data may not fit to any specific probability distribution assumption and it is required to opt for nonparametric tests.

Sign Test

This nonparametric test is used as an alternative to the one-sample t test or paired t test. The sign test is applicable for the *large* samples from symmetrical distribution. This property of symmetry may not be always possible to check with small sample (mean divides the data into equal halves). Thus, median is chosen for the test instead of mean. The null hypothesis $\tilde{\mu} = \tilde{\mu}_o$ is tested against an appropriate alternative hypothesis, where $\tilde{\mu}$ is the median of the sample. To carry out the test, each sample value greater than $\tilde{\mu}_o$ is replaced with '$+$' and each value less than the same is replaced with '$-$'. Any value equal to $\tilde{\mu}_o$ is discarded. The null hypothesis that these plus and minus signs are outcomes of binomial trials with $p = 1/2$ is tested.

Example 6.4.14
The following data for maximum temperature (in °C) at a location is recorded for 10 days. Test the null hypothesis $\tilde{\mu} = 30$ against the alternative hypothesis $\tilde{\mu} > 30$ at the 0.01 level of significance.

Day	1	2	3	4	5	6	7	8	9	10
Temperature (°C)	30	36	34	32	29	28	31	34	36	36

Solution *Null hypothesis*: $\tilde{\mu} = 30$
Alternative hypothesis: $\tilde{\mu} > 30$
Level of significance $\alpha = 0.01$

Replacing each value greater than 30 with a plus sign, each value less than 30 with minus sign and discarding any value equal to 30, the following table is obtained,

Temperature	30	36	34	32	29	28	31	34	36	36
Sign	*discarded*	+	+	+	−	−	+	+	+	+

The number of plus signs(x) is equal to seven. Considering $n = 9$ and $p = 0.5$, the probability of $X \geq 7$ can be evaluated using binomial distribution.

$$P(X \geq 7) = 1 - P(X < 7) = 1 - {}^{n}C_x p^x (1 - p)^{n-x}$$
$$= {}^{9}C_7 0.5^7 0.5^2 = 1 - 0.91 = 0.089$$

As 0.089 is greater than 0.01, the null hypothesis cannot be rejected.

Thereby, the median of maximum temperature at the location does not exceed 30° C.

Rank-Sum Test

The rank-sum test includes two types of test, namely U test and H test. The U test (also known as *Wilcoxon test or Mann–Whitney test*) is used as an alternative to two sample t test and H test (also known as Kruskal–Wallis test) is used to check whether n samples come from identical population against an alternative hypothesis that the populations are not identical.

U Test/Wilcoxon test/Mann–Whitney test

In case of U test, the null hypothesis to be tested is that the two samples come from identical population. To satisfy the above condition, the sum of the ranks assigned to the values of both the samples should be more or less same. The test statistics used for the study is

$$Z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}} \tag{6.18}$$

where $\mu_{U_1} = \frac{n_1 n_2}{2}$ and $\sigma_{U_1} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$, n_1 and n_2 are the respective sample size for sample 1 and 2. U_1 is evaluated as follows:

$$U_1 = W_1 - \frac{n_1 (n_1 + 1)}{2} \tag{6.19}$$

To compute W_1, first the data from both the samples are considered together and ranks are provided. Next W_1 is computed as the sum of the ranks values for the data in the first sample. We can also compute W_2 (and subsequently U_2) in the same way for the second sample but either one of the W_1 and W_2 is sufficient for the test. Hence, only W_1 is computed.

Conditions for rejection of the null hypothesis: The Z statistic in Eq. 6.18 follows approximate standard normal distribution. If the null hypothesis states that population 1 is stochastically identical to population 2, then the rejection zones are $(-\infty, -Z_{\alpha/2}]$ and $[Z_{\alpha/2}, \infty)$, where α is the significance level. When the alternative hypothesis states that population 2 is stochastically larger than population 1, then the rejection zone is $[-Z_\alpha, \infty)$, as small values of U_1 corresponds to small values of W_1. Similarly, when the alternative hypothesis states that population 1 is stochastically larger than population 2, then the rejection zone is $(\infty, Z_\alpha]$. Considering p_1 as population 1 and p_2 as population 2, the rejection criteria are shown in the following table:

H_o	***H_a***	**Rejection Region**
p_1 and p_2	p_1 and p_2 are not stochastically identical	$\lvert z \rvert > Z_{\alpha/2}$
are stochastically	p_1 is stochastically less than p_2	$z < -Z_\alpha$
identical	p_1 is stochastically greater than p_2	$z > Z_\alpha$

Example 6.4.15

Let us consider the data provided in Example 6.4.9. Use the U test to show that the rainfall values at two stations belong to same/identical population at a significance level of 0.05. The table is shown here again

Year	1998	1999	2000	2001	2002	2003	2004	2005
Rainfall A (mm)	1225	1075	1260	1100	1125	1275	1300	1205
Rainfall B (mm)	1276	1135	1288	1255	–	1365	1345	1310

Solution *Null hypothesis*: Population are identical.

Alternative hypothesis: The population are not identical.

Level of significance: $\alpha = 0.05$ (given)

The data from both the stations (A and B) are considered together, and ranks are provided. The ranks are shown in the parentheses below each data in the following table:

Year	1998	1999	2000	2001	2002	2003	2004	2005
Rainfall A (mm)	1225 (6)	1075 (1)	1260 (8)	1100 (2)	1125 (3)	1275 (9)	1300 (12)	1205 (5)
Rainfall B (mm)	1276 (10)	1135 (4)	1288 (11)	1255 (7)	–	1365 (15)	1345 (14)	1310 (13)

The sum of ranks assigned to the first sample designated as W_1 is 46. In this case, the Z statistics can be used as follows:

$$Z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}}$$

$$U_1 = W_1 - \frac{n_1(n_1+1)}{2} = 46 - \frac{8(8+1)}{2} = 10$$

$$\mu_{U_1} = \frac{n_1 n_2}{2} = 28$$

$$\sigma^2_{U_1} = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = 74.67$$

Therefore,

$$Z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}} = \frac{10 - 28}{\sqrt{74.67}} = -2.083$$

At a significance level of 0.05 for two-sided test, the rejection zone is $(-\infty, -1.96]$ and $[1.96, \infty)$. As the Z statistic falls in the rejection zone, the null hypothesis should be rejected.

Thereby, it can be concluded that the rainfall at both the stations are not essentially from identical population at a significance level of 0.05.

H Test/Kruskal–Wallis test

H test is a generalized form of U test, used to test if k-independent random samples are drawn from identical populations. The null hypothesis to be tested is that the populations are identical against alternative hypothesis that all populations are not identical. The test statistic used is as follows:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

where n_i is the number of observations of the ith sample, $n = n_1 + n_2 + \cdots + n_k$, and R_i is the sum of ranks occupied by the observations of the ith sample. It is to be noted that all the observations from each sample are jointly ranked before calculation of the test statistic. The H statistic is approximated by the chi-square distribution with $k - 1$ degrees of freedom. The null hypothesis can be rejected if $H > \chi^2_\alpha(k-1)$, where α is the significance level.

Example 6.4.16

Three sets of soil moisture (in %) are recorded as follows:

Set A	11.0	24.8	13.7	39.7	19.6	31.4	24.7	34.7
Set B	23.7	18.6	22.5	42.5	29.0	21.4	25.6	22.3
Set C	21.4	26.0	22.8	14.6	39.6	25.3	11.3	–

At significance level of 0.05, can we conclude that all the sets of data are collected from statistically similar regions, so that they belong to the same population?

Solution *Null hypothesis*: Populations are identical.

Alternative hypothesis: The populations are not identical.

Level of significance: $\alpha = 0.05$(given)

The observations from all the three sets (A, B, and C) are considered together, and ranks are provided. The ranks are shown in the parentheses below each data in the following table:

Set A	11.0 (1)	24.8 (13)	13.7 (3)	39.7 (21)	19.6 (6)	31.4 (18)	24.7 (12)	34.7 (19)
Set B	23.7 (11)	18.6 (5)	22.5 (9)	42.5 (22)	29.0 (17)	21.4 (7)	25.6 (15)	22.3 (8)
Set C	21.4 (7.5)	26.0 (16)	22.8 (10)	14.6 (4)	39.6 (20)	25.3 (14)	11.3 (2)	–

The sum of ranks assigned to the three sample designated as R_1, R_2, and R_3 are 93, 94, and 73.5, respectively. The H statistics can be evaluated as follows:

$$\begin{aligned}
H &= \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \\
&= \frac{12}{23(23+1)} \sum_{i=1}^{3} \frac{R_i^2}{n_i} - 3(23+1) \\
&= \frac{12}{552} \left(\frac{93^2}{8} + \frac{94^2}{8} + \frac{73.5^2}{7} \right) - 72 \\
&= -7.71
\end{aligned}$$

At significance level of 0.05, the $\chi^2_{0.05}(3-1) = 5.991$. The rejection zone is $[5.991, \infty)$. As the H statistic does not fall in the rejection zone, the null hypothesis cannot be rejected.

Thereby, it can be concluded at a significance level of 0.05, the three sets of data are collected from same population.

Kolmogorov–Smirnov Goodness-of-Fit Test

The Kolmogorov–Smirnov (KS) test is a nonparametric test to access the difference between cumulative distributions. Two types of tests, namely one-sample and two-sample tests can be carried out. In one-sample test, the difference between the observed/empirical *CDF* and a specific *CDF* (e.g., normal distribution, uniform distribution) is tested. This test is generally considered more efficient than chi-square goodness-of-fit test for small samples. In case of two-sample test, the hypothesis whether two independent samples come from identical distributions is tested.

One-sample test is based on the maximum absolute difference between the empirical *CDF* and the specific theoretical *CDF*. The null hypothesis to be tested is if the sample follows the theoretical distribution against the alternative hypothesis that the sample does not follow the specific distribution.

Rejection criteria of the null hypothesis: If $D_{\max} < D_\alpha$, the null hypothesis cannot be rejected, where $D_{\max}$ is the maximum absolute difference between the empirical *CDF* and the theoretical *CDF*. The values of D_α can be obtained from Table B.8.

Example 6.4.17

Daily maximum monthly temperature at a location, for ten months is as follows: 14.8, 25.0, 28.2, 28.7, 23.1, 4.8, 4.4, 2.4, 6.2, and 19.5. It is desired to check whether the data set is uniformly distributed between 0 to 30° C at a significance level of 0.01.

Solution *Null hypothesis*: Sample follows the given uniform distribution

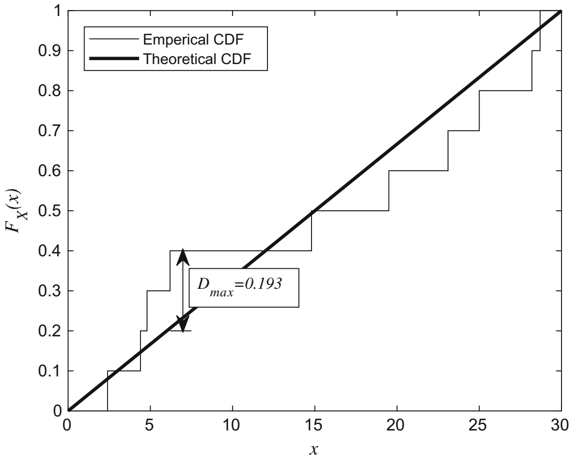
Alternative hypothesis: Sample does not follow the given uniform distribution

Level of significance: $\alpha = 0.01$(given)

The evaluation of the empirical *CDF* ($P_X(x)$) and the *CDF* considering the given uniform distribution with $\alpha = 0$ and $\beta = 30$ ($F_X(x)$) for the given data is shown in the following table:

Data	2.4	4.4	4.8	6.2	14.8	19.5	23.1	25	28.2	28.7
$P_X(x)$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$F_X(x)$	0.080	0.147	0.160	0.207	0.493	0.650	0.770	0.833	0.940	0.956
D	0.020	0.053	0.140	0.193	0.007	−0.05	−0.07	−0.03	−0.04	0.043

The value of $D_{\max}$ can be observed from the last row of the table as 0.193. The value of $D_{\max}$ can also be evaluated using the following figure:



$$D_{\max} = 0.4 - \frac{6.2}{30} = 0.193$$

For $\alpha = 0.01$, the value of $D_\alpha = 0.410$. Since $D_{\max}$ does not exceed 0.410, the null hypothesis cannot be rejected. Thereby, the generated data can be assumed to follow uniform distribution.

Anderson–Darling Goodness-of-Fit Test

The KS tests are not effective for all the cases. Difference in the tails can be easier to detect if the difference between the empirical cumulative distribution $F_X^n(n)$ and $F_X(x)$ is divided by $\sqrt{F_X\,(x)\,(1 - F_X\,(x))}$. In particular, the Anderson–Darling test is based on large values of the statistic,

$$A^2 = \int_{-\infty}^{\infty} \left[F_X^n\,(x) - F_X\,(x)\right]^2 \frac{1}{F_X\,(x)\,(1 - F_X\,(x))} f_X\,(x)\;dx \tag{6.20}$$

The intergration may appear to be different, but A^2 can be computed as

$$A^2 = \frac{\sum_{i=1}^{n}\,(2i - 1)\,(\ln\,(u_i) + \ln\,(1 - u_{n+1} - i))}{n} - n \tag{6.21}$$

where $u_i = F_X(x_i)$ is the value of the theoretical cumulative distribution at the ith largest observation x_i. The null hypothesis is rejected for the large values of the statistic A^2. As a guideline, the large sample 5% point is 2.492 and the 1% points is 3.857. It has been suggested that these critical values are quite accurate even for samples as small as 10.

Example 6.4.18
Using the data provided in Example 6.4.17, check whether the data set follows uniform distribution using Anderson–Darling test at a significance level of 0.01.

Solution *Null hypothesis*: Sample follows the given uniform distribution
Alternative hypothesis: Sample does not follow the given uniform distribution
Level of significance: $\alpha = 0.01$ (given)
For $\alpha = 0.01$, the value of $A^2_\alpha = 3.857$. The test statistic can be evaluated as

$$A^2 = \left((2-1)\left[\ln\left(\frac{2.4}{30} \times \left(1 - \frac{28.7}{30}\right)\right)\right] + (4-1)\left[\ln\left(\frac{4.4}{30} \times \left(1 - \frac{28.2}{30}\right)\right)\right] + \cdots - 10\right)/10$$
$$= 0.5267$$

As $A^2 < A^2_\alpha$, the null hypothesis cannot be rejected at a significance level of 0.01.

Thereby, it can be concluded that the given sample follows uniform distribution.

6.5 MATLAB Examples

This section will provide sample scripts for solving examples using MATLAB. A brief description of each command line is provided at the end of each line after % symbol. The sample code for solving Example 6.4.5 is given in Box 6.1.

Box 6.1 MATLAB script to solve Example 6.4.5

```
clear all;clc;close all
% Inputs
m=570; % mean value
x=[543, 496, 523,450, 576, 590, 505];
% Rainfall depth in mm (Sample data)

% Test the null hypothesis that the data comes from a
%population with mean equal to or greater than 570,against
%the alternative that the mean is less than 570.
[h,p,ci,stats] = ttest(x,m, 'Alpha',0.05,'Tail','left');
% t statistics is used as the standard deviation is to be
%evaluated from the sample.

% Display results
output_file=['output' filesep() 'code_1_results.txt'];
delete(output_file);diary(output_file);diary on;
fprintf(' h= %d\n p= %0.4f \n ci = (%d, %2.3f)\n',h,p,ci(1),ci(2))
    ;
fprintf(' stats\n\t tstat= %1.3f\n\t df=%d\n\t sd=%2.3f\n',...
    stats.tstat,stats.df,stats.sd);
diary off;
```

The result for script provided in Box 6.1 is given in Box 6.2. The returned value of h = 1 indicates that ‘ttest’ rejects the null hypothesis at the 5% significance level, in favor of the alternate hypothesis The value of p signifies the p-value. The concluding remark and the p-value are the same as evaluated in the solution of Example 6.4.5.

Box 6.2 Results for script provided in Box 6.1

```
h= 1
p= 0.0266
ci = (-Inf, 561.634)
stats
  tstat= -2.401
  df=6
  sd=48.323
```

Similarly, the sample code for solving Example 6.4.17 is provided in Box 6.3.

Box 6.3 MATLAB script to solve Example 6.4.17

```
clear all;clc;close all

% Generation of random variables
x=[4.8, 14.8, 28.2, 23.1, 4.4, 28.7, 19.5, 2.4, 25.0, 6.2];

%Defining the CDF of the uniform distribution
test_cdf = makedist('Uniform','lower',0,'upper',30);

%Fitting the given uniform distribution using KS Test
[h,p,ksstat] = kstest(x,'CDF',test_cdf,'Alpha',0.01);

% Display results
output_file=['output' filesep() 'code_2_results.txt'];
delete(output_file);diary(output_file);diary on;
fprintf(' h=%d\n p=%0.4f \n ksstat = %0.4f\n',h,p,ksstat);
diary off;
```

The output of Box 6.3 is provided in Box 6.4. The returned value h = 0 indicates that ‘kstest’ does not reject the null hypothesis at the 1% significance level. Therefore, the data follows uniform distribution as concluded from the solution of Example 6.4.17.

Box 6.4 Results for script provided in Box 6.3

```
h=0
p=0.7827
ksstat = 0.1933
```

Exercise

6.1 Test the claim that the mean annual rainfall in a semiarid region is 750 mm considering significance level of 5%. Also, evaluate the p-value. Using 20 sample data, the mean and standard deviation are calculated as $\overline{X} = 725.5$ mm, $S = 200$ mm. *(Ans: Can't reject the claim; p-value: 0.6.)*

6.2 The 25 years record of observed wind data at a location shows the standard deviation is 76.85 km/h. Considering 95% confidence level, test the hypothesis whether the standard deviation of wind speed at that location is less than 72.53 km/h. *(Ans: Reject the hypothesis at the given confidence level.)*

6.3 The mean annual evaporation from a reservoir is estimated as 1360 mm with standard deviation of 204 mm using 40 years data. Test the hypothesis that $\mu = 1500$ mm considering the level of significance as 0.01. *(Ans: Reject the hypothesis at the given confidence level.)*

6.4 Test the hypothesis that annual average local sea level at station A is 30 mm higher than another station B at significance level of 0.01 and 0.05. Also, determine whether the variability at station A is greater than station B at 0.01 significance level. Estimated sample statistics are obtained as $\overline{X}_A = 76$ mm, $S_A = 15$ mm and $\overline{X}_B = 59.6$ mm, $S_B = 12.5$ mm using 60 years of data. (Ans: *Reject the hypothesis at both levels of significance; Yes, variability at station A is greater than station B.*)

6.5 The temperature data during the month of June at a city is given in the following table.

Year	2010	2011	2012	2013	2014	2015	2016
Temperature (°C)	43.6	46.4	44.9	45.7	47.1	44.2	42.8

Test the claim that the mean temperature is greater than 45° C at a significant level of (a) 0.01 and (b) 0.05. (Ans: *Can't reject the claim at both levels of significance.)*

6.6 The meteorologists claimed that at least 95% of the stream-flow measuring devices are functioning properly. 160 gauges are examined, and 15 gauges are found to be damaged. Test the claim of meteorologists using a significance level of 0.01 and 0.05. *(Ans: Reject the claim at both levels of significance.)*

6.7 Two teams A and B went to collect soil samples for field measurements of soil moisture. 200 and 120 samples were collected by them, respectively. Later, it is found that 16 and 8 samples are not usable, collected by team A and B, respectively. Test the hypothesis that (a) both the teams were showing equal proficiency in collecting the samples (proportions are equal), (b) If not, which team is more efficient in terms of collecting usable samples? Use $\alpha = 0.05$ for both the cases. *(Ans: (a) Reject the hypothesis that both the teams are showing equal proficiency in collecting the samples, (b) Team A is better.)*

6.8 Two groups of groundwater measuring wells are considered depending on the topographical characteristics. Group I shows a mean depth of 10.32 m and standard deviation of 1.18 m. Similarly, Group II shows a mean depth of 13.30 m and standard deviation of 0.96 m. Find out whether the difference between two groups is significant, using $\alpha = 0.01$ and 0.05. Also, calculate the p-value. *(Ans: The difference between two groups is significant at both significance level; p-value* $= 6.9 \times 10^{-80}$.)

6.9 The following streamflow measurements are taken from two different outlets.

Outlet 1 (cumec)	7268	7130	7351	7070	7346
Outlet 2 (cumec)	6954	7332	7043	6825	7350

Test whether the difference between the means of both the outlets is significant using $\alpha = 0.01$. *(Ans: The difference between observations of two outlets is not significant at given significance level.)*

6.10 The mean annual rainfall at a location was estimated as 1100 cm with a standard deviation of 120 cm during pre-industrialization period. Recently, 20 observations are considered and the mean is estimated as 1030 cm. Test the hypothesis that the mean annual rainfall has not changed, using 0.05 and 0.01 significant levels. Assume that standard deviation remains same. *(Ans: The mean annual rainfall has changed.)*

6.11 60 observations on July rainfall are taken at rainguage station A, and variance is estimated 240 mm^2. Similarly, 100 observations are taken at rain gauge station B and variance is estimated as 160 mm^2. Test the hypothesis that variance at station A is greater than station B using (a) $\alpha = 0.05$ and (b) $\alpha = 0.01$. *(Ans: Variance at station A is not greater than station B at both significance level.)*

6.12 Number of rainy days is obtained from three stations. At station A, 41 out of 120, at station B, 27 out of 80, and at station C, 22 out of 100 days were found to be rainy days. Use 0.05 level of significance to test whether the proportion of rainy days is same at all three stations. *(Ans: The proportions are same at significance level of 0.05.)*

6.13 Before and after the installation of a new rain gauge station, the variances are estimated as 106 mm and 128 mm using monthly data for 1 year. Check if the rainfall measurement remains consistent with respect to the variance before and after the installation, at a significance level of 0.05 and 0.01. *(Ans: There is no significant increase in variability at both the significance level.)*

Chapter 7
Regression Analysis and Curve Fitting

Many applications in hydrology and hydroclimatology deal with studying the relationship between the associated variables. The target variable is known as dependent variable, whereas other variables are known as independent variables. In statistics, the procedure of developing such relationship between dependent and independent variables is called regression analysis. The fitted statistical model is termed as regression model. Such models can be used for assessment of the dependent variable, knowing the independent variables. There are different types of regression models, and every regression model consists of some mathematical formulation with parameters to relate independent variables to dependent variable. All these types of regression models are discussed in this chapter.

7.1 Simple Linear Regression

One of the most commonly used models in hydrology is based on the assumption of a linear relationship between two variables. In this particular model, we aim toward representing a dependent variable in terms of linear equation of single independent variable. For example, let us estimate runoff using precipitation. In this case, runoff is the dependent variable Y, whereas precipitation is the independent variable X. It can be noted that observed values of dependent variable Y may vary even for a specific value of independent variable X owing to the uncertainty associated with it arising from unknown factors. Hence, Y is a random variable whose distribution is dependent on X. In such cases, the relationship between X and the mean of the distribution of Y is referred to as **regression curve of Y on X**.

Considering the regression curve to be linear, the regression equation is given by,

$$Y = \alpha + \beta X + \varepsilon \tag{7.1}$$

R. Maity, *Statistical Methods in Hydrology and Hydroclimatology*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-10-8779-0_7

where ε is the difference between observed Y and the estimated Y (represented as $\hat{Y}$), termed as residual. The value of ε will depend on the error in measurement and the influence of unknown factors on Y. For fitting a linear regression, following assumptions are made,

(i) The relationship between the X and mean of Y is a straight line. Hence, for a specific value of X and Y (represented by x_i and y_i, respectively), the regression model is given by,

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

the random variables y_i are independently normally distributed with a mean of $\alpha + \beta x_i$ and variance equal to standard error (represented by σ^2), for $i = \{1, 2, \ldots, n\}$

(ii) The residuals (ε_i) are independent and normally distributed with a mean of zero and the variance of σ^2.

For Eq. 7.1, the estimates of α and β (say a and b) can be calculated using observed values. Hence, the estimated dependent variable (denoted as $\hat{Y}$) is given by,

$$\hat{Y} = a + bX \tag{7.2}$$

Here, it should be noted that the estimated Y is the most expected value of Y given X. In other words, $\hat{Y}$ is mean of distribution of Y given X.

The above equation is a equation of straight line with slope b and intercept a. This line is called **fitted or estimated regression line**. Further, due to uncertainty of Y, the $\hat{Y}$ differs from Y (the difference is termed residual, as stated before). The ith residual (ε_i) is given by,

$$\varepsilon_i = y_i - \hat{y}_i \tag{7.3}$$

where $\hat{y}_i$ is the estimate for y_i (i.e., ith observation of Y) using the Eq. 7.2. The aim of regression line fitting is to get the estimate of α and β (say a and b), such that the prediction errors are minimum. It is not possible to minimize all the errors simultaneously, and thereby, the sum of squared errors is minimized. However, prediction error may have positive or negative values. Therefore, a sign-independent criterion is needed, such as minimization of either $\sum_{i=1}^{n} |\varepsilon_i|$ or $\sum_{i=1}^{n} \varepsilon_i^2$. Mathematically, working with absolute values is difficult as compared to working with square function. Hence, $\sum_{i=1}^{n} \varepsilon_i^2$ is minimized to get the estimate of α and β (i.e., a and b). As the sum of squared errors is minimized for estimation of regression parameters, this method is called the **method of least squares** and estimated parameters are called **least square estimates** . The method of least square is described as follows. The sum of squared errors can be expressed as a function of parameters as,

$$S(\alpha, \beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2 \tag{7.4}$$

If a and b are estimate of α and β such that the sum of squared errors is minimized when $\alpha = a$ and $\beta = b$, then partial derivative of S with respect to α and β at $\alpha = a$ and $\beta = b$ should be zero.

$$\left.\frac{\partial S}{\partial \alpha}\right|_{\alpha=a,\,\beta=b} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$\text{Hence, } \sum_{i=1}^{n}(y_i - a - bx_i) = 0 \tag{7.5}$$

$$\left.\frac{\partial S}{\partial b}\right|_{\alpha=a,\,\beta=b} = -2\sum_{i=1}^{n}x_i\,(y_i - a - bx_i) = 0$$

$$\text{Hence, } \sum_{i=1}^{n}x_i\,(y_i - a - bx_i) = 0 \tag{7.6}$$

Thereby, by eliminating a and solving Eqs. 7.5 and 7.6 for b, the **least square estimates** of b can be written as,

$$b = \left[\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i/n\right] \Bigg/ \left[\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2 \Bigg/ n\right] \tag{7.7}$$

$$b = \frac{S_{xy}}{S_{xx}} \tag{7.8}$$

where

$$S_{xy} = \sum_{i=1}^{n}x_i y_i - \frac{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n} \quad \text{and} \quad S_{xx} = \sum_{i=1}^{n}x_i^2 - \frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n} \tag{7.9a}$$

Alternatively, if the ith deviation of X and Y from their respective means is given by x_{di} and y_{di}, respectively, then,

$$S_{xy} = \sum_{i=1}^{n}x_{di}\,y_{di} \quad \text{and} \quad S_{xx} = \sum_{i=1}^{n}x_{di}^2 \tag{7.9b}$$

Further, the **least square estimates** of a can be written as,

$$a = \frac{1}{n}\left(\sum_{i=1}^{n}y_i - b\sum_{i=1}^{n}x_i\right) = \overline{Y} - b\overline{X} \tag{7.10}$$

The individual deviations of the observations y_i from their fitted values $\hat{y}_i = a + bx_i$ are called the **residuals**. Thus, ith residual is expressed by,

$$\varepsilon_i = y_i - a - bx_i \tag{7.11}$$

The minimum value of the sum of square prediction errors is called the **residual sum of squares (RSS)** or **sum of squared errors** (SSE).

$$SSE = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{7.12}$$

Example 7.1.1
For a large catchment, the precipitation and runoff are being recorded monthly. The records for 2 years are tabulated in the following table:

The variables are assumed to be linearly related. Work out a relationship between the monthly precipitation and runoff for the location and use the relationship to estimate the expected amount of runoff generated when monthly precipitation is 14 cm (Table 7.1).

Solution The runoff and precipitation at monthly scale are assumed to be linearly related. A scattergram (Fig. 7.1) between the variables reveals that the relationship is linear. Let us consider Y to be a random variable for monthly runoff and X to be a random variable for monthly precipitation. Relationship between X and expected value of Y, being linear, is expressed as (Eq. 7.2),

$$\hat{Y} = bX + a$$

The parameters b and a can be estimated using the Eqs. 7.8, 7.9a and 7.10, respectively. These calculations are tabulated in Table 7.2.

Table 7.1 Monthly precipitation (X in cm) and runoff (Y in cm) for 2 years

Month	X	Y	Month	X	Y
January, 2010	6.9	2.4	January, 2011	5.5	1.4
February, 2010	6.4	1.1	February, 2011	11.4	6.3
March, 2010	6.5	1.7	March, 2011	10.8	4.4
April, 2010	5.1	0.5	April, 2011	7.5	1.8
May, 2010	7.1	1.8	May, 2011	8.2	4.2
June, 2010	7.1	2.0	June, 2011	7.9	2.9
July, 2010	10.2	4.2	July, 2011	4.1	0.0
August, 2010	9.9	3.0	August, 2011	5.0	1.3
September, 2010	8.4	4.7	September, 2011	6.7	1.3
October, 2010	5.8	3.4	October, 2011	4.3	0.0
November, 2010	10.1	4.4	November, 2011	10.4	5.9
December, 2010	7.3	2.8	December, 2011	3.9	2.4

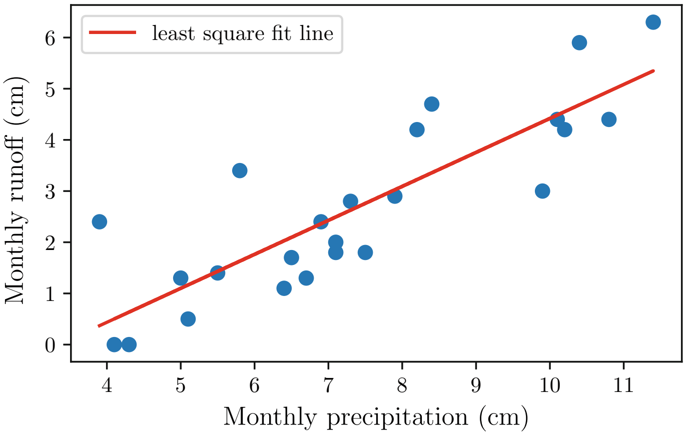


Fig. 7.1 Scattergram between monthly runoff and monthly precipitation

From the table, $n = 24$, $\sum_{i=1}^{n} x_i = 176.5$, $\sum_{i=1}^{n} y_i = 63.9$, $\sum_{i=1}^{n} x_i^2 = 1410.51$, $\sum_{i=1}^{n} x_i y_i = 544.62$ and $\sum_{i=1}^{n} y_i^2 = 239.09$.

Using Eq. 7.8, the parameter b can be calculated as,

$$S_{xy} = \left[\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right) \Big/ n\right] = 544.62 - (176.5 \times 63.9)/24 = 74.689$$

$$S_{xx} = \left[\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n\right] = 1410.51 - (176.5)^2/24 = 112.5$$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{74.689}{112.5} = 0.664$$

Using Eq. 7.10, the parameter a is calculated as,

$$a = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i\right) = (63.9 - 0.664 \times 176.5)/24 = -2.220$$

Hence, the relationship between the expected monthly runoff ($\hat{Y}$ in cm) and monthly precipitation (X in cm) is expressed as (Eq. 7.2),

$$\hat{Y} = 0.664X - 2.220$$

Table 7.2 Calculation for estimating SLR coefficients

S. No.	y_i	x_i	x_i^2	$x_i y_i$	y_i^2
1	2.4	6.9	47.61	16.56	5.76
2	1.1	6.4	40.96	7.04	1.21
3	1.7	6.5	42.25	11.05	2.89
4	0.5	5.1	26.01	2.55	0.25
5	1.8	7.1	50.41	12.78	3.24
6	2.0	7.1	50.41	14.20	4.00
7	4.2	10.2	104.04	42.84	17.64
8	3.0	9.9	98.01	29.70	9.00
9	4.7	8.4	70.56	39.48	22.09
10	3.4	5.8	33.64	19.72	11.56
11	4.4	10.1	102.01	44.44	19.36
12	2.8	7.3	53.29	20.44	7.84
13	1.4	5.5	30.25	7.70	1.96
14	6.3	11.4	129.96	71.82	39.69
15	4.4	10.8	116.64	47.52	19.36
16	1.8	7.5	56.25	13.5	3.24
17	4.2	8.2	67.24	34.44	17.64
18	2.9	7.9	62.41	22.91	8.41
19	0.0	4.1	16.81	0.00	0.00
20	1.3	5.0	25.00	6.50	1.69
21	1.3	6.7	44.89	8.71	1.69
22	0.0	4.3	18.49	0.00	0.00
23	5.9	10.4	108.16	61.36	34.81
24	2.4	3.9	15.21	9.36	5.76
Total	63.9	176.5	1410.51	544.62	239.09

The expected amount of runoff generated by monthly precipitation of 14 cm

$$= 0.664 \times 14 - 2.220 = 7.08 \text{ cm}$$

Example 7.1.2
In a large district, the average monthly air temperature and the average monthly evaporation over 15 water bodies are given below.

The evaporation is expected to increase with temperature. Determine the linear regression equation for estimating the expected evaporation (Y) on the basis of temperature (X) information. Also, calculate the standard error of estimate (Table 7.3).

Solution A linear regression model between the expected monthly evaporation (Y) and average monthly air temperature (X) is given by:

Table 7.3 Average monthly air temperature and evaporation for 15 different water bodies

Location no.	Average monthly temperature (°C)	Average monthly evaporation (mm)
1	22.6	5.2
2	22.1	4.7
3	20.1	2.8
4	29.0	11.3
5	26.7	9.1
6	21.8	4.4
7	23.2	5.8
8	25.6	8.1
9	23.9	6.4
10	26.7	9.1
11	28.4	10.8
12	24.3	6.8
13	29.0	11.3
14	23.6	6.2
15	22.3	4.9

$$\hat{Y} = a + bX$$

The parameters a and b can be estimated using the Eqs. 7.8, 7.9b and 7.10. These calculations are tabulated in Table 7.4.

Hence, $n = 15$, $\sum_{i=1}^{n} x_i = 369.3$, $\sum_{i=1}^{n} y_i = 106.9$, $S_{yy} = \sum_{i=1}^{n} y_{di}^2 = 101.03$, $S_{xx} = \sum_{i=1}^{n} x_{di}^2 = 110.14$, and $S_{xy} = \sum_{i=1}^{n} x_{di} y_{di} = 105.48$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{105.48}{110.14} = 0.958$$

$$a = \overline{Y} - b\overline{X} = (106.9 - 0.958 \times 369.3) \big/ 15 = -16.451$$

Hence, linear regression model between the expected monthly evaporation ($\hat{Y}$ in mm) and average monthly air temperature (X in °C) is given by:

$$\hat{Y} = 0.958X - 16.451$$

Standard error of estimate (s_e) is the sample estimate of σ. $\left(s_e^2\right)$ being an estimate of σ^2, is given by dividing sum of squared errors by $(n - 2)$.

From the table, sum of squared errors $= \sum \left(y_i - \hat{y}_i\right)^2 = \sum (e)^2 = 0.013$. Hence,

Table 7.4 Calculation of SLR parameters for Example 7.1.2

Location	y_i	x_i (°C)	y_{di}	x_{di}	$(y_{di})^3$	$(x_{di})^2$	$x_{di}y_{di}$	$\hat{y}_i$	ε_i	ε_i^2
no.	(mm)		$(y_i-\overline{y})$	$(x_i-\overline{x})$					$(y_i-\hat{y}_i)$	
1	5.2	22.6	−1.93	−2.02	3.71	4.08	3.89	5.199	0.001	0.000
2	4.7	22.1	−2.43	−2.52	5.89	6.35	6.12	4.720	−0.020	0.000
3	2.8	20.1	−4.33	−4.52	18.72	20.43	19.56	2.804	−0.004	0.000
4	11.3	29.0	4.17	4.38	17.42	19.18	18.28	11.330	−0.030	0.001
5	9.1	26.7	1.97	2.08	3.89	4.33	4.10	9.127	−0.027	0.001
6	4.4	21.8	−2.73	−2.82	7.43	7.95	7.69	4.432	−0.032	0.001
7	5.8	23.2	−1.33	−1.42	1.76	2.02	1.88	5.774	0.026	0.001
8	8.1	25.6	0.97	0.98	0.95	0.96	0.95	8.073	0.027	0.001
9	6.4	23.9	−0.73	−0.72	0.53	0.52	0.52	6.444	−0.044	0.002
10	9.1	26.7	1.97	2.08	3.89	4.33	4.10	9.127	−0.027	0.001
11	10.8	28.4	3.67	3.78	13.49	14.29	13.89	10.755	0.045	0.002
12	6.8	24.3	−0.33	−0.32	0.11	0.10	0.10	6.827	−0.027	0.001
13	11.3	29.0	4.17	4.38	17.42	19.18	18.28	11.330	−0.030	0.001
14	6.2	23.6	−0.93	−1.02	0.86	1.04	0.95	6.157	0.043	0.002
15	4.9	22.3	−2.23	−2.32	4.96	5.38	5.17	4.911	−0.011	0.000
Total	106.9	369.3	0.00	0.00	101.03	110.14	105.48	107.01	−0.11	0.013

$$s_e^2 = \frac{1}{n-2}\sum\left(y_i - \hat{y}_i\right)^2 = \frac{0.013}{15-2} = 0.001$$

Standard error of estimate $(s_e) = \sqrt{0.001} = 0.032$

7.2 Curvilinear Regression

In the previous section, the regression equation is considered to be linear that is for a particular value of X, the mean of the distribution of Y is given by $\alpha + \beta x$. In this section, we will consider cases where the regression curve is nonlinear, but the least square method of analysis is still applicable. Such cases of regression are called curvilinear or nonlinear regression. These regression models are classified into two categories:

(i) Model transformable to linear regression;
(ii) Model not transformable to linear regression.

7.2.1 Model Transformable to Linear Regression

Some of the curvilinear regression model, if transformed, can be converted into linear regression model. After transformation, the least square estimates of parameter can be obtained by the method explained in the previous section. Two very commonly used relationships that can be fitted using the least square method after transformation are as follows:

(i) Reciprocal Function:

$$y = \frac{1}{\alpha + \beta x} \tag{7.13}$$

It represents a linear relationship between x and $1\big/y$, namely

$$\frac{1}{y} = \alpha + \beta x \tag{7.14}$$

(ii) Power Function:

$$y = \alpha x^{\beta} \tag{7.15}$$

It represents a linear relationship between $\log(x)$ and $\log(y)$, namely

$$\log\ y = \log\ \alpha + \beta \log\ x \tag{7.16}$$

Example 7.2.1
Multiple models exist for modeling infiltration rate with respect to time. In a flooding type infiltration test, following infiltration capacity data is given in Table 7.5.

The maximum rate at which soil can absorb water at a given time is defined as *infiltration capacity*. It is denoted by f_t. For most of soil, the infiltration capacity at initial time is highest (known as initial infiltration capacity, denoted by f_0), which gradually decreases to steady-state infiltration capacity (also known as constant or ultimate infiltration capacity, denoted by f_c) at $t = t_c$.

(a) Plot the curves for

(i) infiltration capacity versus time;
(ii) infiltration capacity versus cumulative infiltration;
(iii) cumulative infiltration versus time;
(iv) $\ln(f_t - f_c)$ versus time.

(b) Find the least square estimate of Horton's infiltration model parameter given by,

$$f_t = f_c + (f_o - f_c)\ e^{-kt}$$

Table 7.5 Time since start of infiltration experiment and corresponding cumulative infiltration depth

Time since experiment start (minutes)	5	10	20	30	45	60	75	90	105	120
Cumulative infiltration depth (cm)	1.30	2.50	4.30	5.75	7.40	8.75	9.90	10.95	11.90	12.85

where f_t is the infiltration capacity at time t. Similarly, f_0 is the infiltration capacity at $t = 0$ and f_c is the constant infiltration capacity at $t = t_c$.

(c) Fit a Kostiakov infiltration model over the data given by,

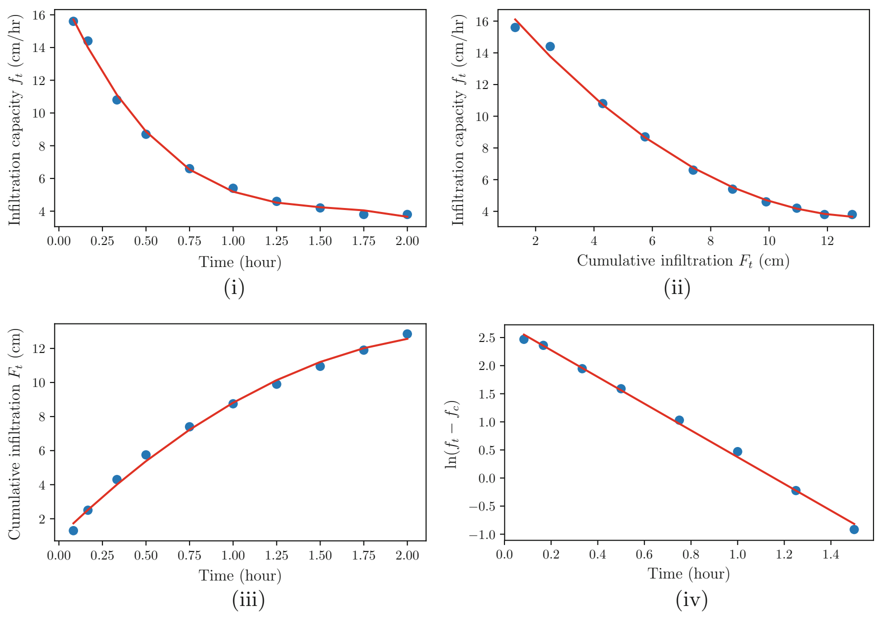
$$F_t = at^b$$

where F_t is the cumulative infiltration capacity at time t.

Solution (a) Incremental infiltration depths along with various other parameters are calculated and shown in following table.

Time in (t) (min)	Cum. depth (F_t) (cm)	Incremental depth (F_{t-1} $-F_t$) (cm)	t in (hrs)	Infiltration capacity (f_t) (cm/h)	$(1/F_t)$	$\ln(f_t - f_c)$
0						
5	1.30	1.30	0.08	15.6	0.77	2.47
10	2.50	1.20	0.17	14.4	0.40	2.36
20	4.30	1.80	0.33	10.8	0.23	1.95
30	5.75	1.45	0.50	8.7	0.17	1.59
45	7.40	1.65	0.75	6.6	0.14	1.03
60	8.75	1.35	1.00	5.4	0.11	0.47
75	9.90	1.15	1.25	4.6	0.10	−0.22
90	10.95	1.05	1.50	4.2	0.09	−0.92
105	11.90	0.95	1.75	3.8	0.08	
120	12.85	0.95	2.00	3.8	0.08	

The relationship between different quantities is shown graphically in following figures.



(i) (ii)

(iii) (iv)

(b) The Horton's infiltration equation can be transformed to linear equation as,

$$f_t = f_c + (f_o - f_c)\ e^{-kt}$$
$$f_t - f_c = (f_o - f_c)\ e^{-kt}$$
$$\ln(f_t - f_c) = \ln\,(f_o - f_c) - kt$$

Hence, by comparing this form of Horton's equation and a linear regression model with $\ln(f_t - f_c)$ as y and t as x,

$$\hat{y} = a + bx$$

From the table, $f_c = 3.8$, $f_o = 15.6$, $a = \ln\,(f_o - f_c) = \ln(15.6–3.8) = 2.468$ and $b = -k$.
Form Eq. 7.10,

$$\begin{aligned} a &= \overline{y} - b\overline{x} \\ \text{or, } 2.468 &= \overline{y} + k\overline{x} \\ \text{or, } 2.468 &= 1.091 + k(0.6975) \\ \text{or, } k &= (2.468 - 1.091)\,\big/0.6975 = 1.9742 \end{aligned}$$

Note that $\overline{x}$ is computed using only first 8 values of x, i.e., t, since only 8 values of y, i.e., $\ln\,(f_t - f_c)$, are available.

Hence, fitted Horton's equation is given by

$$\begin{aligned}\ln(f_t - 3.8) &= 2.468 - 1.974t\\ \text{or, } f_t - 3.8 &= e^{2.468}e^{-1.974t}\\ \text{or, } f_t &= 3.8 + 11.8e^{-1.974t}\end{aligned}$$

(c) The Kostiakov infiltration model can be transformed to linear equation as,

$$\begin{aligned}F_t &= at^b\\ \ln(F_t) &= \ln(a) + b\ln(t)\end{aligned}$$

Hence, $\ln(F_t)$ are $\ln(t)$ in following table.

Time (min)	Cum. Depth (F_t) (cm)	t in (hrs)	$\ln(F_t)$	$\ln(t)$
5	1.30	0.08	0.26	−2.48
10	2.50	0.17	0.92	−1.79
20	4.30	0.33	1.46	−1.10
30	5.75	0.50	1.75	−0.69
45	7.40	0.75	2.00	−0.29
60	8.75	1.00	2.17	0.00
75	9.90	1.25	2.29	0.22
90	10.95	1.50	2.39	0.41
105	11.90	1.75	2.48	0.56
120	12.85	2.00	2.55	0.69

The estimates of parameters of this equation ($\ln(a)$ and b) can be obtained using the Eqs. 7.8 and 7.10 as done in Example 7.1.1.

$$\ln(a) = 2.141 \qquad \text{and} \qquad b = 0.702$$

$$\text{or, } a = \exp(2.141) = 8.508$$

So, Kostiakov infiltration model for the observed infiltration data is given by

$$F_t = 8.508t^{0.702}$$

7.2.2 *Model Not Transformable to Linear Regression*

Secondly, we will consider the case where the functional form of the regression Y on X is not transformable to linear regression. For example, a polynomial fit between Y and X is given by:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon \tag{7.17}$$

where the degree of the equation is determined by inspecting the data. The corresponding coefficients can be calculated by the method of least square as discussed in the previous section. The coefficients of fit (β_i) can be obtained by considering different powers of independent variable as separate independent variable and using the concept of multiple linear regression, which will be discussed in the next section.

7.3 Multiple Linear Regression

In the previous section, we have discussed the relation between a dependent and a single independent variable. However, in many cases, the dependent variable may depend on more than one independent variables. For example, the runoff is dependent on precipitation depth, duration of rainfall, initial losses, and infiltration indices. A general multiple linear regression (MLR) model can be represented as,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{7.18}$$

where Y is the dependent variable and $X_1, X_2, \ldots, X_p$ are the independent variables and $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the unknown parameters. It should be noted that while fitting MLR, assumptions of simple linear regression should hold, and additionally, the data should not have multicollinearity. Multicollinearity represents a situation that linear combination of some inputs (independent variables) results in zero.

Now, a set of observed data will consist of n observations of Y and corresponding n observations of p independent variables. Thereby, the Eq. 7.18 can be written as,

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} \tag{7.19}$$

where Y_i is the ith observation of the dependent variable and $X_{i,j}$ is the ith observation of the jth independent variable. In the form of matrix, it can be written as,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & X_{1,3} & \cdots & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & X_{2,3} & \cdots & X_{2,p} \\ 1 & X_{3,1} & X_{3,2} & X_{3,3} & \ldots & X_{3,p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & X_{n,3} & \cdots & X_{n,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \tag{7.20}$$

or,

$$Y = X\beta \tag{7.21}$$

where Y is an $n \times 1$ vector of the dependent variable, X is an $n \times (p+1)$ matrix of the independent variables, and β is a $(p+1) \times 1$ vector of the unknown parameters. In order to find out the values of the parameters, we can use the least square method as utilized in the earlier sections. Hence by minimizing $\sum_{i=1}^{n} \varepsilon_i^2$, we can obtain $\hat{\beta}$. In matrix form, the sum of squared error can be written as,

$$\sum_{i=1}^{n} \varepsilon_i^2 = e^T e = \left(Y - X\hat{\beta}\right)^T \left(Y - X\hat{\beta}\right) \tag{7.22}$$

Differentiating the above equation with respect to $\hat{\beta}$ and setting the value of the expression to zero, we get

$$X^T Y = X^T X \hat{\beta} \tag{7.23}$$

The solution for $\hat{\beta}$ can be obtained by multiplying both sides of the equation with $\left(X^T X\right)^{-1}$. We finally obtain

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{7.24}$$

Example 7.3.1

The average monthly evapotranspiration is estimated using the average temperature and average wind speed by the following model using the data given in Table 7.6.

$$E\left(Y|X_1, X_2\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where Y: average monthly evapotranspiration (in mm)
X_1: average wind speed (in kmph)
X_2: average temperature (in °C)

Determine β_0, β_1, and β_2.

Table 7.6 The average monthly evapotranspiration and wind speed and temperature for 10 months

Observation no.	Evapotranspiration (Y) (mm)	Wind speed (X_1) (kmph)	Temperature (X_2) (°C)
1	7	12	22.30
2	6	10	24.50
3	5	8	22.30
4	11	15	21.90
5	13	19	25.60
6	12	22	26.20
7	26	25	27.80
8	11	14	23.80
9	13	18	29.00
10	11	13	27.40

Solution The transpose of independent variable matrix X is given by

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 12 & 10 & 8 & 15 & 19 & 22 & 25 & 14 & 18 & 13 \\ 22.30 & 24.50 & 22.30 & 21.90 & 25.60 & 26.20 & 27.80 & 23.80 & 29.00 & 27.40 \end{bmatrix}^T$$

Similarly, the transpose of dependent variable matrix is given by:

$$Y = \begin{bmatrix} 7 \; 6 \; 5 \; 11 \; 13 \; 12 \; 26 \; 11 \; 13 \; 11 \end{bmatrix}^T$$

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y = \begin{bmatrix} -8.507 \\ 0.882 \\ 0.249 \end{bmatrix}$$

Hence, the relationship is given by

$$E\left(Y|X_1, X_2\right) = -8.507 + 0.882X_1 + 0.249X_2$$

Example 7.3.2
At any given location, the mean annual temperature is estimated from the average elevation (in m) above mean sea level (MSL) and the latitude (°N) by the following model using the data given in Table 7.7.

$$E\left(Y|X_1, X_2\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where Y: mean annual temperature (in °C)
X_1: average elevation (in m) above MSL
X_2: latitude (°N)

Determine the coefficients β_0, β_1, and β_2, respectively.

Solution The independent and dependent variables matrix (X and Y, respectively) is given by,

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 & \ldots & 1 & 1 & 1 & 1 \\ 600 & 587 & 651 & 574 & \ldots & 591 & 601 & 577 & 629 \\ 33.54 & 26.77 & 31.37 & 29.66 & \ldots & 30.66 & 29.58 & 27.51 & 34.28 \end{bmatrix}^T$$

$$Y = \begin{bmatrix} 25.5 \; 30.0 \; 26.7 \; 28.1 \ldots 27.4 \; 28.1 \; 29.5 \; 24.9 \end{bmatrix}^T$$

$$\left(X^T X\right) = \begin{bmatrix} 20.00 & 12058.00 & 597.53 \\ 12058.00 & 7297068.00 & 361654.95 \\ 597.53 & 361654.95 & 18000.43 \end{bmatrix}$$

Table 7.7 Mean annual temperature, average elevation, and latitude for 20 places

Observation no.	Mean annual temperature (°C)	Average elevation (m)	Latitude (°N)
1	25.5	600	33.54
2	30.0	587	26.77
3	26.7	651	31.37
4	28.1	574	29.66
5	29.1	621	27.92
6	26.2	623	32.37
7	26.0	644	32.48
8	25.0	670	33.90
9	24.9	676	33.86
10	28.3	592	29.21
11	28.5	583	29.01
12	29.0	539	28.61
13	30.0	599	26.69
14	30.3	600	26.09
15	31.0	548	25.48
16	29.0	553	28.54
17	27.4	591	30.66
18	28.1	601	29.58
19	29.5	577	27.51
20	24.9	629	34.28

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 13.3815 & -0.0228 & 0.0149 \\ -0.0228 & 0.00007 & -0.0006 \\ 0.0149 & -0.0006 & 0.0131 \end{bmatrix}$$

$$\left(X^T Y\right) = \begin{bmatrix} 557.500 \\ 335081.400 \\ 16553.339 \end{bmatrix}$$

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y = \begin{bmatrix} 50.002 \\ -0.004 \\ -0.651 \end{bmatrix}$$

Hence, the relationship is given by

$$E\left(Y|X_1, X_2\right) = 50 - 0.004X_1 - 0.651X_2$$

Example 7.3.3
For the data presented in Example 7.2.1, the cumulative infiltration depth and time are found to follow second-degree polynomial regression. Develop a curvilinear regression model for predicting cumulative infiltration depth using time as independent variable.

Solution From Example 7.2.1

Cumulative Depth (F_t) (cm)	t (hr)	t^2 (hr^2)
1.30	0.08	0.0064
2.50	0.17	0.0289
4.30	0.33	0.1089
5.75	0.50	0.2500
7.40	0.75	0.5625
8.75	1.00	1.0000
9.90	1.25	1.5625
10.95	1.50	2.2500
11.90	1.75	3.0625
12.85	2.00	4.0000

Taking t as first independent variable and t^2 as second independent variable, the transpose of independent and dependent variable matrix is given by:

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.08 & 0.17 & 0.33 & 0.50 & 0.75 & 1.00 & 1.25 & 1.50 & 1.75 & 2.00 \\ 0.0064 & 0.0289 & 0.1089 & 0.2500 & 0.5625 & 1.0000 & 1.5625 & 2.2500 & 3.0625 & 4.0000 \end{bmatrix}^T$$

$$Y = [1.30 \quad 2.50 \quad 4.30 \quad 5.75 \quad 7.40 \quad 8.75 \quad 9.90 \quad 10.95 \quad 11.90 \quad 12.85]^T$$

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y = \begin{bmatrix} 0.9253 \\ 9.9545 \\ -2.0674 \end{bmatrix}$$

So the regression model is given by

$$\hat{F}_t = 0.9253 + 9.9545t - 2.0674t^2.$$

7.4 Evaluation of Regression Model

After fitting the regression model over the data, the adequacy of the fitted regression model is required to be checked. This can be checked by determining how much of the variability in dependent variable is explained by the regression model. The individual value of observed Y, i.e., y_i can be expressed as sum of three components as,

$$\begin{aligned} y_i &= \overline{Y} + (\hat{y}_i - \overline{Y}) + (y_i - \hat{y}_i) \\ \text{or, } \left(y_i - \overline{Y}\right) &= \left(y_i - \hat{y}_i\right) + \left(\hat{y}_i - \overline{Y}\right) \end{aligned} \tag{7.25}$$

Squaring both sides and summing for all values of Y,

$$\sum \left(y_i - \overline{Y}\right)^2 = \sum \left(y_i - \hat{y}_i\right)^2 + \sum \left(\hat{y}_i - \overline{Y}\right)^2 + 2\sum \left(y_i - \hat{y}_i\right)\left(\hat{y}_i - \overline{Y}\right) \tag{7.26}$$

As $\sum \left(y_i - \hat{y}_i\right)\left(\hat{y}_i - \overline{Y}\right) = 0$ and $\sum \left(y_i - \overline{Y}\right)^2 = \sum y_i^2 - n\overline{Y}^2$, so above equation can be written as,

$$\sum y_i^2 = n\overline{Y}^2 + \sum \left(y_i - \hat{y}_i\right)^2 + \sum \left(\hat{y}_i - \overline{Y}\right)^2 \tag{7.27}$$

Hence, from Eq. 7.27, the total sum of squares of dependent variables ($\sum y_i^2$) can be expressed into three following components:

(a) $n\overline{Y}^2$ – sum of squares due to mean (SSM);
(b) $\sum \left(y_i - \hat{y}_i\right)^2$ – sum of squared errors or regression residual (SSE);
(c) $\sum \left(\hat{y}_i - \overline{Y}\right)^2$ – sum of squares due to regression (SSR).

So, total variability in the dependent variable is the sum of variability explained by regression and variability due to residuals/errors. The adequacy of regression model can be expressed as ratio of variability explained by the regression model $\sum \left(\hat{y}_i - \overline{Y}\right)^2$ and total variability in observed dependent variable $\sum \left(y_i - \overline{Y}\right)^2$. The ratio is called coefficient of determination and represented as r^2 or R^2.

$$\begin{aligned} R^2 &= \frac{\text{Sum of variability in dependent variable explained by regression}}{\text{Total variability in dependent variable}} \\ &= 1 - \frac{\text{Sum of Squared Error}}{\text{Total variability in dependent variable}} \\ &= \frac{\sum \left(\hat{y}_i - \overline{Y}\right)^2}{\sum \left(y_i - \overline{Y}\right)^2} = \frac{\sum \left(a + bx_i - a - b\overline{X}\right)^2}{\sum \left(y_i - \overline{Y}\right)^2} = b^2 \frac{S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}} \end{aligned} \tag{7.28}$$

The ratio of variability explained by the regression model can never be greater than total variability of the dependent variable. Hence, the coefficient of determination ranges between 0 and 1. Closer the R^2 to 1, the better the model is.

In the case of multiple linear regression, coefficient of determination also called coefficient of multiple determination can be calculated using Eq. 7.28. However, with increase in independent variables, the R^2 will automatically and spuriously increase. This may lead to wrong interpretation for the model having large number of independent variables. Hence, R^2 need to be adjusted for increased number of independent variables. The adjusted R^2 is always smaller than R^2 and may be negative also. Adjusted R^2 (R^2_{adj}) is expressed as

$$\begin{aligned} R^2_{adj} &= 1 - \frac{\text{Sum of Squared Error}}{\text{Total variability in dependent variable}} \times \frac{n-1}{n-p-1} \\ &= 1 - \left(1 - R^2\right) \times \frac{n-1}{n-p-1} \end{aligned} \tag{7.29}$$

It should be noted that R^2_{adj}, unlike R^2, does not show measure of fit. Rather, R^2_{adj} is useful for selecting the variables to be included in a MLR model.

Example 7.4.1

Find the coefficient of determination for the linear regression model obtained in Example 7.1.1.

Solution Variability in X is given by $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$ where, $\left(\sum_{i=1}^{n} x_i\right)^2 = (176.5)^2 = 31152.25$, $\sum_{i=1}^{n} x_i^2 = 1410.51$ and $n = 24$. Hence, $S_{xx} = 1410.51 - \frac{31152.25}{24} = 1410.51 - 1298.01 = 112.5$

Similarly, for Y,

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 239.9 - \frac{(63.9)^2}{24} = 68.956$$

$$R^2 = (b)^2 \frac{S_{xx}}{S_{yy}} = (0.664)^2 \frac{112.5}{68.956} = 0.719$$

Coefficient of determination for linear regression model developed in Example 7.1.1 is 0.719, or in other words, the developed model is able to explain 71.90% of variability in dependent variable.

Example 7.4.2

Find the coefficient of determination for the multiple linear regression model obtained in Example 7.3.1.

Solution Sum of squared of error (SSE) = $\sum \left(Y - \hat{Y}\right)^2 = \sum (Y - X\beta)^2 = 70.246$

Total variance in dependent variable = $\sum \left(Y - \overline{Y}\right)^2 = 308.50$

Coefficient of determination = $1 - \frac{\sum\left(Y-\hat{Y}\right)^2}{\sum\left(Y-\overline{Y}\right)^2} = 1 - 70.246/308.50 = 0.7723$

Adjusted coefficient of determination can be calculated using Eq. 7.29.

$$R^2_{adj} = 1 - \left(1 - R^2\right) \times \frac{n-1}{n-p-1} = 1 - (1 - 0.7723)\frac{10-1}{10-2-1} = 0.7072.$$

7.5 Correlation and Regression

Coefficient of correlation is a measure of linear association between dependent and independent variable. Mathematically, sample correlation coefficient is defined as the sum of product of standardized variable divided by $(n-1)$.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\overline{X}}{\sqrt{S_{xx}}}\right)\left(\frac{y_i-\overline{Y}}{\sqrt{S_{yy}}}\right) = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}} = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}\,\frac{S_{xy}}{S_{xx}} = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}b \tag{7.30}$$

where b is least square estimate of slope for simple linear regression model (β). It can be observed that if most of the deviation from mean in either of X or Y is of same sign, then the r will be having positive value. In other words, if both the variables deviate from mean in similar trend (one increases then other also increases and vice versa), then the linear association is high. The correlation coefficient can also become negative if most of the deviation from mean in either of X or Y has opposite signs. The magnitude of r ranges between -1 and 1. Following inferences about the linear relationship between the variables can be drawn based on the value of correlation coefficient.

(i) The magnitude and sign of r represent the strength of linear association and direction of slope of straight line fit between variables.

(ii) A value of r closer to zero represents very weak linear association between the variables involved. In such cases, linear regression may not be able to model the relationship between the variables.

7.6 Correlation and Causality

A high observed correlation does not suggest anything about a cause-and-effect relationship. Hence, the observation that two variables tend to vary simultaneously in the same direction does not imply a direct relationship between them. Both variables may depend upon on unknown variables, and positive correlation is being produced due the mutual relationship with other variables. These unknown variables are called lurking variables. Lurking variables are often overlooked when mistaken claims are made about X causing Y. Hence, the correlation coefficient should not be taken as a measure of relationship or causality. Sometimes, a causal relationship may also exist that is opposite to the observed correlation.

7.7 Confidence Interval

The confidence interval of least square estimates of α and β depends upon the estimate of standard error. The standard error σ^2 is estimated from the deviation of sample points from the estimated least square line. The estimate of σ^2 from a sample is given by standard error of estimate (S_e). Standard error of estimate is the residual sum of squares or the sum of squared errors divided by $n - 2$ and is expressed as,

$$S_e^2 = \frac{1}{n-2}\sum\left(y_i - \hat{y}_i\right)^2 = \frac{S_{yy} - \left(S_{xy}\right)^2 / S_{xx}}{n-2} \qquad (7.31)$$

For studying the statistics for inference about α and β, i.e., the least square estimators of the regression coefficients, two following random variables are defined as,

$$t^a = \frac{a-\alpha}{S_e}\sqrt{\frac{nS_{xx}}{S_{xx} + n\left(\overline{x}\right)^2}} \quad \text{and} \quad t^b = \frac{(b-\beta)}{S_e}\sqrt{S_{xx}} \qquad (7.32)$$

These statistics t^a and t^b follow t distribution with $n - 2$ degrees of freedom. To construct confidence intervals of $(1-\alpha)100\%$ for the regression coefficients α and β, we substitute for the middle term of $-t_{\alpha/2} < t < t_{\alpha/2}$ for the appropriate t statistic, leads us to,

$$-t_{\alpha/2} < \frac{a-\alpha}{S_e}\sqrt{\frac{nS_{xx}}{S_{xx} + n\left(\overline{x}\right)^2}} < t_{\alpha/2} \quad \text{and} \quad -t_{\alpha/2} < \frac{(b-\beta)}{S_e}\sqrt{S_{xx}} < t_{\alpha/2}$$

$$\alpha : a \pm t_{\alpha/2}S_e\sqrt{\frac{1}{n} + \frac{\left(\overline{x}\right)^2}{S_{xx}}} \quad \text{and} \quad \beta : b \pm t_{\alpha/2}S_e\frac{1}{\sqrt{S_{xx}}} \qquad (7.33)$$

The estimate $\hat{y}$ $(= a + bx)$ follows a $t-$distribution with mean $a + bx$, variance $S_e^2\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right)$, and degrees of freedom $n - 2$. Thus, $(1-\alpha)100\%$ confidence interval of the estimated value $(\hat{y})$ is given by,

$$Y\colon (a + bx) \pm t_{\alpha/2}S_e\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}. \qquad (7.34)$$

Example 7.7.1
For the Example 7.1.1, find the 95% confidence interval of the parameter α.

Solution From the solution of Examples 7.1.1 and 7.4.1, $S_{xx} = 112.499$, $S_{yy} = 68.956$, $S_{xy} = 74.689$, $\sum_{i=1}^{n} x_i = 176.5$, and $n = 24$

$$S_e^2 = \frac{S_{yy} - \left(S_{xy}\right)^2 \Big/ S_{xx}}{n-2} = \frac{68.956 - (74.689)^2/112.499}{24-2} = 0.8805$$

$$S_e = \sqrt{0.8805} = 0.9383$$

For $(n-2) = 24 - 2 = 22$ degrees of freedom $t_{0.975} = 2.0739$, so 95% confidence limit for parameter α

$$\alpha : a \pm t_{\alpha/2} S_e \sqrt{\frac{1}{n} + \frac{(\overline{x})^2}{S_{xx}}} = (-2.220) \pm 2.0739 \times 0.9383 \sqrt{\frac{1}{24} + \frac{(176.5/24)^2}{112.499}}$$

Hence, the confidence interval is given by,

$$-3.6265 \leq \alpha \leq -0.8134.$$

Example 7.7.2
For the Example 7.1.1, is parameter β equal to unity at 5% level of significance?

Solution The null and alternate hypothesis can be expressed as

- *Null hypothesis*: $\beta = 1$;
- *Alternative hypothesis*: $\beta \neq 1$;
- *Level of significance* $= 0.05$.

For $(n-2) = 24-2 = 22$ degrees of freedom, $t_{0.975} = 2.0739$ and $t_{0.025} = -2.0739$.
Hence, the critical zone is given by $(\infty, 2.0739] \cup [-2.0739, -\infty)$.
The test statistic is given by

$$t^b = \frac{(b-\beta)}{S_e}\sqrt{S_{xx}} = \frac{(0.664-1)}{0.9383}\sqrt{112.499} = -3.799$$

Since the statistic fall in the critical zone, the null hypothesis must be rejected.

Example 7.7.3
For the Example 7.1.2, check whether the regression line passes through origin at 0.01 level of significance?

Solution If the regression line passes through origin, then its intercept on the y-axis should be 0. Hence,

- *Null hypothesis*: $\alpha = 0$;
- *Alternative hypothesis*: $\alpha \neq 0$;
- *Level of significance* $= 0.01$.

Given (from Example 7.1.2): $n = 15$, $\sum_{i=1}^{n} x_i = 369.3$, $\sum_{i=1}^{n} y_i = 106.9$,
$S_{yy} = \sum_{i=1}^{n} y_{di}^2 = 101.029$, $S_{xx} = \sum_{i=1}^{n} x_{di}^2 = 110.14$ and
$S_{xy} = \sum_{i=1}^{n} x_{di} y_{di} = 105.482$

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{369.3}{15} = 24.62$$

$$s_e = \sqrt{\frac{S_{yy} - \left(S_{xy}\right)^2 / S_{xx}}{n-2}} = \sqrt{\frac{101.029 - (105.482)^2 / 110.144}{15-2}} = 0.030$$

For $(n-2) = 15-2 = 13$ degrees of freedom, $t_{0.995} = 3.0123$ and $t_{0.005} = -3.0123$. Hence, the critical zone is given by $(\infty, 3.0123] \cup [-3.0123, -\infty)$.
The test statistic is given by,

$$t^a = \frac{a-\alpha}{S_e}\sqrt{\frac{nS_{xx}}{S_{xx} + n\left(\overline{x}\right)^2}} = \frac{-16.459-0}{0.030}\sqrt{\frac{15 \times 110.14}{110.14 + 15(24.62)^2}} = -232.46$$

Since the test statistics fall in critical zone, the null hypothesis is rejected at 1% significance level.

7.8 MATLAB Examples

Simple and multiple linear regression can be done in MATLAB using ‘`regress`’ function. The ‘`regress`’ function needs at least two inputs and produces a number of outputs like regression parameters, their confidence interval, etc. The ‘`regress`’ function for solving $Y = X\beta$ (Eq. 7.21) is expressed as:

```
[b,bint,r,rint,stats] = regress(y,X,alpha)
```

Inputs:

`y` – vector of values of dependent variables $(n \times 1)$,
`X` – $(n \times (p+1))$ matrix of n values of p independent variable where first column contains all ones and 2nd to $(p+1)$th column contain values of independent variables,
`alpha` (optional) – level of significance for least square estimates. If user do not provide ‘`alpha`’ then its default value is 0.05.

Outputs:

`b` – vector of regression parameter,
`bint` – confidence interval of regression parameter at given level of significance,
`r` –residual for each value of dependent variable,
`rint` - confidence interval of residual for each value of dependent variable,
`stats` – some statistical measures (R^2), F statistic, p-value, and estimate of standard error about the fitted regression model.

A sample code for solving simple linear regression Example 7.1.1 and related Examples 7.4.1, 7.7.1 and 7.7.2 is shown in Box 7.1. The output of above code is shown in Box 7.2.

Box 7.1 Sample MATLAB code for Example 7.1.1 and related examples

```
clc;close all;clear

%% Inputs
precipitation=[6.9;6.4;6.5;5.1;7.1;7.1;10.2;9.9;8.4;5.8;...
  10.1;7.3;5.5;11.4;10.8;7.5;8.2;7.9;4.1;5;6.7;4.3;10.4;3.9];
runoff=[2.4;1.1;1.7;0.50;1.8;2;4.2;3.0;4.7;3.4;4.4;2.8;...
  1.4;6.3;4.4;1.8;4.2;2.9;0;1.3;1.3;0;5.9;2.4];

%% Scattergram
scatter(precipitation,runoff);box on;
xlabel('Monthly Precipitation (cm)');
ylabel('Monthly Runoff (cm)');
max_val=ceil(max(max(precipitation),max(runoff)));
h=lsline;set(h,'color','r');
legend(h,'least square fit line')
axis([0 max_val 0 max_val]);

%% Regression Fitting
Y=runoff;
X=[ones(size(precipitation,1),1) precipitation];
alpha=0.05;
[b,bint,r,rint,stats] = regress(Y,X,alpha);

%% Display Results
output_file=['output' filesep() 'code_1_result.txt'];
delete(output_file);diary(output_file);diary on;
disp('The regression Parameters:');
fprintf('a = %2.3f and b = %2.3f\n',b(1), b(2));
disp('The confidence Interval of parameters:');
fprintf('a : %2.3f and %2.3f\n',bint(1,1), bint(1,2));
fprintf('b : %2.3f and %2.3f\n',bint(2,1), bint(2,2));
fprintf('Residuals: ');fprintf('%2.2f, ',r);fprintf('\n');
disp('Statistical Measures for the developed model');
fprintf('R^2: %1.3f, \nF Statistics: %1.3f, \np-value: %1.3f, \
    nError Variance estimate: %3.2f \n',stats);
diary off;
```

Box 7.2 Output of sample MATLAB code provided in Box 7.1

```
The regression Parameters:
a = -2.220 and b = 0.664
The confidence Interval of parameters:
a : -3.626 and -0.813
b : 0.480 and 0.847
Residuals: 0.04, -0.93, -0.40, -0.67, -0.69, -0.49, -0.35, -1.35,
    1.34, 1.77, -0.09, 0.17, -0.03, 0.95, -0.55, -0.96, 0.98,
    -0.12, -0.50, 0.20, -0.93, -0.63, 1.22, 2.03,
Statistical Measures for the developed model
R^2: 0.719,
F Statistics: 56.318,
p-value: 0.000,
Error Variance estimate: 0.88
```

As we can see from Box 7.2, the regression model is

$$\hat{Y} = 0.664X - 2.220$$

which we also obtained in Example 7.1.1. Other output of the code can be cross checked with the Examples 7.7.1 and 7.7.2.

Similarly, the regress function can be used for solving examples based on multiple linear regression, e.g., Example 7.3.1. A sample code to solve the Examples 7.3.1 and 7.4.2 is given in Box 7.3. The output of above code is shown in Box 7.4.

Box 7.3 Sample MATLAB code for Example 7.3.1 and associated examples

```
clc;close all;clear

%% Inputs
wind_speed=[12;10;8;15;19;22;25;14;18;13];
temperature
     =[22.30;24.50;22.30;21.90;25.60;26.20;27.80;23.80;29;27.40];
evapotranspiration=[7;6;5;11;13;12;26;11;13;11];

%% Regression Fitting
Y=evapotranspiration;
X=[ones(size(wind_speed)) wind_speed temperature];
alpha=0.05;
[b,bint,r,rint,stats] = regress(Y,X,alpha);

%% Display Results
output_file=['output' filesep() 'code_2_result.txt'];
delete(output_file);diary(output_file);diary on;
disp('The regression Parameters:');
fprintf('%2.2f, ',b);fprintf('\n');
disp('The confidence Interval of parameters:');
disp(bint)
fprintf('Residuals: ');fprintf('%2.2f, ',r);fprintf('\n');
disp('Statistical Measures for the developed model');
fprintf('R^2: %1.3f, \nF Statistics: %1.3f, \np-value: %1.3f, \
     nError Variance estimate: %3.2f \n',stats)
diary off;
```

Box 7.4 Output of sample MATLAB code provided in Box 7.3

```
The regression Parameters:
-8.51, 0.88, 0.25,
The confidence Interval of parameters:
  -35.6356   18.6208
    0.2844    1.4806
   -1.0222    1.5198

Residuals: -0.63, -0.41, 0.90, 0.82, -1.63, -5.43, 5.53, 1.23,
     -1.59, 1.22,
Statistical Measures for the developed model
R^2: 0.772,
F Statistics: 11.871,
p-value: 0.006,
Error Variance estimate: 10.04
```

The developed regression model is

$$E\left(Y|X_1, X_2\right) = -8.51 + 0.88X_1 + 0.25X_2$$

The other outputs of the MATLAB code can be cross-checked from the Examples 7.3.1 and 7.4.2.

Exercise

7.1 In a certain catchment of area 40 km^2, the following rainfall and direct runoff depth over the catchment has been observed for 16 isolated rainfall events.

Rainfall Event no.	Rainfall Depth (mm)	Runoff (mm)	Rainfall Event no.	Rainfall Depth (mm)	Runoff (mm)
1	42.39	13.26	9	47.08	22.91
2	33.48	3.31	10	47.08	18.89
3	47.67	15.17	11	40.89	12.82
4	50.24	15.50	12	37.31	11.58
5	43.28	14.22	13	37.15	15.17
6	52.60	21.20	14	40.38	10.40
7	31.06	7.70	15	45.39	18.02
8	50.02	17.64	16	41.03	16.25

These measurements are made at a culvert present in the downstream of the catchment. Develop a linear regression model taking runoff as dependent and rainfall as independent variable. Using the developed relationship, answer the following:

(a) A precipitation event generated direct runoff of 1.2 Mm^3 from the basin. What is the corresponding rainfall depth? (Ans 6.24 mm)

(b) For a rainfall event of 3 hr with average intensity 12.7 mm/hr, what is the corresponding direct runoff depth? (Ans 11.49 mm)

(c) How much percentage of variance in runoff is being accounted for in the developed model? (Ans 65.94%).

7.2 For a catchment, following observations are made for 24 consecutive months.

Precipitation (mm)	Surface Air Temperature (°C)	Precipitable Water Content (kg/m^2)	Pressure at surface (mb)
0.00	19.11	11.15	964.00
0.03	20.43	12.39	964.21
5.99	21.81	15.54	961.21
4.97	28.66	15.59	960.19
9.50	31.61	20.99	957.69
3.94	34.45	21.41	954.03
145.06	32.41	36.46	951.22
241.36	25.58	49.36	951.11
413.60	24.08	49.66	952.21
216.46	24.14	41.48	956.19
41.37	24.19	31.27	961.91
44.43	22.40	20.99	963.17
0.81	21.58	20.30	964.59
1.94	22.36	18.51	963.70
6.50	24.52	21.08	963.30
1.56	28.55	17.42	960.72
0.57	34.27	23.29	955.75
2.50	36.26	23.74	952.15
67.74	31.10	43.73	950.79
422.32	24.28	53.09	949.85
370.69	23.86	51.05	951.75
237.83	23.75	48.86	955.06
210.00	23.04	39.29	958.84
0.00	20.13	18.09	962.57

Taking precipitation as dependent variable, check:

(a) Which variable between surface air temperature and surface pressure has stronger linear relationship with the precipitation? (Ans Pressure at surface)

(b) Derive a linear regression model between precipitation (Y) and precipitable water content (X), and evaluate its adequacy? (Ans $\hat{Y} = -169.18 + 9.24X$, $R^2 = 0.82$).

7.3 From historical records, for rainfall depth of 15 cm, the runoff generated along with basin characteristics for 10 basins is tabulated below.

Basin Area (km^2)	Length of longest Stream (km)	Drainage Density (km$\big/$km^2)	Generated Runoff (cm)
118.71	20.02	14.42	10.8
92.72	15.22	14.95	8.5
81.14	17.43	14.87	8.4
64.90	8.18	16.31	6.8
58.71	10.17	16.00	6.1
68.20	9.82	16.29	7.0
85.89	19.03	16.50	8.6
73.08	14.60	16.81	8.0
106.66	19.84	13.97	8.3
102.96	18.89	13.86	9.8

(a) Fit a simple regression model for drainage density (Y) using basin area (X) as input. Also, calculate the percentage of variance in drainage density explained by the fitted simple linear regression model. (Ans $\hat{Y} = 19.22 - 0.045X$, $R^2 = 0.64$)

(b) The generated runoff from the basin area is expected to follow power relationship as expressed below,

$$Q = JA^b$$

where Q and A are runoff and basin area, respectively, and J and b are model parameters. Fit a curvilinear regression model for generated runoff using basin area as input. (Ans $J = 0.41$ and $b = 0.674$).

7.4 For the infiltration data given in Example 7.2.1, fit a Philip two-term model. The model is expressed as:

$$f_t = \frac{S}{2\sqrt{t}} + A$$

where f_t is the infiltration capacity at time t. S and A are the model parameters. Also, find the coefficient of determination for the developed model. (Ans $S = 9.59$, $A = 0.89$, $R^2 = 0.93$).

7.5 For the data presented in Exercise 7.2, develop an MLR model taking precipitation as dependent variable (Y) and all other variables as independent (X_1: surface air temperature, X_2: precipitable water content and X_3: pressure at surface). Calculate the coefficient of determination and SSE for the developed model. Compare the developed MLR model with the SLR model developed in Exercise 7.2 (in terms of goodness-of-fit). Is the inclusion of extra variables justified?

(Ans $\hat{Y} = 14103.7 - 13.77X_1 + 5.30X_2 - 14.41X_3$
$R^2 = 0.89$, $\sum \varepsilon_i^2 = 51830.6$
MLR model is fitting better than SLR model fitted in Exercise 7.2.
$R^2_{adj} = 0.87$ is higher as compared to Exercise 7.2, so inclusion of extra variables is justified.)

7.6 Develop an MLR model for predicting direct runoff (Y) by using length of longest stream (X_1), drainage density (X_2) as inputs for the data presented in Exercise 7.3. Calculate the coefficient of determination and SSE for the developed model. (Ans $\hat{Y} = 7.54 + 0.23X_1 - 0.19X_2$, $R^2 = 0.73$, $\sum \varepsilon_i^2 = 4.66$).

7.7 Air temperature and evaporation for a water body are recorded for 20 consecutive summer days at a location.

Days	Temperature (°C)	Evaporation (mm/day)
1	25.64	3.4*
2	32.67	10.6
3	31.71	10.2
4	32.15	11.2
5	31.42	10.4
6	29.19	8.9*
7	28.91	2.1*
8	33.16	11.7
9	28.17	4.2*
10	33.15	10.9
11	32.89	10.6
12	33.52	12.3
13	30.06	4.5*
14	31.36	11
15	34.51	13
16	29.44	6.8*
17	25.60	3.6*
18	33.67	10.9
19	31.31	9.8
20	28.55	4.7*

Evaporation (Y) is dependent upon the temperature (X). Develop a SLR model for the data and check if

(a) The regression line between the evaporation and the temperature has a slope of 45° at 5% level of significance.
(b) The intercept of regression line is 0 mm at 1% level of significance.
(c) The evaporation corresponding to 25° C air temperature is 0 mm at 99% confidence interval.

(Ans $\hat{Y} = 1.21X - 28.70$.

(a) At $\alpha = 0.05$, the regression line has a slope of 45°.
(b) At $\alpha = 0.01$, the intercept is not 0 mm.
(c) Yes, the evaporation corresponding to 25° C air temperature is 0 mm at 99% confidence interval.)

7.8 For the data presented in Exercise 7.7, the rows having * show the cloudy days. Develop an SLR model using data from cloudy days only, considering daily evaporation as dependent (Y) and daily air temperature as independent variable (X). Check that the slope and intercept of developed SLR model differ from the value of slope and intercept, respectively, of SLR model developed in Exercise 7.7 at 5% level of significance. Comment on the statement that "on cloudy days evaporation rate with respect to air temperature is lower compared to average/normal condition".

(Ans $\hat{Y} = 0.54X - 10.44$. At $\alpha = 0.05$ the slope and intercept of developed SLR model is not different compared to model developed in Exercise 7.7. Hence, on cloudy days, evaporation rate with respect to air temperature is not lower compared to average/normal condition.)

Chapter 8
Multivariate Analysis

Often many hydroclimatic variables are associated with each other and such associations are complex. Many a times several hydroclimatic variables are required to be analyzed simultaneously. Several techniques related to multiple hydroclimatic variables are discussed in this chapter. Different techniques include principal component analysis, supervised principal component analysis, canonical correlation analysis, empirical orthogonal function, one-way and two-way analysis of variance. All these techniques are explained in this chapter with illustrative examples.

8.1 Principal Component Analysis

Principal component analysis (PCA) is the transformation of p correlated variables into p uncorrelated orthogonal components through their linear combination. The resulting uncorrelated orthogonal components are known as principal components (PCs). Most often, a set of hydrologic or hydroclimatic variables (used as input for another target variable) may be significantly correlated with each other. This implies that information available in one variable may also be partially available from other variables. In general, the objective of PCA is data compression, in such a way that resulting PCs are uncorrelated to each other and total variance of the original data is redistributed. The 1st PC contains maximum amount of variance, and variance gradually decreases for the subsequent components. In hydroclimatology, the PCA is used for either dimensionality reduction or identification of covariance structure. PCA reduces the dimensionality as the first few components explain most of the variance of the original data set. The PCA for p variables tries to reorient the pth dimensional space (or Cartesian orthogonal coordinate system) to satisfy the aforementioned redistribution of variance. The transformation of axes is geometrically illustrated for two variables in Fig. 8.1. For a set of p variables, $X = [X_1, X_2, \ldots, X_p]$ each having n observations, the set of principal components $[Z_1, Z_2, \ldots, Z_p]$ are:

R. Maity, *Statistical Methods in Hydrology and Hydroclimatology*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-10-8779-0_8

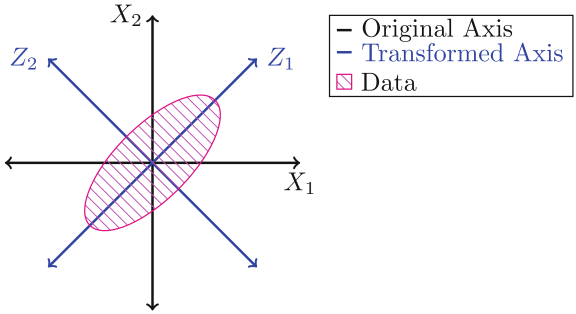


Fig. 8.1 Projection of axis in PCA (Z_1 and Z_2 are principal components of X_1 and X_2)

$$\begin{aligned} Z_1 &= Xa_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ Z_2 &= Xa_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ Z_p &= Xa_p = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned} \tag{8.1}$$

where a_i ($i = 1, 2, \ldots, p$) is a $p \times 1$ vector, $[a_{i1}, a_{i2}, \ldots a_{ip}]^T$, known as loading vector (also called projection or transformation vector) for ith principal component (Z_i). Geometrically, the loading vector shows the direction of orientation of the PC axis. Being a direction vector, its magnitude, i.e., sum of squared terms given by $a_i^T a_i$ is 1. If $U = [a_1, a_2, \ldots, a_p]$ is the orthogonal projection matrix, the PC matrix can be expressed as:

$$Z = XU \tag{8.2}$$

PCs being uncorrelated, the covariance of any two different principal components is zero. From Eq. 8.1, the variance of the PCs can be calculated as:

$$\text{Var}(Z_i) = \text{Var}(Xa_i) = a_i^T \text{Cov}(X) a_i = a_i^T S_x a_i \tag{8.3}$$

where S_x represents the covariance matrix of X.

8.1.1 Determination of Principal Components

PCs can be determined by maximizing the variance of ith principal component with the constraint that sum of square of loadings is unity. This optimization problem can be expressed as

$$\text{maximize } \mathrm{Var}(Z_i) \text{ or, maximize } a_i^T S_x a_i \tag{8.4}$$
$$\text{subjected to } a_i^T a_i = 1$$

Further, the estimated ith and jth principal component should be such that $a_i^T S_x a_i \geq a_j^T S_x a_j$ for $i < j$. The optimization problem can be solved by using the method of Lagrange multiplier. Method of Lagrange multiplier is an optimization technique to find maxima/minima for a function subjected to equality constraint. If the optimization problem is expressed as:

$$\text{maximize } f(x) \tag{8.5}$$
$$\text{subjected to } g(x) = 0$$

then the Lagrangian function with Lagrange multiplier λ is given by:

$$L(x, \lambda) = f(x) - \lambda g(x) \tag{8.6}$$

The solution of the optimization problem is given by:

$$\frac{dL}{dx} = 0 \tag{8.7}$$

Hence, the solution of Eq. 8.4 is given by:

$$L = a_i^T S_x a_i - \lambda(a_i^T a_i - 1) \tag{8.8}$$
$$\frac{dL}{da_i} = 0$$
$$\text{or, } (S_x - \lambda I)a_i = 0 \tag{8.9}$$

The characteristic equation is obtained by taking determinant of the above equation.

$$|S_x - \lambda I| = 0 \tag{8.10}$$

The roots obtained from the above equation (λ_i, for $i \in \{1, 2, \ldots, p\}$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$) are eigenvalues of covariance matrix of X. The variance of ith principal component is given by λ_i. Hence, by arranging the eigenvalues in descending order, one can ensure that the variance of ith principal component is more compared to jth principal component if $i < j$. The principal component loading vector for ith principal component (a_i) is given by:

$$(S_x - \lambda_i I)a_i = 0 \qquad\qquad \text{Such that } a_i^T a_i = 1 \tag{8.11}$$

Thus, the principal component is obtained using following expression:

$$Z_i = Xa_i \tag{8.12}$$

where Z_i is the principal component with variance λ_i. According to properties of covariance matrix, the trace of matrix (sum of diagonal elements) is the total variance of all the variables in X variable $(X_1, X_2, \ldots, X_p)$. Moreover, from the properties of matrix, sum of all eigenvalues is equal to the trace of matrix. Hence, the variance explained by ith principal component is given by:

$$\text{Variance explained by } Z_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} = \frac{\lambda_i}{tr(S_x)} \tag{8.13}$$

Similar expression can be obtained for the principal components from the correlation matrix. However, the principal components obtained from correlation and covariance matrix are not same.

Example 8.1.1

Calculate the principal component loading vectors using covariance matrix of observed monthly hydroclimatic data given in Table A.1 (p. 429). Also, find the variance explained by each principal component.

Solution From the data given in Table A.1, suppose that $X = [X_1, X_2, \ldots, X_9]$ represent the data such that X_1 represents precipitation, X_2 represents surface air temperature and so on. The covariance matrix of X is given by:

$$\text{Cov}(X) = S_x = \begin{bmatrix} 21878.87 & -73.79 & 2031.88 & \ldots & -4463.97 & 199.33 & -18.61 \\ -73.79 & 23.27 & 4.14 & \ldots & -114.67 & 4.86 & 0.66 \\ 2031.88 & 4.14 & 235.45 & \ldots & -570.38 & 22.14 & 0.01 \\ -491.46 & -14.84 & -63.88 & \ldots & 226.8 & -9.38 & -0.49 \\ -105.98 & 22.95 & -0.10 & \ldots & -103.39 & 4.44 & 0.58 \\ 673.09 & -1.95 & 77.12 & \ldots & -168.49 & 5.93 & -0.36 \\ -4463.97 & -114.67 & -570.38 & \ldots & 1951.57 & -80.89 & -4.13 \\ 199.33 & 4.86 & 22.14 & \ldots & -80.89 & 4.54 & 0.17 \\ -18.61 & 0.66 & 0.01 & \ldots & -4.13 & 0.17 & 0.71 \end{bmatrix}_{9\times 9}$$

Hence, the characteristic equation for calculating eigenvalues of S_x is given by:

$$|S_x - \lambda I| = 0$$

$$\text{or,}\quad \begin{bmatrix} 21878.87-\lambda & -73.79 & 2031.88 & \cdots & 199.33 & -18.61 \\ -73.79 & 23.27-\lambda & 4.14 & \cdots & 4.86 & 0.66 \\ 2031.88 & 4.14 & 235.45-\lambda & \cdots & 22.14 & 0.01 \\ -491.46 & -14.84 & -63.88 & \cdots & -9.38 & -0.49 \\ -105.98 & 22.95 & -0.10 & \cdots & 4.44 & 0.58 \\ 673.09 & -1.95 & 77.12 & \cdots & 5.93 & -0.36 \\ -4463.97 & -114.67 & -570.38 & \cdots & -80.89 & -4.13 \\ 199.33 & 4.86 & 22.14 & \cdots & 4.54-\lambda & 0.17 \\ -18.61 & 0.66 & 0.01 & \cdots & 0.17 & 0.71-\lambda \end{bmatrix} = 0$$

or, $\lambda = 23063.59, 1062.88, 35.89, 5.79, 1.06, 0.63, 0.17, 0.02$ and 0.01.

Hence, loading vector corresponding to first principal component (a_1) is given by:

$$(S_x - \lambda_1 I)a_1 = 0 \qquad \text{subjected to } a_1^T a_1 = 1$$

$$\text{or,}\quad \begin{bmatrix} -1184.72 & -73.79 & \cdots & 199.33 & -18.61 \\ -73.79 & -23040.32 & \cdots & 4.86 & 0.66 \\ 2031.88 & 4.14 & \cdots & 22.14 & 0.01 \\ -491.46 & -14.84 & \cdots & -9.38 & -0.49 \\ -105.98 & 22.95 & \cdots & 4.44 & 0.58 \\ 673.09 & -1.95 & \cdots & 5.93 & -0.36 \\ -4463.97 & -114.67 & \cdots & -80.89 & -4.13 \\ 199.33 & 4.86 & \cdots & -23059.05 & 0.17 \\ -18.61 & 0.66 & \cdots & 0.17 & -23062.88 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \\ a_{14} \\ a_{15} \\ a_{16} \\ a_{17} \\ a_{18} \\ a_{19} \end{bmatrix} = 0$$

$$\text{or, } a_1 = [0.97 \quad 0 \quad 0.09 \quad -0.02 \quad 0 \quad 0.03 \quad -0.21 \quad 0.01 \quad 0]^T$$

Similarly,

$$U = \begin{bmatrix} 0.97 & 0.22 & -0.08 & 0 & -0.01 & 0.01 & 0 & 0 & 0 \\ 0 & -0.12 & -0.36 & 0.59 & 0.11 & 0.05 & -0.07 & 0.69 & -0.14 \\ 0.09 & -0.13 & 0.77 & 0.43 & 0.32 & 0.01 & -0.29 & -0.06 & 0.02 \\ -0.02 & 0.12 & 0.04 & -0.06 & 0.01 & 0 & -0.03 & 0.29 & 0.95 \\ 0 & -0.12 & -0.38 & 0.58 & 0.05 & -0.04 & 0.11 & -0.65 & 0.27 \\ 0.03 & -0.02 & 0.33 & 0.23 & -0.42 & -0.19 & 0.78 & 0.13 & 0 \\ -0.21 & 0.94 & 0.03 & 0.22 & 0.08 & -0.02 & 0.01 & -0.03 & -0.1 \\ 0.01 & -0.04 & -0.08 & -0.17 & 0.76 & -0.49 & 0.37 & 0.05 & -0.02 \\ 0 & -0.01 & 0.02 & -0.05 & 0.34 & 0.85 & 0.4 & -0.02 & 0.01 \end{bmatrix}$$

The variance explained by first principal component $= \lambda_1 \big/ \sum \lambda = 0.954$.

Similarly, the variance explained by second and third principal components is 0.044 and 0.001, respectively. For all other principal components, the explained variance is negligible.

Example 8.1.2

Calculate the loading vectors of principal components using correlation matrix for the data set used in the last example. Are the loading vectors obtained same as last example? Also, find the variance explained by each of the principal components.

Solution Correlation matrix can be calculated from covariance matrix using the following equation:

$$C_{i,j} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}}$$

where C_{ij} and S_{ij} represent the elements of correlation and covariance matrix in ith row and jth column. Hence, the correlation matrix ($\mathrm{Cor}(X)$ or C_x) is given by:

$$C_x = \begin{bmatrix} 1 & -0.103 & 0.895 & -0.645 & -0.15 & 0.886 & -0.683 & 0.633 & -0.149 \\ -0.103 & 1 & 0.056 & -0.598 & 0.997 & -0.079 & -0.538 & 0.473 & 0.161 \\ 0.895 & 0.056 & 1 & -0.809 & -0.001 & 0.979 & -0.841 & 0.677 & 0.001 \\ -0.645 & -0.598 & -0.809 & 1 & -0.553 & -0.705 & 0.997 & -0.855 & -0.114 \\ -0.15 & 0.997 & -0.001 & -0.553 & 1 & -0.133 & -0.491 & 0.437 & 0.143 \\ 0.886 & -0.079 & 0.979 & -0.705 & -0.133 & 1 & -0.743 & 0.542 & -0.082 \\ -0.683 & -0.538 & -0.841 & 0.997 & -0.491 & -0.743 & 1 & -0.860 & -0.111 \\ 0.633 & 0.473 & 0.677 & -0.855 & 0.437 & 0.542 & -0.860 & 1 & 0.093 \\ -0.149 & 0.161 & 0.001 & -0.114 & 0.143 & -0.082 & -0.111 & 0.093 & 1 \end{bmatrix}$$

Corresponding characteristic equation for eigenvalues are given by:

$$|C_x - \lambda I| = 0$$

$$\text{or,} \begin{vmatrix} 1-\lambda & -0.103 & 0.895 & -0.645 & -0.15 & 0.886 & -0.683 & 0.633 & -0.149 \\ -0.103 & 1-\lambda & 0.056 & -0.598 & 0.997 & -0.079 & -0.538 & 0.473 & 0.161 \\ 0.895 & 0.056 & 1-\lambda & -0.809 & -0.001 & 0.979 & -0.841 & 0.677 & 0.001 \\ -0.645 & -0.598 & -0.809 & 1-\lambda & -0.553 & -0.705 & 0.997 & -0.855 & -0.114 \\ -0.15 & 0.997 & -0.001 & -0.553 & 1-\lambda & -0.133 & -0.491 & 0.437 & 0.143 \\ 0.886 & -0.079 & 0.979 & -0.705 & -0.133 & 1-\lambda & -0.743 & 0.542 & -0.082 \\ -0.683 & -0.538 & -0.841 & 0.997 & -0.491 & -0.743 & 1-\lambda & -0.860 & -0.111 \\ 0.633 & 0.473 & 0.677 & -0.855 & 0.437 & 0.542 & -0.860 & 1-\lambda & 0.093 \\ -0.149 & 0.161 & 0.001 & -0.114 & 0.143 & -0.082 & -0.111 & 0.093 & 1-\lambda \end{vmatrix} = 0$$

or, $\lambda = 5.139, 2.453, 0.958, 0.302, 0.109, 0.032, 0.004, 0.000$ and 0.000.

The corresponding loading vectors are given by

$$a_1 = [\,-0.349 \quad -0.195 \quad -0.398 \quad 0.430 \quad -0.173 \quad -0.361 \quad 0.433 \quad -0.386 \quad -0.026\,]^T$$

$$U = \begin{bmatrix} -0.349 & -0.341 & -0.047 & 0.191 & 0.831 & -0.178 & -0.027 & 0.014 & 0.006 \\ -0.195 & 0.559 & -0.146 & -0.175 & 0.220 & 0.209 & 0.194 & 0.68 & -0.075 \\ -0.398 & -0.252 & 0.073 & -0.235 & -0.115 & 0.327 & 0.739 & -0.222 & 0.036 \\ 0.430 & -0.123 & 0.000 & 0.106 & 0.195 & 0.447 & 0.055 & 0.116 & 0.731 \\ -0.173 & 0.572 & -0.169 & -0.155 & 0.234 & 0.194 & -0.187 & -0.661 & 0.159 \\ -0.361 & -0.333 & 0.030 & -0.407 & -0.122 & 0.434 & -0.605 & 0.151 & 0.003 \\ 0.433 & -0.082 & -0.019 & 0.080 & 0.252 & 0.535 & 0.032 & -0.119 & -0.659 \\ -0.386 & 0.098 & 0.002 & 0.819 & -0.244 & 0.323 & -0.087 & 0.025 & -0.005 \\ -0.026 & 0.194 & 0.970 & -0.014 & 0.132 & 0.028 & -0.041 & -0.002 & 0.001 \end{bmatrix}$$

These loading vectors are different from the loading vectors obtained using covariance matrix.

Variance explained by 1st principal component is $\lambda_1 \big/ \sum \lambda = 0.571$.

Similarly, the variance explained by next five principal components is 0.272, 0.106, 0.033, 0.012, and 0.003, respectively. The variance explained by last two principal components is negligible, as evident from corresponding eigenvalues. It should be noted that the variance explained by different principal components is different from previous example.

8.2 Supervised Principal Component Analysis

Supervised principal component analysis (SPCA) is a technique to find the linear combination of independent variables that leads to maximum correlation with target or response variable. SPCA differs from PCA by the fact that there is no target variable involved in PCA; hence, the obtained loading vector maximizes the individual variance of principal components. However, in SPCA the loading vectors are such that they maximize the association with the target variables. Hence, SPCA is more useful in studies which try to establish relationship between two data sets. Interestingly, PCA can be considered as one of the special cases of SPCA where target variable is identity matrix. SPCA has similar loading equation as PCA (Eq. 8.1). However, the procedure to estimate loading vector (a_i) differs as it takes target variable data set in consideration.

Let us assume that there are p independent variables (X_i, $i \in \{1, 2, \ldots, p\}$) and l dependent or response variables (Y_j, $j \in \{1, 2, \ldots, l\}$) each having n observations with individual mean of zero. Let $X = [X_1, X_2, \ldots, X_p]^T$ and $Y = [Y_1, Y_2, \ldots, Y_l]^T$. Hence, the matrices X and Y are of sizes $p \times n$ and $l \times n$, respectively. Then, the correlation between X and Y as per Hilbert–Schmidt Independence Criterion (HSIC) is given by $(n-1)^2 tr(KHLH)$, where K and L are kernel for $U^T X$ and Y, respectively; i.e., $K = XUU^T X$, and $L = Y^T Y$, and H is a centering matrix

($H = I - n^{-1}ee^T$, where I is identity matrix of order n and e is an all-ones matrix of size $1 \times n$). The details about HSIC and Hilbert spaces can be found elsewhere. Hence, the SPCA can be obtained by solving the following optimization problem:

$$\begin{aligned} &\text{maximize } tr(KHLH) && \text{Subjected to } U^TU = 1 \\ &\text{or, maximize } tr(U^TXHLHX^TU) && \text{Subjected to } U^TU = 1 \end{aligned} \tag{8.14}$$

The difference between SPCA and PCA is evident from the above equation. In PCA, the covariance matrix of X is used for optimization; however, in SPCA, the covariance of XHY^T is optimized. Similar to PCA, the above-stated optimization problem can be solved using Lagrangian multiplier discussed in Sect. 8.1.1. The Lagrangian function for optimization problem using the Lagrangian multiplier λ is given by:

$$L(X, Y, U, \lambda) = tr(U^TXHLHX^TU) - \lambda(U^TU - 1) \tag{8.15}$$

After maximizing the Lagrangian function (differentiation with respect to U and equating it to zero), the following expression is obtained:

$$(XHLHX^T - \lambda I)U = 0 \tag{8.16}$$

The characteristic equation is given by:

$$|XHLHX^T - \lambda I| = 0 \tag{8.17}$$

Corresponding to p roots of above equation (λ_i such that $i \in \{1, 2, \ldots, p\}$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$), p projection or loading vectors (a_i such that $i \in \{1, 2, \ldots, p\}$) can be calculated using the following relationship:

$$(XHLHX^T - \lambda_i I)a_i = 0 \qquad \text{Subjected to } a_i^Ta_i = 1 \tag{8.18}$$

Using the ith loading or projection vector, the ith supervised principal component (Z_i) is given by:

$$Z_i = a_i^TX = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p \tag{8.19}$$

Example 8.2.1
From the Table A.1 (p. 429), considering the monthly precipitation as target variable and other variables as independent variables and calculate the SPCA loading vectors and corresponding Supervised Principal Component (SPC).

Solution Let Y be the monthly precipitation, and $X_1, X_2 \ldots, X_8$ are independent variables; hence, $X = [X_1,\ X_2,\ \ldots,\ X_8]^T$. The characteristic equation for calculation of SPCA is given by:

$$|XHLHX^T - \lambda I| = 0$$

where $L = Y^T Y$, and $H = I - n^{-1}ee^T = I - \frac{1}{24}ee^T = I - 0.0417ee^T$

Hence,

$$XHLHX^T = \begin{bmatrix} 0.29 & -7.93 & 1.92 & \dots & 17.43 & -0.78 & 0.07 \\ -7.93 & 218.40 & -52.83 & \dots & -479.82 & 21.42 & -2.00 \\ 1.92 & -52.83 & 12.78 & \dots & 116.06 & -5.18 & 0.48 \\ 0.41 & -11.39 & 2.76 & \dots & 25.03 & -1.12 & 0.10 \\ -2.63 & 72.35 & -17.50 & \dots & -158.95 & 7.10 & -0.66 \\ 17.43 & -479.82 & 116.06 & \dots & 1054.14 & -47.07 & 4.39 \\ -0.78 & 21.42 & -5.18 & \dots & -47.07 & 2.10 & -0.20 \\ 0.07 & -2.00 & 0.48 & \dots & 4.39 & -0.20 & 0.02 \end{bmatrix}_{8\times 8} \times 10^7$$

The first eigen value (λ_1) is 13.123×10^9 and all other eigenvalues are insignificant.

As the number of response variable is one and only the first eigenvalue is significant, so, only one SPC is selected. The corresponding SPC loading vector is calculated by Eq. 8.18.

$$(XHLHX^T - \lambda_1 I)a_1 = 0$$

$$\text{or}, a_1 = [\,-0.148 \quad 0.407 \quad -0.098 \quad -0.021 \quad 0.135 \quad -0.896 \quad 0.040 \quad -0.003\,]$$

Corresponding SPC is calculated using Eq. 8.19.

$$Z_1 = a_1^T X$$

Hence, Z_1 = [−807.04, −798.28, −786.16, −758.74, −731.69, −692.21, −691.27, −701.56, −737.46, −778.75, −805.80, −803.13, −810.25, −802.53, −787.62, −758.28, −723.20, −688.51, −694.76, −694.19, −734.85, −780.34, −795.36, −820.96].

8.3 Dimensionality Reduction using PCA and SPCA

PCA and SPCA can be used for dimensionality reduction. Many a times, a threshold for explained variance is used for the PCA and SPCA analysis. Minimum number of principal components is selected that explain the threshold variance. In practice, for many studies, first two or three PCs are enough to explain most of the variance required for analysis. Similarly, the first l SPCs may found to be enough if target variable (Y) is of size ($l \times n$). Some of the techniques used to select the number of PC are as follows:

(i) **Total variance explained criteria**: Depending upon the prediction problem and accuracy of the data measurement, a threshold cumulative percentage of total variance can be selected (say V_T). First, k PCs are selected if they are able to explain at least the threshold amount of total variance as expressed below:

$$\text{Select the minimum value of } k \text{ for which } \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{i=1}^{p} \lambda_i} \geq V_T \tag{8.20}$$

(ii) **Average eigen value criteria**: First k PCs corresponding to which the eigen value is above mean eigenvalue are selected.

$$\text{Select } k \text{ PC, if } \lambda_k \geq \overline{\lambda} \text{ and } \lambda_{k+1} < \overline{\lambda} \tag{8.21}$$

(iii) **Scree plot**: Plot between PCs and variance explained by each PC is called scree plot. The first ith PCs are selected for which scree plot shows significant slope.

Based on the scree plot, a hypothesis test can be used for selecting PCs. If k is number of selected PCs, then the null hypothesis is the equality of all remaining eigen vectors (as they are measure of variance explained by any PC). The test statistics is given by:

$$D = n\left[(p-k)\ln(\overline{\lambda}_k) - \sum_{j=k+1}^{p} \ln(\lambda_j)\right] \tag{8.22}$$

where $\overline{\lambda}_k = \frac{\sum\limits_{j=k+1}^{p} \lambda_j}{p-k}$. The test statistics D follows χ^2 distribution with $0.5(p-k-1)(p-k+2)$ degrees of freedom. For a hypothesis test at α level of significance, the null hypothesis is rejected if $D > \chi^2_{(\alpha)}(0.5(p-k-1)(p-k+2))$.

Example 8.3.1
For the Examples 8.1.1 and 8.1.2, select the minimum number of principal components required to explain 95% variance of the data.

Solution From the Example 8.1.1, the variance explained by first principal component is 95.4%, so only one principal component is enough for explaining 95% variance. From the Example 8.1.2, cumulative variance explained by different principal components is given by:

Principal component	Variance explained	Cumulative variance explained
1st	0.571	0.571
2nd	0.272	0.843
3rd	0.106	0.949
4th	0.033	0.982
5th	0.012	0.994
6th	0.003	0.997

It should be noted that the last two principal components are insignificant and are only explaining 0.3% of total variance. From the table, for explaining 95% of variance, first four principal components are enough.

Example 8.3.2
For a data set having 40 observations and 8 variables, the eigenvalues of its covariance matrix are given by 20.75, 13.88, 7.57, 1.07, 1.02, 0.93, 0.87, and 0.71. Check whether the last five eigenvalues differ significantly at 5% level of significance.

Solution According to question,

$$\lambda_1 = 20.75,\ \lambda_2 = 13.88,\ \lambda_3 = 7.57,\ \lambda_4 = 1.07,\ \lambda_5 = 1.02,\ \lambda_6 = 0.93,\ \lambda_7 = 0.87$$
$$\text{and } \lambda_8 = 0.71.$$

Null Hypothesis: Last five eigenvalues are equal, i.e., $(\lambda_4 = \lambda_5 = \lambda_6 = \lambda_7 = \lambda_8)$
Alternative Hypothesis: At least one eigen value out of last five is not equal to other.
Level of Significance: $\alpha = 5\%$.

The test statistics is given by Eq. 8.22. For Eq. 8.22, p=8 and $k = 8 - 5 = 3$.

$$\overline{\lambda}_3 = \frac{\sum\limits_{j=k+1}^{p} \lambda_j}{p-k} = \frac{\sum\limits_{j=4}^{8} \lambda_j}{8-3} = 0.92$$

$$\sum_{j=k+1}^{p} \ln(\lambda_j) = \sum_{j=4}^{8} \ln(\lambda_j) = -0.47$$

$$\begin{aligned} D = n\left[(p-k)\ln(\overline{\lambda}_k) - \sum_{j=k+1}^{p} \ln(\lambda_j)\right] &= 40\left[(8-3)\ln(\overline{\lambda}_3) - \sum_{j=4}^{8} \ln(\lambda_j)\right] \\ &= 40(5\ln(0.92) + 0.47) \\ &= 2.12 \end{aligned}$$

Test statistics D is supposed to follow χ^2 distribution with $0.5(p-k-1)(p-k+2) = 14$ degrees of freedom.

$$\chi^2_\alpha(0.5(p-k-1)(p-k+2)) = \chi^2_{0.05}(14) = 23.68$$

As $D < 23.68(\chi^2_{0.05}(14))$; hence, null hypothesis of the eigenvalues being equal is accepted.

8.4 Canonical Correlation Analysis

Canonical correlation analysis is a procedure to find a linear combination of two different set of variables X and Y such that their correlation is maximum. Suppose that X is $n \times p_1$ and Y is $n \times p_2$ where n is the number of observations and $p_2 < p_1$. Further, suppose $Z = [X \; Y]$ with variance S_z. The variance S_z can be partitioned into variance of X (S_{xx}), variance of Y (S_{yy}), and covariance of X and Y (S_{xy}).

$$S_z = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix} \tag{8.23}$$

where S_{xx} and S_{yy} are of size $p_1 \times p_1$ and $p_2 \times p_2$. S_{xy} and S_{yx} are of size $p_1 \times p_2$ and $p_2 \times p_1$, respectively. Suppose that a_i and b_i are linear transformation vector (with variance 1) for X and Y, respectively, which result in series U_i and V_i having maximum correlation. The correlation between the U_i and V_i is given by

$$\mathrm{Cor}(U_i, V_i) = \frac{\mathrm{Cov}(U_i, V_i)}{\sqrt{\mathrm{Var}(U_i)\mathrm{Var}(V_i)}} = \frac{a_i^T S_{xy} b_i}{\sqrt{a_i^T S_{xx} a_i b_i^T S_{yy} b_i}} \tag{8.24}$$

As correlation can be negative, so the square of correlation (Γ) is required to be optimized for highest value. Using the technique of Lagrange multiplier (Sect. 8.1.1), the characteristic equation of this optimization problem is written as:

$$|\Gamma - \lambda I| = 0 \tag{8.25}$$

where $\Gamma = S_{yy}^{-1} S_{xy}^T S_{xx}^{-1} S_{xy}$ and λ is Lagrange multiplier. The transformation vector b_i for transforming Y to V_i is given by:

$$(\Gamma - \lambda I) b_i = 0 \tag{8.26}$$

The corresponding transformation vector a_i can be found as:

$$a_i = \frac{S_{xx}^{-1} S_{xy} b_i S_{yx}}{\sqrt{\lambda_i}} \tag{8.27}$$

Example 8.4.1
Holehonnur town is situated 50 km downstream of Bhadra Reservoir. The temperature in Holehonnur town is assumed to be affected by the air temperature at Bhadra Reservoir, as it is a major water body in vicinity. Using the data provided in Table A.3 (p. 432), calculate the canonical correlation loading vectors by considering the minimum and maximum temperature for Holehonnur town as target variable and observed temperature at Bhadra Reservoir (both minimum and maximum) as independent variable.

Solution Assume minimum and maximum temperature at Bhadra reservoir as X and temperature at Holehonnur town as Y. Further, $Z = [XY]$, and the covariance of Z is given by:

$$S_z = \begin{bmatrix} 5.54 & -0.11 & -0.62 & 1.65 \\ -0.11 & 2.95 & -3.50 & -0.45 \\ -0.62 & -3.50 & 5.71 & 0.25 \\ 1.65 & -0.45 & 0.25 & 2.94 \end{bmatrix}$$

Hence,

$$S_{xx} = \begin{bmatrix} 5.54 & -0.11 \\ -0.11 & 2.95 \end{bmatrix} \text{ and } S_{yy} = \begin{bmatrix} 5.71 & 0.25 \\ 0.25 & 2.94 \end{bmatrix}$$

Similarly, as the S_z is symmetrical so $S_{xy}^T = S_{yx} = \begin{bmatrix} -0.62 & -3.50 \\ 1.65 & -0.45 \end{bmatrix}$.

Loading vector for Y (b_i) can be found by using Eqs. 8.25 and 8.26.

$$\Gamma = S_{yy}^{-1} S_{xy}^T S_{xx}^{-1} S_{xy} = \begin{bmatrix} 0.74 & 0.04 \\ 0.04 & 0.18 \end{bmatrix}$$

The corresponding eigenvalues are 0.75 and 0.18. The loading vectors are given by column of matrix $B = \begin{bmatrix} 0.997 & -0.084 \\ 0.079 & 0.996 \end{bmatrix}$. The loading vectors for X can be calculated by Eq. 8.27 as column of matrix $A = \begin{bmatrix} 0.924 & 0.007 \\ 0.381 & 0.999 \end{bmatrix}$.

8.5 Empirical Orthogonal Function

Hydroclimatic data, apart from time-variability, also has spatial variation. PCA can also be utilized for studying these spatio-temporal variation in hydroclimatic data. PCA helps in understanding the contribution of each of the variables (X_i) to total variability (S_{xx}) (through the coefficient of loading vector). If same hydroclimatic variable across many locations is treated as different variables, then PCA can be used to study the relative contribution of different locations in total variability. This spatial analysis is known as empirical orthogonal function (EOF) analysis.

For EOF analysis, hydrological data set is collected in 3-D matrix. First two dimensions show the grid on which the hydroclimatic variable is observed, and the last dimension shows the time steps. Hence, data set for a hydroclimatic variable has dimensions of $p_1 \times p_2 \times n$, where p_1 and p_2 are the number of grid points in x and y directions, respectively, and n is the number of time steps. First, the data is converted to $p \times n$ matrix where $p = p_1 \times p_2$ by rearranging all the grid points. PCA loadings and corresponding principal components are calculated using the procedure discussed in Sect. 8.1.1. Further, the variance explained by each of the principal component is obtained by Eq. 8.13. The loading vector obtained for ith principal component has p loadings. If the square of loading vector is arranged in $p_1 \times p_2$ matrix, it shows the relative contribution of hydroclimatic variable at a grid point to the spatial distribution of variance of ith principal component. On the other hand, principal components obtained in EOF analysis shows the variability across the space.

Example 8.5.1
Average monthly sea surface temperature (SST) for 25 locations in Arabian Sea is recorded for 2 years as given in Table A.2. Calculate the EOF loading for SST and variance explained by individual EOFs.

Solution The monthly average sea surface temperature for 25 monitoring station is 24 × 25 matrix (say X). Empirical orthogonal functions are calculated from the corresponding anomaly matrix (say X_d) which is obtained by subtracting each of the column with its mean. The mean of different columns of X represented by $\overline{X}$ is [26.92, 27.20, 27.45, 27.60, 27.63, 27.03, 27.32, 27.63, 27.90, 28.05, 27.19, 27.46, 27.78, 28.08, 28.33, 27.39, 27.68, 27.97, 28.27, 28.50, 27.68, 27.97, 28.25, 28.50, 28.70].

The covariance matrix of the matrix X_d is given by:

$$\text{Cov}(X_d) = \begin{bmatrix} 2.13 & 2.13 & 2.05 & 1.97 & \dots & 1.25 & 1.09 & 0.94 & 0.79 \\ 2.13 & 2.18 & 2.14 & 2.09 & \dots & 1.17 & 1.02 & 0.88 & 0.72 \\ 2.05 & 2.14 & 2.15 & 2.12 & \dots & 1.05 & 0.92 & 0.78 & 0.62 \\ 1.97 & 2.09 & 2.12 & 2.1 & \dots & 0.96 & 0.85 & 0.72 & 0.56 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 1.25 & 1.17 & 1.05 & 0.96 & \dots & 0.98 & 0.86 & 0.76 & 0.69 \\ 1.09 & 1.02 & 0.92 & 0.85 & \dots & 0.86 & 0.76 & 0.67 & 0.6 \\ 0.94 & 0.88 & 0.78 & 0.72 & \dots & 0.76 & 0.67 & 0.59 & 0.54 \\ 0.79 & 0.72 & 0.62 & 0.56 & \dots & 0.69 & 0.6 & 0.54 & 0.5 \end{bmatrix}_{25\times 25}$$

The first five highest eigenvalues of the covariance matrix are 31.50, 2.95, 0.11, 0.03, and 0.02. All other eigenvalues are close to zero and hence are insignificant. The sum of all eigenvalues is 34.66. The loadings for EOFs corresponding to the eigenvalues can be obtained by using Eq. 8.11. The loading matrix thus obtained is

$$U = \begin{bmatrix} 0.26 & 0.1 & 0.41 & 0.22 & \ldots & -0.12 & 0.06 & -0.33 & 0 \\ 0.25 & 0.22 & 0.17 & -0.04 & \ldots & 0.05 & 0.02 & 0.22 & -0.05 \\ 0.24 & 0.32 & -0.07 & -0.21 & \ldots & 0.14 & -0.28 & 0.14 & 0.29 \\ 0.23 & 0.38 & -0.19 & -0.12 & \ldots & 0.22 & 0.01 & -0.12 & -0.22 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0.16 & -0.21 & -0.12 & -0.25 & \ldots & 0.2 & 0.39 & 0.11 & -0.02 \\ 0.14 & -0.18 & -0.23 & -0.24 & \ldots & -0.08 & 0.01 & -0.04 & -0.39 \\ 0.12 & -0.18 & -0.29 & -0.01 & \ldots & -0.17 & -0.16 & 0.07 & -0.08 \\ 0.11 & -0.21 & -0.29 & 0.2 & \ldots & -0.21 & 0.09 & -0.19 & 0.07 \end{bmatrix}_{25\times 23}$$

Last two eigenvectors could not determined since corresponding eigenvalues are zero. The variance explained by any EOF is the ratio of corresponding eigen value and trace of covariance matrix (Eq. 8.13). Hence, variance explained by first EOF is $\lambda_1 \big/ \sum \lambda \times 100\% = 90.88\%$, and similarly, the variance explained by the other EOFs (2nd to 5th) is 8.51%, 0.32%, 0.08% and 0.06%, respectively. It can be observed that the variance explained by first two EOFs is more than 99%; hence, the EOF analysis also leads to dimensionality reduction (instead of using data from 25 locations, two EOFs are sufficient).

8.6 Data Generation

In hydroclimatology, sometimes we need to generate data based on statistical properties of observed data. Data generation is specially needed if the observed data is less; however, the length of observed data should be sufficient to draw inferences about its population statistics. Data generation in general depends upon the fact that the cumulative probability of any random variable is uniformly distributed between 0 and 1 irrespective of nature of probability distribution function. Data can be generated for univariate and multivariate (with required correlation) case.

8.6.1 Univariate Data Generation

For generating the data for single variable, its probability distribution should be known. In general the univariate data can be generated as:

(i) From the observed data set, fit a probability distribution and calculate the parameters of cumulative distribution function.
(ii) Generate uniformly distributed random number between 0 and 1.

(iii) Taking the generated random number as the value of the CDF, calculate the value of the random variable by taking the inverse of the CDF.

For generating uniformly distributed random numbers between 0 and 1, random number generator of the following form is used.

$$\begin{aligned} R_{i+1} &= \text{Remainder of } (aR_i + b)\big/m \\ Y_{i+1} &= \frac{R_{i+1}}{m} \end{aligned} \tag{8.28}$$

where a and b are some integers and m is a very large integer ($m >> a, b$). The range of random variable R_i is 0 to $m - 1$, and hence, the range of Y_i is 0 to $1 - 1/m$. Since m is very large, the range of Y_i is effectively 0 to 1. Using the above expression, a series of Y_i can be generated.

For the last step the series of Y_i is equated to cumulative distribution function and the values of the variable are generated through inverse CDF. Hence, if F_X is the CDF for random variable X, then the value of X_i given Y_i is calculated as:

$$\begin{aligned} Y_i &= F_X(X_i) \\ X_i &= F_X^{-1}(Y_i) \end{aligned} \tag{8.29}$$

For the distributions, if the associated cumulative distribution function is not directly invertible like normal, gamma, and other, solution of Eq. 8.29 is done numerically.

Example 8.6.1

At a location, the daily average air temperature is found to be normally distributed with mean 15° C and standard deviation 2° C. Generate 20 new values of daily average air temperature.

Solution The 20 random numbers between 0 and 1 (using Eq. 8.28) generated are 0.44, 0.38, 0.77, 0.80, 0.19, 0.49, 0.45, 0.65, 0.71, 0.75, 0.28, 0.68, 0.66, 0.16, 0.12, 0.50, 0.96, 0.34, 0.59, and 0.22.

From normal distribution table, the corresponding standard normal variate (Z) is −0.151, −0.305, 0.739, 0.842, −0.878, −0.0250, −0.126, 0.385, 0.553, 0.674, −0.583, 0.468, 0.412, −0.994, −1.175, 0, 1.751, −0.412, 0.228, and −0.772.

The standard normal variate can be converted into normally distributed random variable having mean 15 and standard deviation 2 as:

$$Y = 15 + 2Z$$

Hence, corresponding daily average air temperature (in °C) is 14.698, 14.390, 16.478, 16.684, 13.244, 14.950, 14.748, 15.770, 16.106, 16.348, 13.834, 15.936, 15.824, 13.012, 12.650, 15.000, 18.502, 14.176, 15.456, and 13.456.

Example 8.6.2

At a location, the daily rainfall is following a mixed distribution with a probability mass at 0. The nonzero daily rainfall is found to follow exponential distribution with $\lambda = 1.5$. Generate 15 values of nonzero daily rainfall for the location.

Solution The 15 random numbers (between 0 and 1) generated are 0.595, 0.262, 0.603, 0.711, 0.222, 0.117, 0.297, 0.319, 0.424, 0.508, 0.086, 0.262, 0.801, 0.029, and 0.929.

The cumulative distribution function for exponential distribution with $\lambda = 1.5$ is given by:

$$F_X(x) = 1 - e^{-x/\lambda}$$
$$\text{or, } x = -\lambda \ln(1 - F_X(x)) = -1.5 \ln(1 - F_X(x))$$

Replacing $F_X(x)$ with the 15 generated random numbers will generate the exponentially distributed random numbers, which are 1.356, 0.456, 1.386, 1.862, 0.377, 0.187, 0.5290, 0.576, 0.8270, 1.064, 0.1350, 0.456, 2.422, 0.044, and 3.968, respectively.

8.6.2 Multivariate Data Generation

Multivariate data may have a correlation structure associated with them. For the case in which little or no correlation exists between the variable, the univariate data generation procedure is repeated multiple times. For multivariate data having significant correlation among them, depending upon whether to follow normal distribution or not, different procedures are used. It should be noted that the techniques discussed in this section are simple and can always yield the desired results. For more generalized approach for multivariate data generation (to conserve nonlinear association if any using the joint distribution), the copula is used as discussed in Sect. 10.10.1.

Correlated and Normally Distributed Random Variables

In this case, the correlation matrix along with the mean and standard deviation of all variables to be generated should be known. To conserve the correlation structure, the theory of Principal Component Analysis is used during data generation. Suppose that the observed multivariate data set is X having a size $n \times p$ with mean $[\mu_1, \mu_2, \ldots, \mu_p]$ and standard deviation $[\sigma_1, \sigma_2, \ldots, \sigma_p]$. The matrix X can be standardized into Y by subtracting respective column mean and dividing by respective column standard deviation. Suppose Z is principal component matrix of Y with transformation matrix U and eigen values $[\lambda_1, \lambda_2, \ldots, \lambda_p]$; hence, $Z = YU$. Y can be calculated as ZU^T as the transformation matrix is orthogonal. Further, if size of Y is $n \times p$, then size of Z is $n \times p$ with ith column or ith principal component has a mean of zero and standard deviation of λ_i. These observations can be utilized in data generation. The procedure of data generation for p correlated and normally distributed variables is given below:

(i) From the observed data set, calculate the mean and standard deviation vectors and correlation matrix.
(ii) Principal component transformation matrix (U) and corresponding eigenvalues $[\lambda_1, \lambda_2, \ldots, \lambda_p]$ are calculated from known correlation matrix.
(iii) p different normal distributed random variables with mean 0 and standard deviation λ_i are generated for the length n. The matrix of these variables (Z) is considered principal components as they are uncorrelated.
(iv) The standardized variable matrix Y is calculated from Z and U.

$$Y = ZU^T \tag{8.30}$$

(v) The standardized variable Y can be transformed into the X by multiplying ith column with the corresponding standard deviation and adding the column mean.

As the linear transformation does not change the correlation structure, the above procedure provides the multivariate normally distributed data with required mean, standard deviation, and correlation structure.

Correlated and Non-normal Random Variables

Correlated non-normal data can be generated by generating the normally distributed data and transforming it to other distribution. The procedure of data generation is explained below:

(i) From the observed data, fit an appropriate distribution.
(ii) Calculate the correlation matrix from the observed data.
(iii) Using the procedure discussed in the last section, the correlated standard normal multivariate data set is obtained.
(iv) The data set so obtained is converted to the cumulative probability using the standard normal distribution.
(v) The cumulative probability is back-transformed into the multivariate data set using the known probability distribution of observed data.

It should be noted that if the last step of transformation is nonlinear, the correlation structure may change in generated data set.

Example 8.6.3
From a historical records of daily precipitation anomaly, precipitable water anomaly and pressure anomaly, the following correlation structure is obtained.

$$\mathrm{Cor}(X) = \begin{bmatrix} 1 & 0.776 & 0.623 \\ 0.776 & 1 & 0.637 \\ 0.623 & 0.637 & 1 \end{bmatrix}$$

Assuming that all of three variables are distributed normally with mean 0 and standard deviation 1, generate the anomaly data set for 12 months in such a way that the correlation structure is preserved.

Solution Considering the steps of normally distributed multivariate data set generation with defined correlation structure, calculation is done in the following steps:

Step 1 Calculation of correlation structure
According to the example, the correlation matrix is given by

$$\mathrm{Cor}(X) = \begin{bmatrix} 1 & 0.776 & 0.623 \\ 0.776 & 1 & 0.637 \\ 0.623 & 0.637 & 1 \end{bmatrix}$$

Step 2 Calculation of principal component loadings
Corresponding to correlation matrix $\mathrm{Cor}(X)$, the eigenvalues and principal component loading matrix are calculated using the methodology of Example 8.1.2. The eigenvalues for the correlation matrix is

$$\lambda = [2.360, 0.417, 0.223]$$

Correspondingly, the loading matrix is

$$U = \begin{bmatrix} 0.590 & -0.416 & -0.692 \\ 0.593 & -0.358 & 0.721 \\ 0.548 & 0.836 & -0.036 \end{bmatrix}$$

Step 3 Generation of random principal components with required mean and standard deviation.
For generating the normally distributed variable anomaly, first three principal components (normally distributed) are generated with mean 0 and variance λ. Using the procedure discussed in Example 8.6.1 the following set of principal components are generated.

$$Z = \begin{bmatrix} -1.08 & 0.57 & -0.24 \\ 0.26 & 0.54 & -0.09 \\ -0.99 & -0.4 & 0.94 \\ 1.85 & 1.18 & -0.3 \\ -2.42 & -1.22 & 0.41 \\ 1.83 & -0.14 & -0.01 \\ 1.18 & -0.12 & -0.01 \\ 1.68 & -0.38 & -0.98 \\ -0.75 & -0.34 & 0.48 \\ 0.64 & 0.57 & 0.05 \\ 0.22 & -0.51 & -0.24 \\ -2.42 & 0.24 & -0.01 \end{bmatrix}$$

Step 4 Calculate the values following standard normal distribution with the correlation matrix (Cor$X$) (Eq. 8.30)

$$\text{Hence, } Y = ZU^T = \begin{bmatrix} -0.71 & -1.02 & -0.11 \\ -0.01 & -0.1 & 0.6 \\ -1.07 & 0.23 & -0.91 \\ 0.81 & 0.46 & 2.01 \\ -1.2 & -0.7 & -2.36 \\ 1.14 & 1.13 & 0.89 \\ 0.75 & 0.74 & 0.55 \\ 1.83 & 0.43 & 0.64 \\ -0.63 & 0.02 & -0.71 \\ 0.11 & 0.21 & 0.83 \\ 0.51 & 0.14 & -0.3 \\ -1.52 & -1.53 & -1.13 \end{bmatrix}$$

Step 5 Transform Y by multiplying with standard deviation and mean to each of the column.

As variables are already standardized (has a mean 0 and standard deviation 1), so no transformation is required. Hence, Y is the required generated data set.

Example 8.6.4

In the last example, assume the standardized precipitable water and standardized pressure for a season are the following exponential distribution with $\lambda = 2$ and $\lambda = 0.6$, respectively. Generate the data set preserving the correlation structure.

Solution From Example 8.6.3

$$Y = \begin{bmatrix} -0.71 & -1.02 & -0.11 \\ -0.01 & -0.1 & 0.6 \\ -1.07 & 0.23 & -0.91 \\ 0.81 & 0.46 & 2.01 \\ -1.2 & -0.7 & -2.36 \\ 1.14 & 1.13 & 0.89 \\ 0.75 & 0.74 & 0.55 \\ 1.83 & 0.43 & 0.64 \\ -0.63 & 0.02 & -0.71 \\ 0.11 & 0.21 & 0.83 \\ 0.51 & 0.14 & -0.3 \\ -1.52 & -1.53 & -1.13 \end{bmatrix}$$

The corresponding normal distribution cumulative probability is

$$F(Y) = \begin{bmatrix} 0.24 & 0.15 & 0.46 \\ 0.50 & 0.46 & 0.72 \\ 0.14 & 0.59 & 0.18 \\ 0.79 & 0.68 & 0.98 \\ 0.11 & 0.24 & 0.01 \\ 0.87 & 0.87 & 0.81 \\ 0.77 & 0.77 & 0.71 \\ 0.97 & 0.66 & 0.74 \\ 0.26 & 0.51 & 0.24 \\ 0.54 & 0.58 & 0.8 \\ 0.69 & 0.56 & 0.38 \\ 0.06 & 0.06 & 0.13 \end{bmatrix}$$

The second and third columns of $F(Y)$ (corresponding to standardized precipitable water and standardized pressure) can be transformed back using the inverse function of their cumulative distribution function as done in Example 8.6.2. Hence, the generated data is

$$X = \begin{bmatrix} -0.71 & 0.33 & 0.37 \\ -0.01 & 1.23 & 0.76 \\ -1.07 & 1.78 & 0.12 \\ 0.81 & 2.28 & 2.35 \\ -1.2 & 0.55 & 0.01 \\ 1.14 & 4.08 & 1.00 \\ 0.75 & 2.94 & 0.74 \\ 1.83 & 2.16 & 0.81 \\ -0.63 & 1.43 & 0.16 \\ 0.11 & 1.74 & 0.97 \\ 0.51 & 1.64 & 0.29 \\ -1.52 & 0.12 & 0.08 \end{bmatrix}$$

8.7 Analysis of Variance in Hydrology and Hydroclimatology

Analysis of variance (ANOVA) is a statistical procedure to check the significance of variance in different samples of same population, and thus, check the null hypothesis that mean of all samples is equal. ANOVA is also used to study the spatial homogeneity of hydroclimatic data. Under ANOVA, the variance is partitioned into a number of sources/factors. Depending upon the number of sources of variance (referred as attributes), apart from the system error (experimental or measurement errors), one-way ANOVA or two-way ANOVA is used.

8.7.1 One-Way Analysis of Variance

One-way ANOVA is used when apart from system error, only a single factor/attribute contributes to the variance. Due to this attribute, the sample mean differs from population mean. The sample mean can be written as sum of population mean and effect of attribute.

$$\overline{x_i} = \mu + \alpha_i \tag{8.31}$$

where $\overline{x_i}$ is sample mean of ith sample, μ is the mean of all the samples or population mean and α_i is effect of attribute on the sample mean. In one-way ANOVA, the null hypothesis is that all the sample means are equal to population mean.

$$\begin{aligned} H_0 &: \overline{x_1} = \overline{x_2} = \cdots = \overline{x_a} = \mu \\ \text{or, } H_0 &: \alpha_i = 0 \qquad \text{for all } i \in \{1, 2, \ldots, a\} \end{aligned} \tag{8.32}$$

where $\alpha_i = \overline{x_i} - \mu$ for $i = 1, 2, \ldots, a$.

Correspondingly, the alternative hypothesis states that at least one sample mean is not equal to the population mean.

$$H_1 : \alpha_i \neq 0 \qquad \text{for at least one value of } i \tag{8.33}$$

Null hypothesis can only be true if all the variability is primarily contributed due to chance or random error. Hence, the variance in the data set needs to be separated into variance due to attribute and variance due to random error. Ratio of mean variance contribution from these two categories is the test statistic for testing null hypothesis. Using Eq. 8.31, an observation from ith sample can be expressed as sum of population mean, attribute effect, and the random error.

$$x_{ij} = \overline{x_i} + e_{ij} = \mu + \alpha_i + e_{ij} \tag{8.34}$$

where e_{ij} is the random error associated with the element x_{ij}. This equation can be written in terms of deviation from sample or population mean as:

$$\begin{aligned} x_{ij} - \mu &= \alpha_i + e_{ij} \\ \text{or, } (x_{ij} - \mu) &= (\overline{x_i} - \mu) + (x_{ij} - \overline{x_i}) \end{aligned} \tag{8.35}$$

If different samples are available and ith sample contains n_i observations, then the total sum of squares (sum of squared deviation of all observations from the population mean) is obtained by squaring both sides and taking the summation for all the observations.

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(x_{ij} - \mu)^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\overline{x_i} - \mu)^2 + \sum_{i=1}^{a}\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i})^2 \tag{8.36}$$

$$\text{SST} = \text{SSA} + \text{SSE} \tag{8.37}$$

where a is the number of different attributes/samples, n_i is the number of elements in ith sample, SSA shows the variance contributed by attribute, and SSE shows the variance due to random error. Hence, Eq. 8.37 shows that the total variance is partitioned into variance due to attribute and random errors. Degrees of freedom for each term can be evaluated as follows. While calculating SSA one mean, μ is computed from all SSA attributes (total a). So, one degree of freedom is lost. Hence, degrees of freedom for SSA is $a - 1$. Similarly, if total number of observations is N, then degrees of freedom for SSE and SST are $N - a$ and $N - 1$, respectively. Using these degrees of freedom, the mean square error and attribute can be calculated.

$$\text{MSE} = \frac{\text{SSE}}{N - a} \qquad \text{MSA} = \frac{\text{SSA}}{a - 1} \tag{8.38}$$

The test statistics in one-way ANOVA is given by the ratio of MSA and MSE.

$$F = \frac{\text{MSA}}{\text{MSE}} \tag{8.39}$$

A large value of test statistics (F) indicates the effect of attribute is prominent as compared to the effect of the random error over total variance, and thus, all the sample means are not equal. The test statistics follows F-distribution with $(a-1)$ and $(N-a)$ degrees of freedom. Hence, the null hypothesis is rejected if $F > F_\alpha(a - 1, N - a)$ at α level of significance. Details of the one-way ANOVA is summarized in Table 8.1. Sum of square can be obtained using the following equations also.

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} x_{ij}^2 - C \tag{8.40}$$

Table 8.1 A typical one-way ANOVA table

Source of variation	Degree of freedom	Sum of squares	Mean square	F
Attribute	$a-1$	SSA	MSA	$^{\text{MSA}}/_{\text{MSE}}$
Error	$N-a$	SSE	MSE	
Total	$N-1$	SST		

$$\text{SSA} = \sum_{i=1}^{a} \frac{T_i^2}{n_i} - C \tag{8.41}$$

$$\text{SSE} = \text{SST} - \text{SSA} \tag{8.42}$$

$$C = \frac{\left(\sum_{i=1}^{a} T_i\right)^2}{N} \tag{8.43}$$

where T_i represents the sum of observations in ith sample $\left(T_i = \sum_{j=1}^{n_i} x_{ij}\right)$ and C is called correction term for the mean.

Example 8.7.1

For three different locations, the following average monthly meridional wind speed (in m/s) was recorded for a year (Table 8.2).

Test at 0.05 level of significance whether the difference among the means is significant or not.

Solution In this dataset, the location is the only source of variation; hence, it can be termed as the attribute for ANOVA analysis. Denoting x_i ($i = 1, 2, 3$) as the wind speed variable from ith location, μ as overall mean and $\alpha_i = \overline{x}_i - \mu$

Null Hypothesis: Means do not differ significantly. $\alpha_i = 0$ for $i \in \{1, 2, 3\}$
Alternative Hypothesis: $\alpha_i \neq 0$ for at least one value of i.
Level of Significance: $\alpha = 5\%$

Table 8.2 Monthly meridional wind speed (in m/s)

Location	Months											
	1	2	3	4	5	6	7	8	9	10	11	12
1	2.21	0.62	2.03	0.8	0.84	1.52	0.57	1.39	2.3	1.78	2.17	1.72
2	0.87	1.65	0.74	3.52	2.27	2.15	1.33	1.87	1.93	2.48	1.44	1.03
3	1.89	3.03	1.85	−0.29	0.68	2.76	1.03	0.88	1.03	2.49	0.88	1.17

Different quantities of one-way ANOVA table can be calculated using Eq. 8.37. Let X_1, X_2 and X_3 represent average monthly meridional wind speed. Further, the individual value for jth month for ith location is denoted by $x_{i,j}$.

$$\text{Overall mean}(\mu) = \frac{\sum\limits_{i=1}^{3}\sum\limits_{j=1}^{12} x_{i,j}}{12 \times 3} = 1.573$$

Mean of monthly meridional wind at different locations are given by:

$$\overline{x_1} = \frac{\sum\limits_{j=1}^{12} x_{1,j}}{12} = 1.496 \text{ m/s}$$

Similarly, $\overline{x_2} = 1.773$ m/s and $\overline{x_3} = 1.450$ m/s.

$$\text{SST} = \sum_{i=1}^{3}\sum_{j=1}^{12}(x_{i,j} - \mu)^2 = 22.393$$

$$\text{SSA} = 12\sum_{i=1}^{3}(\overline{x_i} - \mu)^2 = 0.734$$

$$\text{SSE} = \sum_{i=1}^{3}\sum_{j=1}^{12}(x_{i,j} - \overline{x_i})^2 = 21.659$$

The degrees of freedom for SST, SSA, and SSE are $36 - 1 = 35$, $3 - 1 = 2$, and $36 - 3 = 33$, respectively. Hence,

$$\text{MSA} = \frac{\text{SSA}}{2} = \frac{0.734}{2} = 0.367$$

$$\text{MSE} = \frac{\text{SSE}}{33} = \frac{21.659}{33} = 0.656$$

$$\text{F} = \frac{\text{MSA}}{\text{MSE}} = \frac{0.367}{0.656} = 0.559$$

The one-way ANOVA table shown in Table 8.3.

The test statistics for ANOVA analysis (F) follows F-distribution with $(a - 1)$ and $(N - a)$ degrees of freedom.

$$F_\alpha((a-1),(N-a)) = F_{(0.05)}(2, 33) = 3.285$$

Table 8.3 One-way ANOVA table for Example 8.7.1

Source of variation	Degree of freedom	Sum of squares	Mean square	F
Attribute	2	0.734	0.367	0.559
Error	33	21.659	0.656	
Total	35	22.393		

As $0.559 < 3.285$ ($F_{(0.05)}(2, 33)$); hence, null hypothesis cannot be rejected; i.e., the mean of monthly meridional wind speed does not differ across the locations at 5% significance level.

Example 8.7.2
The average annual sea surface temperature is recorded for four locations and is given in Table 8.4. Test at 1% significance level whether the mean sea surface temperature differs across the locations.

Solution Different locations are the only source of variance in sea surface temperature; hence, they can be considered as the attribute for the ANOVA analysis. Denoting x_i ($i = 1, 2, 3, 4$) as the annual sea surface temperature from ith location, μ as overall mean and $\alpha_i = \overline{x_i} - \mu$

Null Hypothesis: Means do not differ significantly across the locations, i.e., $\alpha_i = 0$ for $i \in \{1, 2, 3, 4\}$
Alternative Hypothesis: $\alpha_i \neq 0$ for at least one value of i.
Level of Significance: $\alpha = 1\%$

Let X_1, X_2, X_3 and X_4 represent average annual sea surface temperature at locations 1, 2, 3, and 4, respectively. Further, the individual value for year 2005 at ith location is denoted by $x_{i,1}$ and value for year 2010 is denoted by $x_{i,6}$. Different quantities of one-way ANOVA table can be calculated using Eqs. 8.40–8.43.

$$T_1 = \sum_{j=1}^{6} x_{1,j} = 129.95$$

Table 8.4 Sea surface temperature for 6 years

Location	Year					
	2005	2006	2007	2008	2009	2010
A	20.39	19.84	20.48	20.56	25.51	23.17
B	31.57	30.52	25.64	25.97	28.53	22.68
C	27.94	26.71	24.94	28.47	26.20	23.97
D	23.45	18.68	20.65	24.19	26.57	24.67

Similarly, $T_2 = 164.91$, $T_3 = 158.23$ and $T_4 = 138.21$

$$C = \frac{\left(\sum_{i=1}^{a=4} T_i\right)^2}{N} = \frac{(129.95 + 164.91 + 158.23 + 138.21)^2}{4 \times 6} = 14568.15$$

$$\text{SST} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} x_{i,j}^2 - C = 14840.07 - 14568.15 = 271.92$$

$$\text{SSA} = \sum_{i=1}^{a} \frac{T_i^2}{n_i} - C = 14703.51 - 14568.15 = 135.35$$

$$\text{SSE} = \text{SST} - \text{SSA} = 271.92 - 135.35 = 136.57$$

The MSE and MSA can be calculated by dividing SSE and SSA with their respective degrees of freedom. The degrees of freedom for SSA and SSE are $(a - 1) = 3$ and $(N - a) = 24 - 4 = 20$. Hence,

$$\text{MSA} = \frac{\text{SSA}}{a - 1} = \frac{135.52}{3} = 45.17$$

$$\text{MSE} = \frac{\text{SSE}}{N - a} = \frac{136.57}{20} = 6.83$$

$$F = \frac{\text{MSA}}{\text{MSE}} = \frac{45.17}{6.83} = 6.61$$

Summarizing this, one-way ANOVA is shown in Table 8.5:

The test statistics F is supposed to follow F-distribution with $(a-1)$ and $(N-a)$, i.e., 3 and 20 degrees of freedom.

$$F_\alpha(a - 1, N - a) = F_{(0.01)}(3, 20) = 4.94$$

Table 8.5 One-way ANOVA table for Solution 8.7.2

Source of variation	Degree of freedom	Sum of squares	Mean square	F
Attribute	3	135.35	45.17	6.61
Error	20	136.57	6.61	
Total	23	271.92		

As 6.61 > 4.94 ($F_{(0.01)}(3, 20)$); hence, null hypothesis must be rejected at 1% level of significance; i.e., sea surface temperature at different locations is different.

8.7.2 *Two-Way Analysis of Variance*

Two-way ANOVA is used when apart from random error (white noise), two other factors, called as the attribute-1 and attribute-2, also contribute to the variance. Under this condition, the sample mean for ith attribute-1 and jth attribute-2 is given by:

$$\overline{x_i} = \mu + \alpha_i + \beta_j \tag{8.44}$$

where $\overline{x_i}$ is sample mean of ith sample, μ is the mean of all the samples or population mean, α_i and β_j are the effects of ith component of attribute-1 and jth component of attribute-2 on sample mean. In two-way ANOVA, the null hypothesis is that the effects across the different components of each attribute are same. Thus,

$$H_0 : \alpha_i = 0 \text{ and } \beta_j = 0 \qquad \text{for all } i \in \{1, 2, \ldots, a\} \text{ and } j \in \{1, 2, \ldots, b\} \tag{8.45}$$

where a and b are the number of components of attribute-1 and attribute-2, respectively. Correspondingly, the alternative hypothesis states that at least one sample mean is not equal to population mean.

$$H_a : \text{ Null hypothesis is not true} \tag{8.46}$$

Null hypothesis can only be true if all the variability is primarily contributed due to chance or random error. Hence, the variance in the data set needs to be separated into variance due to components of attribute-1 and attribute-2 and random error. Using Eq. 8.44 an observation from ith sample is expressed as:

$$x_{ij} = \overline{x_i} + e_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \tag{8.47}$$

where e_{ij} show the random error associated with the element x_{ij}. Using the above equation, the total variance can be expressed as:

Table 8.6 A typical two-way ANOVA table

Source of variation	Degree of freedom	Sum of squares	Mean square	F
Attribute-1	$a-1$	SSA_1	MSA_1	MSA_1/MSE
Attribute-2	$b-1$	SSA_2	MSA_2	MSA_2/MSE
Error	$(a-1)(b-1)$	SSE	MSE	
Total	$N-1$	SST		

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\overline{x})^2 = b\sum_{i=1}^{a}(\overline{x_i}-\overline{x})^2 + a\sum_{j=1}^{b}(\overline{x_j}-\overline{x})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\overline{x_i}-\overline{x_j}+\overline{x})^2 \tag{8.48}$$

$$\text{SST} = \text{SSA}_1 + \text{SSA}_2 + \text{SSE} \tag{8.49}$$

where $\overline{x}$ is overall mean, $\overline{x_i}$ is mean of all the observations for ith component of attribute-1, $\overline{x_j}$ is the mean of all the observations for jth component of attribute-2, a is the number of components in attribute-1 and b is number of components in attribute-2. SSA_1 and SSA_2 represents variance contributed due to the effect of attribute-1 and attribute-2, respectively. The degrees of freedom for SSA_1, SSA_2 and SSE are $(a-1)$, $(b-1)$ and $(a-1)(b-1)$, respectively. The degrees of freedom for SST are $ab-1$. Using the degrees of freedom, the mean of SSA_1, SSA_2 and SSE (i.e., MSA_1, MSA_2 and MSE) is calculated. The test statistics for two-way ANOVA is defined by:

$$F_1 = \frac{\text{MSA}_1}{\text{MSE}} \tag{8.50}$$

$$F_2 = \frac{\text{MSA}_2}{\text{MSE}} \tag{8.51}$$

F_1 follows F-distribution with $(a-1)$ and $(a-1)(b-1)$ degrees of freedom. Similarly, F_2 follows F-distribution with $(b-1)$ and $(a-1)(b-1)$ degrees of freedom. Null hypothesis of no significant difference in mean cannot be rejected at α significance level if $F_1 > F_\alpha((a-1),(a-1)(b-1))$ and $F_2 > F_\alpha((b-1),(a-1)(b-1))$. Details of two-way ANOVA are summarized in Table 8.6.

Different partitions of variance can be calculated using the following relationships also.

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b} x_{ij}^2 - C \tag{8.52}$$

$$\mathrm{SSA}_1 = \frac{1}{b}\sum_{i=1}^{a} T_{i\bullet}^2 - C \tag{8.53}$$

$$\mathrm{SSA}_2 = \frac{1}{a}\sum_{j=1}^{b} T_{\bullet j}^2 - C \tag{8.54}$$

$$\mathrm{SSE} = \mathrm{SST} - \mathrm{SS(Tr)} - \mathrm{SSB} \tag{8.55}$$

$$C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b} x_{ij}\right)^2}{ab} \tag{8.56}$$

where $T_{i\bullet}$ and $T_{\bullet j}$ represent the sum of observations for ith component of attribute-1 and sum of observations for jth component of attribute-2, respectively, and C is called correction term for the observation mean.

Example 8.7.3
It is required to analyze the effect of global circulation models (GCMs) and hydrological models (HMs) on the variation of peak flow at the outlet of a study basin. The following table shows the magnitude of peak flow (Mm^3) at the outlet for different GCM and HM combinations.

Global Circulation Models (GCMs)	Hydrological Model (HM)		
	HM-1	HM-2	HM-3
GCM-1	450	435	515
GCM-2	480	461	525
GCM-3	495	505	537
GCM-4	435	372	497

Check whether that the mean of peak streamflow differs either due to GCM or HM selected at 1% level of significance.

Solution There are two sources of variances, one being GCM and other being HM. Different types of GCMs and different HMs can be considered as first and second attribute. Hence, $a = 4$ and $b = 3$.

Null Hypothesis: Mean of peak streamflow does not differ significantly due to GCM or HM selected. $\alpha_i = 0$ for $i \in \{1, 2, 3, 4\}$ and $\beta_j = 0$ for $j \in \{1, 2, 3\}$
Alternative Hypothesis: Null hypothesis is not true
Level of Significance: $\alpha = 1\%$

Different test statistics for two-way ANOVA can be calculated using Eqs. 8.52–8.56.

$$T_{1\bullet} = (450 + 435 + 515) = 1400$$

Similarly, $T_{2\bullet} = 1466$, $T_{3\bullet} = 1537$, $T_{4\bullet} = 1304$, $T_{\bullet 1} = 1860$, $T_{\bullet 2} = 1773$ and $T_{\bullet 3} = 2074$

$$C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b} x_{ij}\right)^2}{ab} = 2714154.08$$

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b} x_{ij}^2 - C = 24378.92$$

$$\text{SSA}_1 = \frac{1}{b}\sum_{i=1}^{a} T_{i\bullet}^2 - C = \frac{8171941}{3} - 2714154.08 = 9826.25$$

$$\text{SSA}_2 = \frac{1}{a}\sum_{j=1}^{b} T_{\bullet j}^2 - C = \frac{10904605}{4} - 2714154.08 = 11997.17$$

$$\text{SSE} = \text{SST} - \text{SSA}_1 - \text{SSA}_2 = 2555.5$$

The mean square sum (MSA_1, MSA_2 and MSE) can be calculated by dividing respective sum of squares (SSA_1, SSA_2, and SSE) with their respective degrees of freedom. The degrees of freedom for SSA_1, SSA_2, and SSE are $(a-1) = 3$, $(b-1) = 2$, and $(a-1)(b-1) = 6$, respectively. Hence,

$$\text{MSA}_1 = \frac{\text{SSA}_1}{a-1} = \frac{9826.25}{3} = 3275.42$$

$$\text{MSA}_2 = \frac{\text{SSA}_2}{b-1} = \frac{11997.17}{2} = 5998.58$$

$$\text{MSE} = \frac{\text{SSE}}{(a-1)(b-1)} = \frac{2555.5}{6} = 425.92$$

$$F_1 = \frac{\text{MSA}_1}{\text{MSE}} = \frac{3275.42}{425.92} = 7.69$$

$$F_2 = \frac{\text{MSA}_2}{\text{MSE}} = \frac{5998.58}{425.92} = 14.08$$

All these values are summarized in the following two-way ANOVA table.

The test statistics F_1 is supposed to follow F-distribution with $(a-1)$ and $(a-1)(b-1)$, i.e., 3 and 6 degrees of freedoms.

$$F_\alpha((a-1), (a-1)(b-1)) = F_{0.01}(3, 6) = 9.78$$

Source of variation	Degree of freedom	Sum of squares	Mean squares	F
1st Attribute	3	9826.25	3275.42	7.69
2nd Attribute	2	11997.17	5998.58	14.08
Error	6	2555.5	425.92	
Total	11	24378.92		

The test statistics F_2 is supposed to follow F-distribution with $(b-1)$ and $(a-1)(b-1)$, i.e., 2 and 6 degrees of freedoms.

$$F_{\alpha}((b-1),(a-1)(b-1)) = F_{0.01}(2,6) = 10.92$$

Since $F_1 < 9.78$, it indicates that there is no significant difference between GCMs. However, $F_2 > 10.92$, it indicates that there is significant difference between hydrologic variables at 1% significance level.

8.7.3 *Multiple Comparisons*

The ANOVA discussed in last section checks the significance of null hypothesis that sample mean does not differ from population mean; however, it does not provide any information about the sample whose mean differs significantly. Many a time in hydroclimatology, the investigator needs to investigate spatial inhomogeneity and find the location that has mean significantly different from overall mean. Significance of difference in mean for data from two locations can be tested using t test. For 'k' different locations or k different variables, the difference in mean needs to be tested for all possible pairs. Hence, a total of ${}^{k}C_2 = k(k-1)\big/2$ two sample t tests are required, which is very large number, even if k is relatively small. Other issues for these tests will be to ensure the independence between the tests and to assign an overall significance level. For overcoming these difficulties, many multiple comparison procedures have been proposed. One of popular method for multiple test is **Boneferroni method**. In this method, level of significance is equally distributed between all the t tests; hence, each t test is conducted at $2\alpha\big/k(k-1)$ level of significance.

8.8 MATLAB Examples

Examples solved in this chapter can also be solved using MATLAB scripts. Following built-in function is helpful in this regard:

- Principal components can be calculated using '`pca`' in-built function.

  - `[loadings,pc,eigen_val]= pca(data)`

    This function can be used for calculating the principal component loading matrix, principal component, and eigenvalues of covariance matrix. It should be noted that this function calculate the principal components using covariance of anomaly of ‘`data`’ matrix. For calculating the principal components from correlation matrix, the input matrix (‘`data`’) should be standardized (mean is subtracted from every column and followed by division with column standard deviation), as done in example script given in Box 8.1. Different inputs and outputs of the function are explained as the following:

    `data`: $n \times p$ matrix of n observations and p variables.
    `loading`: $p \times p$ loading matrix for principal components.
    `pc`: Principal component matrix of size $n \times p$.
    `eigen_val`: Eigenvalues of covariance matrix corresponding to p principal components.

- The random numbers following different distributions can be generated using different in-built functions. Some of these functions are discussed as follows:

  - `X = rand(Sz1,Sz2,...,Szn)`
    The function generates uniformly distributed random number between 0 and 1. The output matrix X is of size `Sz1` $\times$ `Sz2` $\times \cdots \times$ `Szn`. It should be noted that for command `X = rand(n)`, the output X will be $n \times n$ matrix.
  - `X = randi(Sz1,Sz2,...,Szn)`
    The function generates uniformly distributed random number greater than 1.
  - `X = randn(Sz1,Sz2,...,Szn)`
    The function generates normally distributed random number with mean 0 and standard deviation 1.
  - `rng(seed)`
    The function initializes the function `rand`, `randi` and `randn` with a non-negative number as seed.
  - `X = normrnd(mu,sigma,[Sz1,Sz2,...,Szn])`
    This function generate normally distributed random number matrix X with mean `mu` and `sigma`. The mean (`mu`) and standard deviation (`sigma`) can be vector also.
  - `X = exprnd(mu,[Sz1,Sz2,...,Szn])`
    This function generate exponentially distributed random number matrix X with mean `mu`.
  - `X = gamrnd(A,B,[Sz1,Sz2,...,Szn])`
    This function generate gamma-distributed random number matrix X with $\alpha =$ `A` and $\beta =$ `B`.

- One-way or two-way ANOVA analysis can be done using `anova1` and `anova2` functions. These functions also generate standard ANOVA tables and return the p-value of the tests.

For instance, Examples 8.1.1, 8.1.2 and 8.3.1 can be solved using the MATLAB script in Box 8.1.

Box 8.1 Sample MATLAB script for solving Example 8.1.1 and associated examples

```
clear;clc;close all;
output_file=['output' filesep() 'code_1_result.txt'];
load(['data' filesep() 'umb_diff_var.mat']);

cov_data=cov(data);
[cov_loading,cov_PC,cov_eig_val]=pca(data);
variance_explain_cov_pc=cov_eig_val./sum(cov_eig_val);

std_data=(data-mean(data))./std(data);corr_data=cov(std_data);
[corr_loading,corr_PC,corr_eig_val]=pca(std_data);
variance_explain_corr_pc=corr_eig_val./sum(corr_eig_val);

Y=data(:,1)';X=data(:,2:end)';
[spca_eigen_vec,spc,spca_eigen_val]=SPCA(X,Y);

%Display results
delete(output_file);diary(output_file);diary on;
disp('Covariance Matrix of the data');
disp(cov_data)
disp('Correlation Matrix of the data');
disp(corr_data)
disp('loading matrix for PC obtained from covariance matrix');
disp(cov_loading)
disp('Variance explained by PC obtained from covariance matrix');
disp(variance_explain_cov_pc')
disp('loading matrix for PC obtained from correlation matrix');
disp(corr_loading)
disp('Variance explained by PC obtained from correlation matrix');
disp(variance_explain_corr_pc')
disp('SPC loading considering precipitation as dependent variable');
disp(spca_eigen_vec');disp('SPC Values');
spc_val_text=[];
for i=1:length(spc)
    spc_val_text=sprintf('%s %3.2f, ',spc_val_text,spc(i));
    if mod(i,6)==0
        spc_val_text=sprintf('%s\n',spc_val_text);
    end
end
disp(spc_val_text);diary off;
```

It should be noted that the provided code calls a user-defined function ‘SPCA’. This function is defined in a function M-file, and it should be placed in the same directory. The ‘SPCA’ function definition is provided in Box 8.2

Box 8.2 MATLAB function for calculating SPCA

```
function [spca_eigen_vec,spc,spca_eigen_val]=SPCA(X,Y)
spca_eigen_vec=[];spca_eigen_val=[]; spc=[]; %#ok<NASGU>

%% Data validation

% X should be p*n matrix
% where p is number of independent variables, n is number of observations

% Y should be l*n matrix
% where l is number of dependent variables, n is number of observations

num_observations=max(size(X));
num_dependent_var=size(Y,1);
%%%% Data Check Complete

%% SPCA Part starts
H=eye(num_observations)-(ones(num_observations,1)*ones(num_observations,1)')/num_observations;
L=Y'*Y;
Q=X*H*L*H*X';
[eig_vec, eig_val]=eig(Q);

%pick up top l eigen vector as SPCA coefficient
eig_vec=real(eig_vec);
[eig_val_sorted,previous_eig_loc]=sort(diag(real(eig_val)),'descend');

spca_eigen_vec=eig_vec(:,previous_eig_loc);
spca_eigen_val=eig_val_sorted(1:num_dependent_var);

spca_eigen_vec=spca_eigen_vec(:,1:num_dependent_var);
for i=1:num_dependent_var
    if corr((spca_eigen_vec(:,i)'*X)',Y(i,:)')<0
        spca_eigen_vec(:,i)=-spca_eigen_vec(:,i);
    end
end

spc=spca_eigen_vec'*X;
%%%%%SPCA part ends
end
```

The results for the script given in Box 8.1 are provided in Box 8.3. The results obtained match with the Example 8.1.1, 8.1.2 and 8.3.1.

Box 8.3 Results for Box 8.1

```
Covariance Matrix of the data
  # Not shown as it is 9*9 matrix

Correlation Matrix of the data
  # Not shown as it is 9*9 matrix

loading matrix for PC obtained from covariance
    matrix
  # Not shown as it is 9*9 matrix

Variance explained by PC obtained from covariance
    matrix
     0.9542     0.0440     0.0015     0.0002     0.0000
           0.0000     0.0000     0.0000     0.0000

loading matrix for PC obtained from correlation
    matrix
  # Not shown as it is 9*9 matrix

Variance explained by PC obtained from correlation
    matrix
     0.5710     0.2727     0.1064     0.0336     0.0121
           0.0036     0.0005     0.0001     0.0000

SPC loading considering precipitation as dependent
    variable
    -0.0148     0.4080    -0.0987    -0.0213     0.1351
         -0.8963     0.0400    -0.0037

SPC Values
 -807.04,   -798.28,   -786.16,   -758.74,   -731.69,
     -692.21,
 -691.27,   -701.56,   -737.46,   -778.75,   -805.80,
     -803.13,
 -810.25,   -802.53,   -787.62,   -758.28,   -723.20,
     -688.51,
 -694.76,   -694.19,   -734.85,   -780.34,   -795.36,
     -820.96,
```

Similarly, Example 8.5.1 can be solved using MATLAB script provided in Box 8.4.

Box 8.4 Sample MATLAB script for solving Example 8.5.1

```
clear;clc;close all;
load(['data' filesep() 'arabian_sea_sst.mat']);

cov_data=cov(data);
[loading,eof_score,eig_val]=pca(data);
variance_explained=eig_val./sum(eig_val);

%Display results
output_file=['output' filesep() 'code_2_result.txt'
    ];
delete(output_file);diary(output_file);diary on;
disp('Covariance Matrix for data');
disp(cov_data)
disp('Loading matrix for EOF');
disp(loading)
disp('Variance explained by EOF');
disp(variance_explained');
diary off;
```

The results for the script given in Box 8.4 are given in Box 8.5. The results match with the Example 8.5.1.

Box 8.5 Results for Box 8.4

```
Covariance Matrix for data
    # Not Shown as it is 25*25 matrix

Loading matrix for EOF
    # Not Shown as it is 25*25 matrix

Variance explained by EOF

    0.9089    0.0852    0.0034    0.0012    0.0007    0.0002
         0.0001    0.0001    0.0001    0.0000    0.0000    0.0000
            0.0000    0.0000
  # Not shown as all other are zero
```

For solving Example 8.7.2, the sample MATLAB script is provided in Box 8.6. It should be noted that this example script does not use the built-in function to calculate one-way ANOVA rather shows the steps for calculations.

Box 8.6 Sample MATLAB script for solving Example 8.7.2

```
clc;clear;close all;
alpha=0.01;

data=[20.39,19.84,20.48,20.56,25.51,23.17;...
    31.57,30.52,25.64,25.97,28.53,22.68;...
    27.94,26.71,24.94,28.47,26.20,23.97;...
    23.45,18.68,20.65,24.19,26.57,24.67];

T=sum(data,2);
C=sum(T)^2/numel(data);
SST=sum(data(:).^2)-C;
SS_Tr=sum(T.^2/size(data,2))-C;
SSE=SST-SS_Tr;

MSE=SSE/20;
MS_Tr=SS_Tr/3;

F=MS_Tr/MSE;

%%% Display the results
output_file=['output' filesep() 'code_3_result.txt'];
delete(output_file);diary(output_file);diary on;
fprintf('The test statistic (F) is %2.2f.\n',F);
fprintf('The critical value of test statistic is %1.3f.\n',...
    finv(1-alpha,3,20))
if F > finv(1-alpha,3,20)
    info=sprintf('%2.2f > %1.3f, so the null hypothesis is rejected', ...
        F, finv(1-alpha,3,20));
else
    info=sprintf('%2.2f < %1.3f, so the null hypothesis is accepted', ...
        F, finv(1-alpha,3,20));
end
fprintf('%s at %0.2f level of significance.\n',info,alpha);
diary off
```

The output of the code in Box 8.6 is provided in the Box 8.7. The decision for rejecting the null hypothesis matches with the Example 8.7.2.

Box 8.7 Results for Box 8.6

```
The test statistic (F) is 6.61.
The critical value of test statistic is 4.938.
6.61 > 4.938, so the null hypothesis is rejected at 0.01 level of significance.
```

Exercise

8.1 For Upper Mahanadi Basin, the mean monthly rainfall, air temperature, precipitable water, pressure, geo-potential height at 925 mb and wind speed at 925 mb are presented for the year 1971 in Table 8.7.

Table 8.7 Monthly average data for upper Mahanadi basin

Month	Precipitation (mm)	Air Temperature (°C)	Precipitable water (kg/m^2)	Pressure (mb)	Geo-potential height (m)	Wind speed (m/s)
1	9.51	20.1	18.65	962.88	788.11	33.98
2	15.42	22.61	17.81	960.73	770.48	27.96
3	8.72	26.52	14.9	961.25	778.71	22.81
4	29.55	31.5	24.61	955.72	731.13	17.91
5	36.77	33.25	29.84	953.51	712.47	12.12
6	375.43	26.94	50.06	950.38	678.67	11.82
7	331.02	24.52	52.71	950.7	681.07	13.80
8	312.48	23.84	50.57	952.33	695.3	13.32
9	103.19	23.9	44.11	955.3	722.38	9.26
10	104.47	22.43	38.7	959.04	755.85	2.21
11	0.54	20.06	15.81	964.97	806.14	22.69
12	0.00	19.11	11.15	964	797.28	23.47

Calculate the principal components for the data set. How many principal components are enough for explaining 90% of total variance of the data set? (Ans. The first principal component explains 97% of variability.)

8.2 In Bay of Bengal, the sea surface temperature, zonal and meridional wind speed, pressure and specific humidity are monitored for 3 years as presented in Table 8.8.

Calculate the loading matrix for principal components (calculated using correlation matrix of data) and variance explained by them.

$$\text{Ans.}\quad U = \begin{bmatrix} 0.4582 & 0.1203 & 0.3245 & 0.0239 & -0.0437 & 0.7108 & -0.4032 \\ 0.3762 & 0.2086 & 0.5513 & 0.4130 & -0.1914 & -0.5511 & 0.0070 \\ 0.3807 & -0.2796 & -0.2631 & 0.3371 & 0.7498 & -0.1187 & -0.1330 \\ -0.3805 & -0.3416 & 0.0830 & 0.7577 & -0.1660 & 0.3217 & 0.1627 \\ -0.0676 & 0.7823 & -0.4804 & 0.3693 & -0.0206 & 0.0519 & -0.1144 \\ 0.4851 & 0.0447 & -0.1705 & -0.0093 & -0.1052 & 0.2175 & 0.8217 \\ -0.3449 & 0.3649 & 0.5048 & -0.0671 & 0.6001 & 0.1531 & 0.3241 \end{bmatrix}$$

Variance explained by principal components in order is 56.12, 16.71, 12.51, 6.58, 5.67, 1.57, and 0.79%.

8.3 With respect to data presented in Exercise 8.2, check that whether the last three principal components do not explain same amount of variance. (Ans. The last three principal components statistically explain same amount of variance.)

8.4 With respect to data presented in Exercise 8.2, considering the sea surface temperature as dependent variable and all other as independent variables, calculate the loading vector for supervised principal component in the data set. (Ans. The SPC loadings are 0.040, 0.995, −0.022, −0.003, 0.074, and −0.041.)

Table 8.8 Monthly average sea surface temperature and other hydroclimatic variables

Month	Sea surface temperature (°C)	Air temperature (°C)	Rainfall (mm)	Zonal wind speed (m/s)	Meridional wind speed (m/s)	Pressure (mb)	Specific humidity (Kg/m^3)
1	21.15	25.67	9.72	3.05	−0.37	951.46	13.08
2	21.66	27.59	0.65	3.45	−0.05	952.49	17.53
3	21.89	28.32	8.68	3.02	−0.47	953.89	17.6
4	22.68	30.15	0.26	−0.53	0.21	955.89	16.54
5	24.65	28.86	7.95	0.03	−1.48	960.02	11.17
6	24.59	30.89	278.70	−2.13	−1.99	964.55	7.71
7	26.04	31.76	229.60	−1.63	−0.47	965.82	5.66
8	25.32	33.32	239.60	0.22	−0.26	963.47	3.54
9	25.16	30.03	206.63	1.62	−1.55	962.94	3.02
10	23.08	28.35	7.70	0.52	−0.74	960.40	5.37
11	22.50	25.54	0.12	2.40	−2.79	957.02	3.81
12	21.20	24.11	16.17	0.91	−2.31	952.43	6.81
13	22.12	29.08	9.93	3.64	−0.33	951.99	13.35
14	23.05	29.31	0.00	2.85	−0.20	953.16	17.55
15	23.12	30.89	58.92	2.71	−1.84	952.71	17.63
16	24.15	28.95	0.10	2.19	−1.17	956.59	15.82
17	25.65	35.60	5.91	−1.25	−2.20	960.64	11.37
18	27.54	33.51	71.50	−1.69	−0.45	965.28	7.29
19	26.12	33.87	86.58	−1.69	0.15	966.31	6.58
20	25.61	30.50	261.97	0.59	−2.87	964.99	3.43
21	23.85	29.11	253.77	0.23	0.19	962.42	5.88
22	23.23	30.21	25.25	0.44	−0.56	961.14	6.04
23	24.36	31.01	1.92	2.17	−1.22	956.68	5.31
24	22.04	25.83	0.00	1.16	1.22	953.53	10.34
35	20.32	23.68	48.67	3.31	−0.26	953.75	14.26
36	21.94	26.56	30.87	1.25	−0.50	950.47	18.2
37	23.48	27.66	10.71	0.74	−2.64	952.79	20.12
38	23.37	30.21	4.51	−0.35	−0.52	954.92	17.24
39	25.73	31.15	35.19	−1.60	−0.88	959.69	12.43
30	25.06	30.08	73.87	−3.08	0.52	965.23	7.55
31	26.14	28.90	266.09	−3.39	−0.59	966.2	4.83
32	26.07	32.94	332.62	0.69	−3.19	963.52	3.11
33	23.14	29.39	152.16	−0.67	−0.51	963.46	3.69
34	23.50	28.84	74.71	0.59	−0.27	961.39	5.14
35	23.25	30.80	67.11	3.07	−2.23	957.50	5.12
36	20.68	24.73	0.98	1.03	−1.69	952.55	8.02

8.5 Following observations are recorded daily for 20 consecutive days in a city.

Days	Evaporation	Air temperature	Pressure	Specific humidity	Wind speed
	(mm/day)	(°C)	(mb)	(Kg/m^3)	(m/s)
1	3.08	24.32	954.46	11.07	6.75
2	2.43	21.74	955.41	7.59	7.50
3	2.70	20.17	959.64	5.82	5.57
4	3.06	21.82	956.75	7.32	4.27
5	3.17	23.63	955.27	5.36	5.38
6	3.36	27.73	949.16	5.23	0.31
7	4.15	32.11	942.58	6.23	5.51
8	3.09	34.48	937.95	7.82	6.23
9	3.58	32.61	940.43	12.62	1.26
10	2.57	26.90	949.25	15.98	1.84
11	1.48	24.17	955.12	17.62	0.61
12	2.53	24.54	950.78	15.39	2.36
13	4.07	23.70	954.34	12.69	6.19
14	2.66	23.65	955.08	8.80	1.73
15	3.33	28.47	946.57	11.84	4.01
16	2.50	24.32	954.37	11.79	4.67
17	2.09	22.17	956.82	11.72	3.09
18	2.80	23.18	954.10	11.36	3.31
19	3.62	23.67	954.91	9.03	5.79
20	3.01	25.69	951.27	7.02	1.02

Calculate the supervised principal component considering evaporation as dependent variable.

(Ans. Supervised principal component considering evaporation as dependent variable is −679.7, −679.9, −683.3, −681.4, −678.3, −672.9, −665.4, −661.4, −667.3, −677.8, −684.4, −679.7, −680.7, −680.4, −672.8, −680.3, −683.5, −680.8, −679.6, and −676.)

8.6 Considering sea surface temperature and air temperature as dependent variables and other variables as independent variable in Exercise 8.2, calculate the loadings for

(a) Supervised principal component
(b) Canonical correlation component

Ans. (a) Loading vector for supervised principal component

$$U = \begin{bmatrix} 0.996 & -0.022 & -0.002 & 0.080 & -0.037 \\ -0.038 & -0.013 & 0.051 & 0.769 & 0.636 \end{bmatrix}^T$$

(b) Loading vector of independent variables for canonical correlation component

$$U = \begin{bmatrix} -0.0046 & 0.0046 & 0.0044 & -0.0046 & 0.0044 \\ 0.3234 & -0.3242 & -0.2895 & 0.3364 & -0.2987 \\ -0.4699 & 0.4694 & 0.4910 & -0.4615 & 0.4853 \\ 0.7624 & -0.7623 & -0.7644 & 0.7613 & -0.7640 \\ 0.3056 & -0.3057 & -0.3013 & 0.3071 & -0.3025 \end{bmatrix}$$

Similarly, the loading vector for dependent variables is

$$V = \begin{bmatrix} 0.999 & -0.779 \\ -0.047 & 0.627 \end{bmatrix}.$$

8.7 For nine different locations in lower Narmada Basin, the monthly average precipitation (in mm) for 2 years is presented in the following table.

Month	Locations								
	A	B	C	D	E	F	G	H	I
1	16.06	17	27.67	16.63	11.07	19.69	10.49	10.62	11.82
2	13.52	15.47	19.62	17.15	8.44	18.09	9.86	10.84	8.69
3	10.55	8.08	14.91	17.93	7.01	9.66	5.01	5.98	8.23
4	6.00	7.26	15.84	13.6	9.68	8.57	6.28	11.55	10.32
5	17.73	12.26	20.84	22.98	9.81	14.96	13.45	10.35	14.15
6	166.7	163.38	199.45	214.83	182.77	192.6	221.41	179.13	176.47
7	286.17	335.33	331.95	350.86	343.31	336.88	438.6	335.86	312.03
8	298.75	287.19	332.08	351.87	309.21	324.8	461.72	351.64	306.51
9	156.47	159.31	217.76	171.27	189.43	190.72	178.17	170.56	171.18
10	50.04	52.69	54.46	57.13	43.08	49.79	60.96	49.19	46.49
11	7.13	7.01	12.69	10.34	5.39	6.93	4.73	9.11	6.54
12	10.3	7.08	9.82	8.74	5.23	6.65	2.98	3.55	5.49
13	72.11	119.22	26.42	28.08	26.42	0.64	6.81	46.87	29.25
14	49.10	76.07	74.23	27.95	34.45	97.57	32.46	33.77	42.73
15	0.58	26.98	67.47	50.9	36.08	32.61	11.06	61.04	3.85
16	17.69	3.86	14.44	4.37	24.96	16.71	20.41	60.04	2.39
17	17.52	5.23	19.08	59.96	20.29	28.52	58.06	63.06	32.37
18	209.51	178.79	186.59	224.1	203.91	237.33	285.77	172.05	190.35
19	300.24	380.61	320.12	340.26	329.33	356.18	350.21	360.4	312.54
20	250.33	282.35	280.6	288.46	363.54	300.78	288.54	322.15	371.04
21	198.59	195.46	240.21	201.8	286.6	213.29	178.09	211.15	197.2
22	34.64	18.37	80.24	44.21	80.05	80.36	97.56	74.69	43.22
23	18.91	17.44	15.03	19.3	9.8	8.39	12.53	53.68	34.25
24	31.63	23.56	25.63	34.72	16.71	36.64	4.24	16.3	9.53

Calculate the loadings for empirical orthogonal components.
Ans. The loadings for empirical orthogonal components are

$$U = \begin{bmatrix} 0.289 & -0.047 & 0.297 & -0.113 & -0.295 & -0.438 & -0.152 & -0.258 & 0.667 \\ 0.316 & -0.369 & 0.756 & -0.003 & 0.077 & 0.052 & -0.082 & 0.180 & -0.380 \\ 0.318 & -0.063 & -0.193 & -0.491 & 0.429 & -0.086 & 0.175 & 0.574 & 0.255 \\ 0.334 & 0.182 & 0.050 & 0.066 & 0.046 & -0.129 & 0.834 & -0.335 & -0.147 \\ 0.348 & -0.358 & -0.481 & -0.053 & -0.044 & -0.467 & -0.257 & -0.205 & -0.434 \\ 0.335 & 0.020 & -0.087 & -0.503 & -0.135 & 0.655 & -0.145 & -0.398 & -0.006 \\ 0.394 & 0.792 & 0.048 & 0.101 & -0.104 & -0.099 & -0.301 & 0.220 & -0.213 \\ 0.329 & -0.070 & -0.082 & 0.566 & 0.601 & 0.213 & -0.194 & -0.211 & 0.266 \\ 0.327 & -0.251 & -0.229 & 0.394 & -0.573 & 0.282 & 0.170 & 0.409 & 0.137 \end{bmatrix}$$

8.8 For the precipitation data given in Exercise 8.7, the locations 'A' and 'B' are in downstream to all other points. Calculate the loading vector for canonical correlation component considering precipitation at location 'A' and 'B' as dependent variable.

Ans. The canonical correlation loading for independent variables is

$$U = \begin{bmatrix} 1.147 & 0.1470 & 0.1470 & 0.1470 & 0.1460 & 0.1470 & 0.147 \\ 0.582 & 0.5830 & 0.5820 & 0.5820 & 0.5870 & 0.5800 & 0.581 \\ -0.324 & -0.3230 & -0.3240 & -0.3240 & -0.3200 & -0.3260 & -0.325 \\ 0.420 & 0.4200 & 0.4200 & 0.4200 & 0.4220 & 0.4190 & 0.420 \\ -0.374 & -0.3740 & -0.3740 & -0.3740 & -0.3710 & -0.3750 & -0.374 \\ 0.375 & 0.3740 & 0.3740 & 0.3740 & 0.3710 & 0.3760 & 0.375 \\ 0.279 & 0.2800 & 0.2790 & 0.2790 & 0.2810 & 0.2790 & 0.279 \end{bmatrix}$$

The canonical correlation loadings for dependent variables are

$$V = \begin{bmatrix} 0.9940 & -0.7350 \\ -0.1080 & 0.6780 \end{bmatrix}$$

8.9 Calculate the variance explained by EOFs of mean monthly air temperature recorded in five cities (A to E as shown in Table 8.9).

Table 8.9 Air temperature for five monitoring stations

Month	Location				
	A	B	C	D	E
1	21.70	19.94	19.92	25.07	24.52
2	23.60	23.55	18.94	23.51	23.12
3	22.44	25.96	17.40	26.59	26.30
4	20.37	24.29	19.05	23.18	22.09
5	22.73	23.16	20.12	24.06	28.91
6	23.99	20.36	20.67	22.12	22.43
7	21.08	22.07	18.43	20.85	27.03
8	21.62	23.17	15.88	21.59	22.57
9	22.67	19.80	18.20	20.69	24.12

(continued)

Table 8.9 (continued)

Month	Location				
	A	B	C	D	E
10	24.39	23.08	19.86	20.60	25.23
11	21.11	21.91	17.80	23.11	25.94
12	22.02	21.54	15.76	21.95	21.43
13	18.79	22.33	23.05	19.45	19.56
14	22.12	20.06	13.86	22.36	21.55
15	19.31	19.55	18.49	21.54	20.76
16	20.11	21.94	16.10	24.39	22.58
17	21.96	20.54	18.55	20.40	23.24
18	21.01	20.18	20.13	22.29	23.29
19	21.78	19.20	20.52	22.85	21.68
20	23.45	22.25	16.85	24.52	25.62
21	25.23	19.64	21.57	21.06	25.83
22	23.77	23.34	19.55	22.84	23.84
23	22.06	25.86	20.23	19.80	27.33
24	22.57	22.82	17.35	17.92	23.71

(Ans. The variance explained by first five empirical orthogonal components in % is 36.51, 25.59, 15.98, 14.43, and 7.50.)

8.10 At a gauging station, the monthly streamflow is found to follow exponential distribution with $\lambda = 0.5$. Generate a streamflow data for a year.

(Answers may vary depending on random number generated. Refer to Sect. 8.6.)

8.11 Historical data for a location suggests that monthly average rainfall follows exponential distribution with $\lambda = 1.5$ and streamflow follows normal distribution with mean 15 m^3/s and standard deviation of 2.5 m^3/s. The correlation between monthly average precipitation and streamflow is 0.55. Generate the data for 2 years preserving the correlation structure. (Answers may vary depending on random number generated. Refer to Sect. 8.6.)

8.12 Following annual precipitation depths (in cm) is obtained from 4 GCMs for 6 consecutive years.

GCM	Years					
	2025	2026	2027	2028	2029	2030
GCM-1	112.2	117.9	104.3	111.7	112.6	115.2
GCM-2	133.8	117.3	125.4	133.6	128.8	134.8
GCM-3	127.2	88.8	111.6	109.8	115.6	131.4
GCM-4	138.4	111.7	100.8	129.2	124.6	112.1

Check whether the mean annual precipitation depth differs with GCMs at 5% level of significance.

(Ans. Mean annual precipitation depth differs with GCMs at 5% significance level.)

8.13 Different scenarios for GCM result in the different predictions for the hydroclimatic variables. For a location following estimate for average annual streamflow, (Mm^3) is obtained for 4 GCM and four different scenarios.

GCM	Scenarios			
	1	2	3	4
GCM-1	250	261	280	271
GCM-2	265	270	264	259
GCM-3	245	252	250	268
GCM-4	240	255	259	272

Check whether the mean annual streamflow differs with different choices of GCMs and/or scenarios at 1% level of significance.

(Ans. Mean annual streamflow differs with GCMs and/or scenarios at 1% significance level.)

Chapter 9
Time Series Analysis

Hydroclimatic variables such as rainfall intensity, streamflow, air temperature vary with space and time, due to different hydrological/climatic phenomena/processes. As these processes are continuously evolving over time, studying the interdependence in hydroclimatic data with proper consideration of temporal information may lead to better insight into the governing processes. Observations of any variable, recorded in chronological order, represent a time series. A time series is generally assumed to consist of deterministic components (results can be predicted with certainty) and stochastic components (results cannot be predicted with certainty as the outcome depends on chance). Analysis of time series helps to get an insight of the time series that in turn may enhance the prediction of the hydroclimatic processes/variables. The objective of this chapter is to introduce different types of time series analysis techniques. This requires an understanding of time series analysis techniques and time series properties like stationarity, homogeneity, periodicity, which is the subject matter of this chapter.

9.1 Data Representation in Hydroclimatology

Most of the hydrologic time series are continuous in nature but they need to be represented on a discrete time interval. For example, temperature, streamflow, or rainfall depth may vary continuously over time but the records are taken over discrete time interval. There are two methods to represent a continuous time series, $f(t)$, on a discrete time interval.

(i) **Sample Data Representation**: In this representation, value of the function for ith time interval, $X(i)$, is given by instantaneous value of $f(t)$ at the time $i\Delta t$.

$$X(i) = f(i\Delta t) \tag{9.1}$$

Dimension of the pulse data is L^3T^{-1} or LT^{-1}. Most common examples of sample data representation include streamflow, wind speed that are recorded as a series of instantaneous values.

© Springer Nature Singapore Pte Ltd. 2018
R. Maity, *Statistical Methods in Hydrology and Hydroclimatology*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-10-8779-0_9

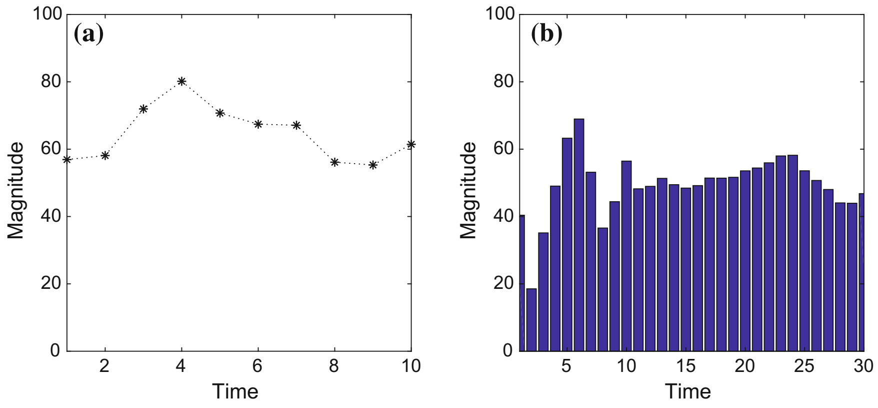


Fig. 9.1 Different types of data representation **a** sample data representation **b** pulse data representation

(ii) **Pulse Data Representation**: In this representation, value of the function for ith time interval, $X(i)$, is given by accumulated value of $f(t)$ during time $i\Delta t$, i.e., between $(i-1)\Delta$ and $i\Delta$.

$$X(i) = \int_{(i-1)\Delta t}^{i\Delta t} f(t)dt \tag{9.2}$$

Dimension of the pulse data is L^3 or L. Most common example of pulse data representation is precipitation that is recorded as a series of accumulated depths. Sometimes pulse data can also be represented as average rate over the interval Δt as follows:

$$X(i) = \frac{1}{\Delta t}\int_{(i-1)\Delta t}^{i\Delta t} f(t)dt \tag{9.3}$$

Example includes precipitation intensity that has a dimension LT^{-1}.

Figure 9.1 explains the two methods of data representation, i.e., sample data representation and pulse data representation.

9.2 Stationary and Non-stationary Time Series

A time series is known to be stationary if the statistical properties of the time series remain constant over time. This property is known as *stationarity*. The order of the stationarity represents the highest central moment (moment around the mean), which remain constant over time. For instance, first-order stationarity indicates time-invariant mean or mean does not change over time. Similarly, if both mean and variance (second-order central moment) remain constant over time, the time series is known to be *second-order stationary* or *weakly stationary*. If mean, variance, and all higher-order moments are constant over time, the time series is called *strict sense stationary* or simply *stationary*. In hydrologic and hydroclimatic applications, second-order stationary can be safely assume to be satisfactory. However, impacts of climate change may impart non-stationarity in many hydrologic time series.

If the statistical properties of a time series change or vary with time, it is known as non-stationary time series. Apart from various other causes, presence of trend, jump, periodicity, and a combination thereof, cause non-stationarity in the time series. These are generally deterministic components that should be removed to obtain the stochastic component of the time series. However, their removal does not always guarantee stationarity. These deterministic components are discussed below:

(i) **Trend**: Trend refers to gradual but continuous change in mean of a time series. Trend may be increasing or decreasing (Fig. 9.2) and may be linear or nonlinear. The cause of trend in time series is gradual change in hydrological and climatic factors or conditions. Sometimes anthropogenic changes (like change in land use and land cover, regulation of river flow using weir) may also lead to development of trend in hydroclimatic time series.

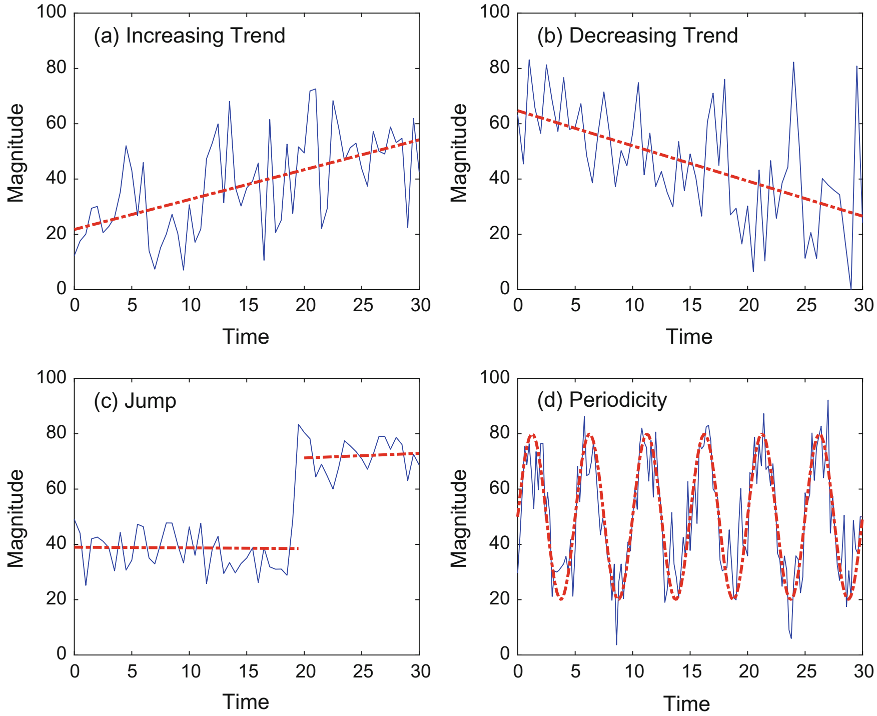


Fig. 9.2 Different types of deterministic components of time series

(ii) **Jump**: An abrupt change in the mean of the time series at some time step is termed as jump (Fig. 9.2). Jump in hydrological time series may occur due to extreme conditions like natural hazards; system errors; inhomogeneity caused by humans or change in experimental method/setup/tools. The removal of jump requires identification of the time step of its occurrence.

(iii) **Periodicity**: Periodicity is a property of time series in which the same or similar values get repeated after some time difference (Fig. 9.2). The periodicity is observed in many hydrological or climatic variable due to seasonality. For example, in India the rainfall is highly seasonal and mostly occurs in the monsoon months (June, July, August, September). On visualization, periodic time series show wave like characteristics. The time series that do not exhibit periodicity is termed as aperiodic.

9.3 Ensemble and Realization

Ensemble refers to a collection of time series representing the same variable. Each of the constituting time series of an ensemble is termed as *realization*. The statistical properties across the different realizations are known as ensemble properties. An ensemble is said to be '*ergodic*' if the statistical properties remain constant across the realizations within the ensemble, otherwise the ensemble is termed as '*non-ergodic*.' Often ensembles are generated in many hydroclimatic simulations through multiple runs of the model, and ensemble average properties are determined.

9.4 Trend Analysis

Deterministic components, if exist, should be treated separately. Thereby, the trend in time series (if any) needs to be identified and removed before applying any time series model.

9.4.1 Tests for Randomness and Trend

The trend (if any) is usually visible in the time series plot. The presence of trend in time series can be checked using the following tests:

(i) **Regression test for linear trend**: If the time series fulfils the assumptions of simple regression model, a linear regression model can be fitted by considering the value of time series as dependent variable and time step as independent variable as per Sect. 7.1. The regression equation is expressed as:

$$X(t) = a + bt + \varepsilon \tag{9.4}$$

where, $X(t)$ is the value of time series at time t, a and b are the intercept and slope parameters of regression model respectively. ε represents the residual or error. Increasing and decreasing (linear) trend with fitted regression line is shown in Fig. 9.2. If the slope of the fitted regression model (b) is not significantly different from zero, then no linear trend exists in the time series. This test for significance of b is explained in Sect. 7.7.

(ii) **Mann–Kendall Test**: Mann–Kendall test is a nonparametric test that identifies the trend in the time series. Being a nonparametric test, the test is widely applied to detect trend in time series following any probability distribution. For a time series $X(t)$, the Mann–Kendall statistic is defined as:

$$S = \sum_{t=1}^{N-1} \sum_{t'=t+1}^{N} sign(X(t') - X(t)) \tag{9.5}$$

where N is number of data and $sign(\bullet)$ represents a *signum* function given by:

$$sign(a) = \begin{cases} \frac{a}{|a|}, & \text{if } x \neq 0 \\ 0, & \text{if } x = 0 \end{cases} = \begin{cases} 1, & \text{if } a > 0 \\ 0, & \text{if } a = 0 \\ -1, & \text{if } a < 0 \end{cases} \tag{9.6}$$

Sign and value of the S statistic show the direction and intensity of the trend. Under the null hypothesis of no trend, the distribution of S statistics is expected to have zero mean. The variance of the statistics is given by:

$$Var(S) = \frac{1}{18}\left[N(N-1)(2N+5) - \sum_{i=1}^{g} t_i(t_i - 1)(2t_i + 5)\right] \tag{9.7}$$

where g is the number of tied groups and t_i represents the number of observations in the tied group. Tied groups are groups having members tied, or, in other words if the frequency of a value is greater than 1 in the frequency table, it constitutes the tied group. For example, in the data set {15, 11, 10, 12, 10, 15, 13, 15} there are two tied groups (10 and 15). Tied group for 10 has 2 members and tied group for 15 has 3 members. However, continuous hydroclimatic variables like precipitation, stream flow, temperature may have very less or no tied group. Under the assumption that there is no tied group, the variance of S statistic becomes:

$$Var(S) = \frac{N(N-1)(2N+5)}{18} \tag{9.8}$$

The test statistics u_c is given by:

$$u_c = \frac{S - sign(S)}{\sqrt{Var(S)}} \tag{9.9}$$

u_c statistic follows standardized normal distribution. The null hypothesis of no trend can be rejected if $|u_c| > Z_{(\alpha/2)}$, where $Z_{(\alpha/2)}$ is standardized normal variate for the non-exceedance probability of $(1 - \alpha/2) \times 100\%$ and α is the level of significance. For no tied group, the test is valid for $N > 10$.

(iii) **Kendall tau (τ) Test (Rank Correlation Test)**: Suppose that for a pair in $(X(i), X(j))$ with $j > i$ in time series $X(t)$, there are p pairs such that $X(j) > X(i)$. These pairs are called concordant pairs.

$$p = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \Im(X(j) > X(i)) \tag{9.10}$$

where, $\Im(\bullet) = 1$ if the argument is true, otherwise $\Im(\bullet) = 0$. The pairs $(X(i), X(j))$ with $j > i$ are called discordant, if $X(i) > X(j)$. It should be noted that the pairs can be neither concordant or discordant if $X(i) = X(j)$. The random variable p, i.e., number of concordant pairs, is supposed to have uniform distribution between minimum possible value (i.e., 0) and maximum possible value. Maximum possible number of concordant pairs (p) will be equal to the number of possible pairs in a strictly increasing time series ($\forall j > i$, $X(j) > X(i)$). If the length of time series is N, then

$$p = (N-1) + (N-2) + \cdots + 1 = \frac{N(N-1)}{2} \tag{9.11}$$

Hence, expected value of p is given as:

$$E(p) = \frac{1}{2}\left(0 + \frac{N(N-1)}{2}\right) = \frac{N(N-1)}{4} \tag{9.12}$$

Test statistic for testing the randomness is defined as Kendall rank correlation coefficient or Kendall τ. The Kendall τ is defined as the difference between the probability of concordant and discordant pairs. If the number of concordant and discordant pairs are p and q, respectively, then the Kendall τ is given by:

$$\tau = \frac{2(p-q)}{N(N-1)} \tag{9.13}$$

If there is no pair that is neither concordant or discordant, then $p + q =$ Total number of pairs $= N(N-1)/2$. In that case, the Eq. 9.13 reduces to,

$$\tau = \frac{4p}{N(N-1)} - 1 \tag{9.14}$$

Kendall τ varies between ±1 with an expected value of 0. The variance of τ is given by:

$$\mathrm{Var}(\tau) = \frac{2(2N+5)}{9N(N-1)} \tag{9.15}$$

With large N, the ratio $\frac{\tau - E(\tau)}{\sqrt{\mathrm{Var}(\tau)}} = \frac{\tau}{\sqrt{\mathrm{Var}(\tau)}}$ follows a standard normal distribution. Hence, if $\left|\frac{\tau}{\sqrt{\mathrm{Var}(\tau)}}\right| < Z_{(\alpha/2)}$ at significance level α then the null hypothesis (time series is random) cannot be rejected. $Z_{(\alpha/2)}$ is the standard normal variate at $(1 - \alpha/2) \times 100\%$ non-exceedance probability.

9.4.2 Trend Removal

The trend can be linear or nonlinear. Removing linear trend by fitting simple regression model is comparatively easier. In case of nonlinear trend, the major problem becomes the estimation of degree of polynomial trend line. With the increase in the degree of fitted polynomial the residual may decrease, however, it needs to be checked whether addition of extra order is statistically significant or not. One alternative approach to remove nonlinear trend is piecewise polynomial regression fitting. Hence, instead of fitting a global polynomial trend line, many lower-order polynomials are fitted in piecewise manner. One popular polynomial fit for piecewise fitting is spline. After fitting an appropriate trend line (either a simple linear regression or polynomial fit) the value of trend can be calculated and subtracted from the observed time series to remove the trend.

Example 9.4.1
The streamflow records (in Mm^3) for 20 consecutive days are 1.1, 0.5, 2.7, 1.3, 1.5, 2.2, 2.1, 3, 2.9, 4.4, 4.6, 3.1, 4.7, 4, 4.6, 5.1, 6.1, 5.3, 6.7, and 5.6. Check the streamflow data for linear trend by using linear regression and comment about the significance of linear trend at 5% significance level.

Solution The given data can be analyzed for fitting a linear regression as given in Table 9.1 (see Example 7.1.2). From the table,

$$N = 20, \sum t = 210, \sum x = 71.5, S_{tt} = \sum {t_d}^2 = 665, \text{ and}$$

$$S_{xx} = \sum {x_d}^2 = 60.73, S_{tx} = S_{xt} = \sum t_d x_d = 188.115$$

Table 9.1 Calculation for fitting linear regression

Days (t)	Streamflow (in Mm^3) (x)	t_d	x_d	t_d^2	x_d^2	$x_d t_d$
1	1.1	−9.5	−2.48	90.25	6.13	23.56
2	0.5	−8.5	−3.08	72.25	9.46	26.18
3	2.7	−7.5	−0.88	56.25	0.77	6.6
4	1.3	−6.5	−2.28	42.25	5.18	14.82
5	1.5	−5.5	−2.08	30.25	4.31	11.44
6	2.2	−4.5	−1.38	20.25	1.89	6.21
7	2.1	−3.5	−1.48	12.25	2.18	5.18
8	3.0	−2.5	−0.58	6.25	0.33	1.45
9	2.9	−1.5	−0.68	2.25	0.46	1.02
10	4.4	−0.5	0.83	0.25	0.68	−0.415
11	4.6	0.5	1.02	0.25	1.05	0.51
12	3.1	1.5	−0.48	2.25	0.23	−0.72
13	4.7	2.5	1.13	6.25	1.27	2.825
14	4.0	3.5	0.42	12.25	0.18	1.47
15	4.6	4.5	1.02	20.25	1.05	4.59
16	5.1	5.5	1.52	30.25	2.33	8.36
17	6.1	6.5	2.52	42.25	6.38	16.38
18	5.3	7.5	1.72	56.25	2.98	12.9
19	6.7	8.5	3.13	72.25	9.77	26.605
20	5.6	9.5	2.02	90.25	4.10	19.19
Total $\sum$: 210	71.5			665	60.73	188.155

Simple linear regression equation for trend is given by

$$x = a + bt$$

$$b = \frac{S_{tx}}{S_{tt}} = \frac{188.115}{665} = 0.283$$

$$\text{and, } a = \overline{x} - b\overline{t} = \frac{71.5 - 0.283 \times 210}{20} = 0.604$$

Hence, the developed linear regression model is

$$x = 0.604 + 0.283t$$

The Sum of squared errors is given by

$$S_e^2 = \frac{S_{xx} - (S_{tx})^2 / S_{tt}}{N-2} = \frac{60.73 - 188.115^2 / 665}{18} = 0.418$$

For checking the significance of the trend in the streamflow time series, we need to prove that the parameter β (population estimate of b; Sect. 7.7) is statically different from 0 at 5% significance level. Hence,

Null Hypothesis: $\beta = 0$
Alternative Hypothesis: $\beta \neq 0$
Level of Significance: $\alpha = 5\%$

For $N - 2 = 18$ degrees of freedom, $t_{0.025}(18) = 2.10$.
The test statistics is given by

$$t = \frac{b - \beta}{S_e}\sqrt{S_{tt}} = \frac{0.283}{\sqrt{0.418}}\sqrt{665} = 11.29$$

As $11.29 > t_{0.025}(18)$, so the null hypothesis is rejected. Hence, the trend is significant in the streamflow time series at 5% significance level.

Example 9.4.2
For the time series given in Example 9.4.1, test the significance of trend using (a) Mann–Kendall Test and (b) Kendall's tau test at 10% significance level.

Solution The null and alternate hypothesis can be expressed as:

Null Hypothesis: Time Series does not have a trend.
Alternative Hypothesis: Time Series has a trend.
Level of Significance: $\alpha = 10\%$

(i) **Mann–Kendall Test**
From Eq. 9.5, the Mann–Kendall statistics (S) is given by

$$S = \sum_{t=1}^{N-1} \sum_{t'=t+1}^{N} sign(X(t') - X(t)) = 157$$

The variance of S is given by

$$\text{Var}(S) = \frac{N(N-1)(2N+5)}{18} = 950$$

The test statistics u_c is

$$u_c = \frac{S - sign(S)}{\sqrt{\text{Var}(S)}} = \frac{157 - 1}{\sqrt{950}} = 5.06$$

$$Z_{(\alpha/2)} = Z_{0.05} = 1.645$$

As $|u_c| > 1.6450$ ($Z_{0.05}$), so the null hypothesis of no trend is rejected.

(ii) **Kendall tau Test**

The number of pairs of concordant pairs (p) as defined by Eq. 9.10 is 173 in the streamflow time series. From Eq. 9.14, the Kendall tau (τ) is given by

$$\tau = \frac{4p}{N(N-1)} - 1 = \frac{4 \times 173}{20(20-1)} - 1 = 0.82$$

The variance of τ is given by (Eq. 9.15)

$$\text{Var}(\tau) = \frac{2(2N+5)}{9N(N-1)} = 0.0263$$

The test statistics (z) is

$$z = \frac{\tau}{\sqrt{\text{Var}(\tau)}} = \frac{0.82}{\sqrt{0.0263}} = 5.06$$

As $|z| > 1.645$ ($Z_{0.05}$), so the null hypothesis of no trend is rejected.

Example 9.4.3

For the years 1981 to 2010, the global mean annual temperature (in °C) was observed as 0.33, 0.13, 0.30, 0.15, 0.12, 0.19, 0.33, 0.41, 0.28, 0.44, 0.43, 0.23, 0.24, 0.32, 0.46, 0.35, 0.48, 0.64, 0.42, 0.42, 0.55, 0.63, 0.62, 0.55, 0.69, 0.63, 0.66, 0.54, 0.64, and 0.71. Check the claim that global mean annual temperature has no trend using Kendall tau test at 5% significance level.

Solution The null and alternate hypothesis can be expressed as:

Null Hypothesis: Global mean annual temperature does not have a trend.
Alternative Hypothesis: Null hypothesis is not true.
Level of Significance: $\alpha = 5\%$

Kendall tau Test

The number of pairs of concordant (p) as defined by Eq. 9.10 is 362 in the streamflow time series. From Eq. 9.13, the Kendall tau (τ) is given by

$$\tau = \frac{2(p-q)}{N(N-1)} = \frac{4p}{N(N-1)} - 1 = \frac{4 \times 362}{30(30-1)} - 1 = 0.66$$

The variance of τ is given by (Eq. 9.15)

$$\text{Var}(\tau) = \frac{2(2N+5)}{9N(N-1)} = 0.017$$

The test statistics (z) is

$$z = \frac{\tau}{\sqrt{\text{Var}(\tau)}} = \frac{0.66}{\sqrt{0.017}} = 5.06$$

As $|z| > 1.96$ ($Z_{0.025}$), so the null hypothesis of no trend is rejected.

9.5 Analysis of Periodicity

In the domain of hydroclimatology, many time series are having periodicity due to their seasonal behavior. For example, monthly rainfall or wind velocity at a location is expected to have a periodicity of 12 months. If the time period of periodicity is known, then it can be removed using harmonic analysis, otherwise time period of periodicity can be identified using autocorrelation or spectral analysis.

9.5.1 Harmonic Analysis

Any time series can be expanded into Fourier series, i.e., a function of series of sines and cosines.

$$X(t) = a_0 + \sum_{i=1}^{\infty} a_i \cos(2\pi\nu_i t) + \sum_{i=1}^{\infty} b_i \sin(2\pi\nu_i t) \tag{9.16}$$

where ν_i is ith frequency and a_i, b_i are corresponding Fourier coefficients. If the length of data is N, then the coefficients are given by,

$$a_i = \frac{1}{N}\sum_{t=1}^{N} X(t)\cos\left(\frac{2\pi i t}{N}\right)dt \qquad \text{for } t = 0, 1, 2, \ldots \tag{9.17}$$

$$b_i = \frac{1}{N}\sum_{t=1}^{N} X(t)\sin\left(\frac{2\pi i t}{N}\right)dt \qquad \text{for } t = 1, 2, 3, \ldots \tag{9.18}$$

If the periodicity is known (say p) then different harmonics or frequencies are expressed as τ/p where $\tau = 1, 2, \ldots, p$. For hydroclimatic data p depends upon the temporal resolution. For example, p for monthly scale data is 12 and for daily scale data is 365. In discrete form, the harmonic fitted mean of such hydroclimatic time series for a period τ (say m_τ) using first h harmonics is given by,

$$m_\tau = \mu + \sum_{i=1}^{h} a_i \cos\left(\frac{2\pi i \tau}{p}\right) + \sum_{i=1}^{h} b_i \sin\left(\frac{2\pi i \tau}{p}\right) \tag{9.19}$$

where μ is the population mean, a_i and b_i are Fourier parameters and h is total numbers of harmonics considered. The Fourier parameters can be obtained by minimizing the Sum of square of differences between the sample estimate of mean and mean estimated using Eq. 9.19. The parameters are given by:

$$a_i = \frac{2}{p} \sum_{\tau=1}^{p} \overline{X}_\tau \cos\left(\frac{2\pi i \tau}{p}\right), \text{for } i = 1, 2, \ldots, h \tag{9.20}$$

$$b_i = \frac{2}{p} \sum_{\tau=1}^{p} \overline{X}_\tau \sin\left(\frac{2\pi i \tau}{p}\right), \text{for } i = 1, 2, \ldots, h \tag{9.21}$$

If we consider all the harmonics, then the m_τ will be equal to actual periodic mean (x_τ). In practice only first few significant harmonics can explain most of the variance in the data. The number of significant harmonics required can be ascertained by plotting the ratio of cumulative variability explained by individual harmonics to total variability in the time series. The plot of explained cumulative variance with respect to order of harmonics is called **cumulative periodogram**.

$$P_j = \frac{\sum_{i=1}^{j} Var(h_i)}{Var(x)} \tag{9.22}$$

where $Var(h_i)$ and $Var(x)$ are the mean square of deviation of m_τ (for harmonics h_i) and x from their respective means. These quantities are given by:

$$Var(h_i) = \frac{1}{2}(a_i^2 + b_i^2), \text{for } i = 1, 2, \ldots, h \tag{9.23}$$

$$Var(x) = \frac{1}{p} \sum_{\tau=1}^{p} (\overline{x}_\tau - \hat{\mu})^2 \qquad \text{where, } \hat{\mu} = \frac{1}{p} \sum_{\tau=1}^{p} \overline{x}_\tau \tag{9.24}$$

Hence, the cumulative periodogram is plotted between P_j and j. The slope of the cumulative periodogram helps in finding the significant number of harmonics. The m_τ thus obtained can be subtracted from the original time series to remove the periodicity from the time series.

9.5.2 *Spectral Analysis*

Spectral analysis, also called spectral density estimation or frequency domain analysis, is the decomposition of a periodic time series in such a way that its constituent frequency (and their amplitude) is revealed. Spectral density estimation can be done

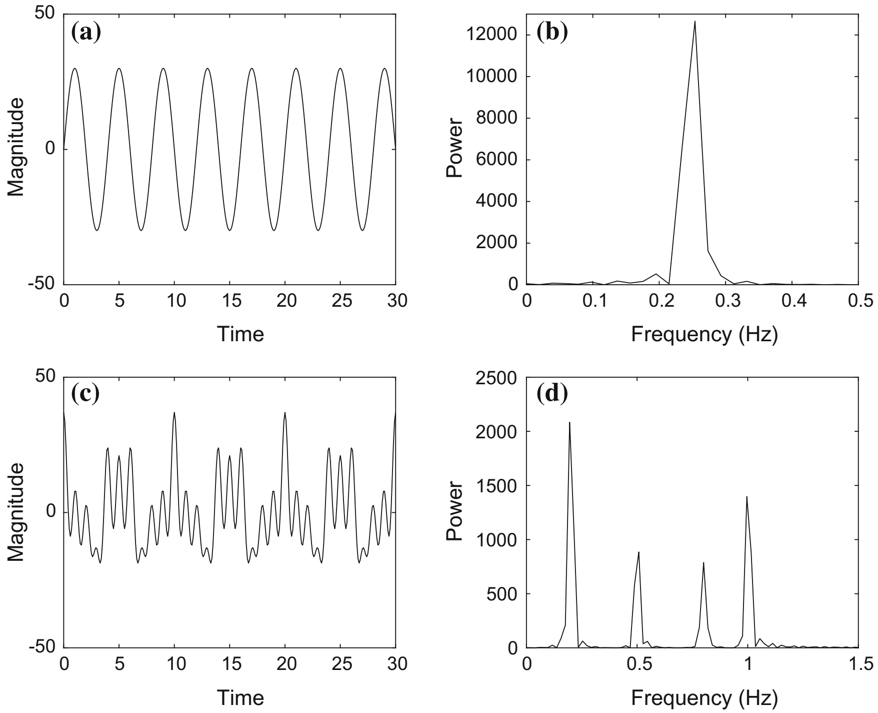


Fig. 9.3 Spectral analysis of periodic functions **a** sine wave **b** spectral density of sine function **c** sum of four cosine functions **d** spectral density of sum of four cosine functions

using the Fourier transformation. The Fourier series is given by:

$$
\begin{aligned}
X(t) &= a_0 + \sum_{i=1}^{\infty} a_i \cos(2\pi\nu_i t) + \sum_{i=1}^{\infty} b_i \sin(2\pi\nu_i t) \\
&= a_0(\cos(0t) + \sin(0t)) + \sum_{i=1}^{\infty} a_i \cos(2\pi\nu_i t) + \sum_{i=1}^{\infty} b_i \sin(2\pi\nu_i t) \\
&= \sum_{i=0}^{\infty} (a_i \cos(2\pi\nu_i t) + b_i \sin(2\pi\nu_i t)) \qquad (9.25)
\end{aligned}
$$

Substituting, $a_i = A_i \sin(\phi_i)$ and $b_i = A_i \cos(\phi_i)$, the Eq. 9.25 can be written as:

$$
X(t) = \sum_{i=0}^{\infty} (A_i \sin(2\pi\nu_i t + \phi_i)) \qquad (9.26)
$$

where A_i and ϕ_i are amplitude and phase for frequency ν_i. The sum of root mean square of sin(•) is $1/\sqrt{2}$, so the variance of the $A_i \sin(2\pi\nu_i t + \phi_i)$ is $A_i^2/2$. Hence, a frequency of ν_i contributes/estimates $A_i^2/2$ of total variance of $X(t)$. The plot of $A_i^2/2$ with respect to frequency ν_i is called power spectrum of time series. For instance, power spectrum of some of circular function is shown in Fig. 9.3. The power spectrum can be used to find significant frequencies or presence of periodicity in a time series. For instance, from power spectrum in Fig. 9.3b the wave has a frequency of 0.25 Hz and hence periodicity of 4 s. Similarly, the power spectrum in Fig. 9.3d clearly shows the frequencies in the wave shown in Fig. 9.3c. These frequencies are 0.2, 0.5, 0.8, and 1 Hz. Generally, in hydroclimatic variables mostly one or two frequencies (related to annual seasonality) will be prominent.

9.6 Data Transformation

Most of the parameter estimation methods are based on the assumption that the time series follows normal probability distribution. Hence, we may need to transform the time series to follow normal probability distribution for some applications.

If we have a time series, $X(t)$ following lognormal distribution, the following transformation yields normal series, $Y(t)$:

$$Y(t) = \ln(X(t)). \tag{9.27}$$

If the time series $X(t)$ follows gamma probability distribution, then the following transformation may yield a normally distributed random variable $Y(t)$.

$$Y(t) = \sqrt{X(t)} \tag{9.28}$$

Power transformation, also known as Box-Cox transformation, can also be used for transforming the data to normal distribution. One-parameter Box-Cox transformation is given by

$$Y(t) = \begin{cases} \frac{(X(t))^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \text{ and } X(t) > 0 \\ \ln(X(t)) & \text{if } \lambda = 0 \end{cases} \tag{9.29}$$

The other two transformation methods discussed before can be considered as special cases of one-parameter Box-Cox transformation.

The two-parameter Box-Cox transformation is given by

$$Y(t) = \begin{cases} \frac{(X(t)+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \text{if } \lambda_1 \neq 0 \text{ and } X(t) > -\lambda_2 \\ \ln(X(t)+\lambda_2) & \text{if } \lambda_1 = 0 \end{cases} \tag{9.30}$$

The parameter λ_1 or λ_2 can be obtained by method of maximum likelihood. It should be noted that any of the transformation procedure discussed in this section does not always result in a time series that is normally distributed. Hence, before further analysis, transformed variable/time series need to be checked if they follow normal distribution using the appropriate test (discussed in next section).

9.6.1 Test for Normal Distribution

For checking that the time series follow normal distribution, the time series can be plotted on normal probability paper. If the plot is close to straight line with slope 1 and intercept 0, then the series can be considered normally distributed. A number of statistical tests exist for checking normality in the data like chi-square (χ^2) test, Kolmogorov–Smirnov test, Anderson–Darling test, and skewness test. Former three tests are discussed in Sect. 6.4.4. The skewness test is explained here.

For skewness test, the skewness coefficient of a time series $X(t)$ is estimated as follows:

$$\hat{S} = \frac{\frac{1}{N}\sum_{i=1}^{N}(X(t) - \overline{X})^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(X(t) - \overline{X})^2\right]^{3/2}} \tag{9.31}$$

where N is the number of sample data and $\overline{X}$ is the sample mean for time series $X(t)$. The skewness test is based on the fact that the skewness coefficient of a normal variable is zero. If the series is normally distributed, $\hat{S}$ is asymptotically normally distributed with the mean of zero, variance of $6/N$, hence, $(1-\alpha) \times 100\%$ confidence limit on skewness is defined as,

$$S \in \left[-Z_{(\alpha/2)}\sqrt{6/N}, Z_{(\alpha/2)}\sqrt{6/N}\right] \tag{9.32}$$

where $Z_{(\alpha/2)}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. Therefore, if $\hat{S}$ falls within the limits of Eq. 9.32, the hypothesis of normality cannot be rejected. The test is found to be reasonably accurate for $N > 150$.

Example 9.6.1
At a location, the rainfall data is found to follow gamma distribution. For 20 consecutive days the recorded rainfall (in mm/day) are 2.89, 7.39, 23.88, 10.59, 5.91, 1.53, 3.48, 56.54, 26.19, 6.35, 38.09, 0.01, 3.03, 41.57, 44.73, 21.39, 15.87, 1.22, 21.75, and 0.21, respectively. Transform the data such that it follows normal distribution. Check whether the transformed data follow normal distribution using skewness test at 5% significance level (as discussed in Sect. 9.6.1).

Solution A gamma distributed random variable can be transformed into normal distribution using the Eq. 9.28. Further, for checking the normality of data using the skewness, the skewness is required to be calculated using the Eq. 9.31. These calculations are shown in Table 9.2.

Table 9.2 Calculation for data transformation and skewness test

S. No.	Rainfall (mm) $X(t)$	Normalized series $Y(t)$	Normalized series deviation $Y_d(t)$	$Y_d(t)^2$	$Y_d(t)^3$
1	2.89	1.70	−1.76	3.10	−5.47
2	7.39	2.72	−0.74	0.55	−0.41
3	23.88	4.89	1.43	2.03	2.89
4	10.59	3.25	−0.21	0.04	−0.01
5	5.91	2.43	−1.03	1.06	−1.09
6	1.53	1.24	−2.22	4.95	−11.01
7	3.48	1.87	−1.60	2.55	−4.07
8	56.54	7.52	4.06	16.46	66.81
9	26.19	5.12	1.66	2.74	4.54
10	6.35	2.52	−0.94	0.89	−0.84
11	38.09	6.17	2.71	7.34	19.9
12	0.01	0.10	−3.36	11.30	−37.99
13	3.03	1.74	−1.72	2.96	−5.10
14	41.57	6.45	2.99	8.92	26.62
15	44.73	6.69	3.23	10.41	33.59
16	21.39	4.62	1.16	1.35	1.57
17	15.87	3.98	0.52	0.27	0.14
18	1.22	1.10	−2.36	5.56	−13.10
19	21.75	4.66	1.20	1.44	1.74
20	0.21	0.46	−3.00	9.02	−27.09
Total	332.62	69.23	0.00	92.94	51.62

From the table, the skewness can be calculated as:

$$S = \frac{\frac{1}{N}\sum_{i=1}^{N}(Y(t)-\overline{Y})^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(X(t)-\overline{X})^2\right]^{3/2}} = \frac{[51.62/20]}{[92.94/20]^{3/2}} = 0.26$$

Test for Normal distributed data using skewness (Sect. 9.6.1)

Null Hypothesis: Data is normally distributed.
Alternative Hypothesis: Data is not normally distributed.
Level of Significance: $\alpha = 5\%$

Null Hypothesis is acceptable for $|S| < Z_{(\alpha/2)}\sqrt{6/N}$ (Eq. 9.32)
As, $Z_{(\alpha/2)}\sqrt{6/N} = Z_{0.025}\sqrt{6/20} = 1.074$
Since $-1.074 < S < 1.074$, so the null hypothesis of data being normally distributed cannot be rejected.

9.7 Time Series Modeling in Hydroclimatology

After removal of deterministic components (like trend, periodicity, or jump) of time series, different time series modeling approaches can be used for modeling stochastic component of the time series. Some of the popular linear models for time series prediction/forecast are following:

(i) Autoregressive model
(ii) Moving average model
(iii) Autoregressive moving average model
(iv) Autoregressive integrated moving average model

Out of these, the first three are linear stationary models used for modeling stationary time series. However, the last model is a linear non-stationary model and is used to model a time series for which dth difference series (Sect. 9.7.2) is stationary. Stationary and non-stationary models are discussed in Sect. 9.7.3. All of these models are linear regression model and try to relate the present value of time series with the previous values. Being linear, these models rely on mutual linear association between time series values. These linear associations are expressed in term of autocorrelation function and partial autocorrelation function in time series.

9.7.1 Measures of Linear Association in Time Series

Hydroclimatic time series often have linear association between its successive values. These linear association can be utilized in developing the structure of the linear models for analysis/prediction of the time series. Two linear association measures for time series are autocorrelation and partial autocorrelation functions.

Autocorrelation Function

Autocorrelation is a measure of linear association between the values of same time series separated by some time lag/steps (say k). For a time series $X(t)$, and the same time series with lag k (represented by $X(t-k)$), the linear association is measured by autocovariance. The term *auto* is used as the values are from same series but with some lags. The autocovariance function for lag k (represented by C_k) is given by:

$$C_k = E(X(t), X(t-k)) \tag{9.33}$$

where E represents the expectation. The autocorrelation function for lag k is defined as:

$$\rho_k = \frac{C_k}{\sqrt{E\left[(X(t)-\overline{X}(t))^2\right]E\left[(X(t-k)-\overline{X}(t-k))^2\right]}} = \frac{C_k}{\sigma_t\sigma_{t-k}} \tag{9.34}$$

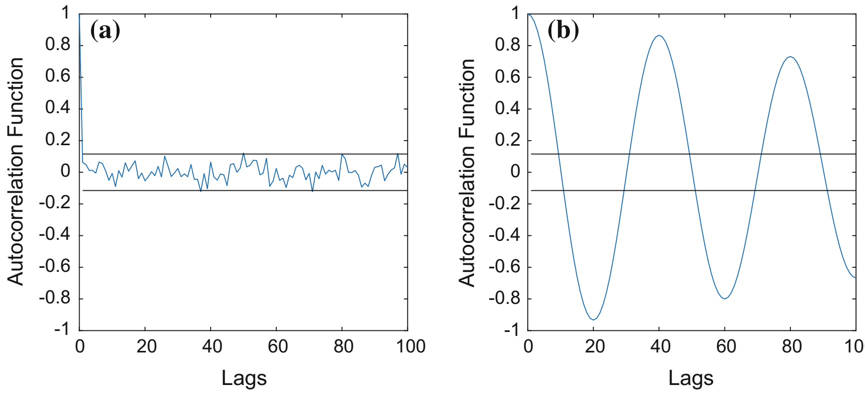


Fig. 9.4 Typical autocorrelogram for **a** random/stationary time series **b** periodic time series

where σ_t and σ_{t-k} are standard deviation for time series $X(t)$ and $X(t-k)$, respectively. If the time series is second-order or higher-order stationary (standard deviation does not change over time), then the Eq. 9.34 can be expressed as:

$$\rho_k = \frac{C_k}{\sigma^2} \tag{9.35}$$

where σ is the standard deviation of time series $X(t)$. A plot of autocorrelation function with corresponding lag is called **autocorrelogram**. For a stationary time series the autocorrelation become insignificant with increasing lag (Fig. 9.4a). However, for a periodic time series the autocorrelation is also periodic and decreases slowly with damping peaks (Fig. 9.4b). Under the assumption of independent time series, autocorrelation at lag k is normally distributed with zero mean and $1\big/(N-k)$ variance, N being the length of time series. The confidence limits of autocorrelation function for α significance level are given as follows,

$$\frac{-Z_{(\alpha/2)}}{\sqrt{N-k}} \leq \rho_k \leq \frac{Z_{(\alpha/2)}}{\sqrt{N-k}} \tag{9.36}$$

where $Z_{(\alpha/2)}$ is standard normal variate at $(1-\alpha/2)\times 100\%$ non-exceedance probability, i.e., $P(Z > Z_{(\alpha/2)}) = \alpha\big/2$. For large value of N $(N \gg k)$ the Eq. 9.36 further reduces to,

$$\frac{-Z_{(\alpha/2)}}{\sqrt{N}} \leq \rho_k \leq \frac{Z_{(\alpha/2)}}{\sqrt{N}} \tag{9.37}$$

Partial Autocorrelation Function (PACF)

Partial correlation is the measure of linear association between two random variables when effect of other random variables is removed. For instance, let X, Y, and Z be three random variables. The partial correlation between X and Y, when the effect of Z is removed, represented as $\rho_{XY/Z}$, is the correlation between the residuals R_y and

R_x resulting from linear regression of Y and X with Z, respectively. Hence, $\rho_{XY/Z}$ is expressed as:

$$\rho_{XY/Z} = \frac{E(R_x R_y)}{\sqrt{Var(R_x)Var(R_y)}} \tag{9.38}$$

With an assumption that all involved variables are multivariate Gaussian distributed, if X is conditionally independent of Y given Z, then $\rho_{XY/Z}$ is zero. Hence, partial correlation is useful in linear models like multiple regression to figure out variables that do not contribute significantly to the prediction.

Partial autocorrelation function (PACF) of a time series X at lag k is defined as:

$$\varphi_k = \rho_{X_0 X_k/\{X_1, X_2, \ldots, X_{k-1}\}} \tag{9.39}$$

The partial autocorrelation at lag 0 (φ_0) is 1. The partial autocorrelation function at higher lag (say k) (φ_k) is calculated using Yule–Walker equation, which is represented as:

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \ldots & \rho_{k-1} \\ \rho_1 & 1 & \rho_2 & \cdots & \rho_{k-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \Phi_1 \\ \Phi_2 \\ \Phi_3 \\ \vdots \\ \Phi_k \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \rho_k \end{bmatrix} \tag{9.40}$$

where ρ_i is autocorrelation function at lag i and Φ_i is ith parameter for autoregressive model (discussed later in Sect. 9.7.4). For the Yule–Walker equation, the last autoregressive parameter (Φ_k) corresponds to φ_k. Thus, for the Eq. 9.40, $\varphi_k = \Phi_k$. *However, it must be noted that* $\varphi_i \neq \Phi_i$, *for* $i \in \{1, 2, \ldots, k-1\}$.

The solution of Yule–Walker equation for calculating partial autocorrelation function at lag k is expressed as:

$$\varphi_k = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 \end{vmatrix}} \tag{9.41}$$

Hence, partial autocorrelation function at lags 1 and 2 is defined as:

$$\varphi_1 = \rho_1 \tag{9.42}$$

$$\varphi_2 = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \tag{9.43}$$

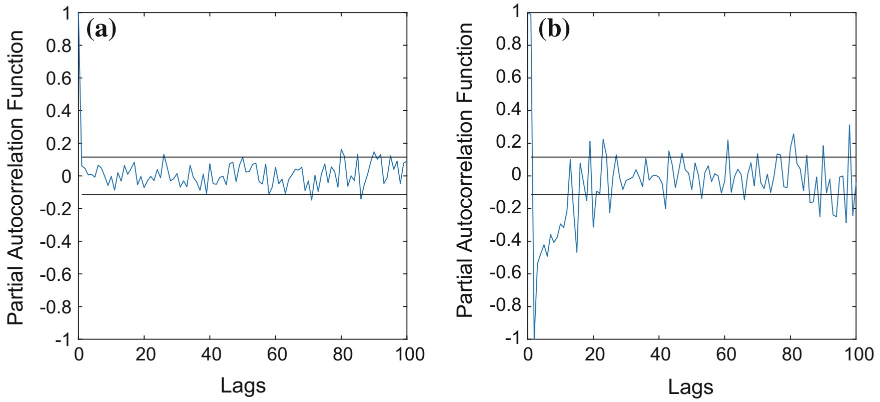


Fig. 9.5 Typical partial autocorrelogram with confidence interval for a **a** random/stationary time series, and a **b** periodic time series

The partial autocorrelation at lag k is expected to follow the normal distribution with mean 0 and standard deviation $1/\sqrt{N-k}$ (like autocorrelation). Thus, $(1-\alpha)\times 100\%$ confidence interval is given by:

$$\frac{-Z_{(\alpha/2)}}{\sqrt{N-k}} \leq \varphi_k \leq \frac{Z_{(\alpha/2)}}{\sqrt{N-k}} \tag{9.44}$$

where $Z_{(\alpha/2)}$ is standard normal variate at $(1-\alpha/2)\times 100\%$ non-exceedance probability, i.e., $P(Z > Z_{(\alpha/2)}) = \alpha\big/2$. The null hypothesis of partial autocorrelation at a lag k being equal to zero can be tested using above equation. For large value of N $(N \gg k)$ the Eq. 9.44 further reduces to,

$$\frac{-Z_{(\alpha/2)}}{\sqrt{N}} \leq \varphi_k \leq \frac{Z_{(\alpha/2)}}{\sqrt{N}} \tag{9.45}$$

Typical examples of partial autocorrelogram with confidence interval for a random/stationary time series and a periodic time series are shown in Fig. 9.5a and Fig. 9.5b respectively.

Example 9.7.1
For the rainfall time series given in Example 9.6.1, calculate the autocorrelation at lags 0, 1, and 2. Calculate the 95% confidence limits for autocorrelation at lags 1 and 2.

Solution Autocorrelation function at lag 0 is 1. Hence,

$$\rho_0 = 1$$

For calculating the autocorrelation at lags 1 and 2, the covariance of the rainfall time series (denoted as X_t) with its 1- and 2-day lagged series (denoted by X_{t-1} and X_{t-2}, respectively) is calculated. Let the covariance of rainfall series with its kth lagged series is represented as cov_k. Hence

$$\text{cov}_1 = \text{cov}\left(X_t, X_{t-1}\right) = \text{cov}\left(\begin{bmatrix} 2.89 & 7.39 \\ 7.39 & 23.88 \\ 23.88 & 10.59 \\ 10.59 & 5.91 \\ 5.91 & 1.53 \\ 1.53 & 3.48 \\ 3.48 & 56.54 \\ 56.54 & 26.19 \\ 26.19 & 6.35 \\ 6.35 & 38.09 \\ 38.09 & 0.01 \\ 0.01 & 3.03 \\ 3.03 & 41.57 \\ 41.57 & 44.73 \\ 44.73 & 21.39 \\ 21.39 & 15.87 \\ 15.87 & 1.22 \\ 1.22 & 21.75 \\ 21.75 & 0.21 \end{bmatrix}\right) = \begin{bmatrix} 293.05 & 9.84 \\ 9.84 & 297.81 \end{bmatrix}$$

And the corresponding autocorrelation matrix at lag 1

$$= \begin{bmatrix} 1 & 0.033 \\ 0.033 & 1 \end{bmatrix}$$

The off-diagonal member of autocorrelation matrix at lag 1 is autocorrelation function at lag 1 (ρ_1), hence

$$\rho_1 = 0.033$$

Similarly, the autocorrelation function at lag 2 (ρ_2) is found to be (-0.282).

The autocorrelation function at lag 1 (ρ_1) is supposed to follow normal distribution with mean 0 and standard deviation $1/\sqrt{N-1} = 0.229$. Hence, the 95% confidence interval for ρ_1 is given by:

$$[-0.229Z_{0.025}, 0.229Z_{0.025}] = [-0.45, 0.45]$$

Similarly, the autocorrelation function at lag 2 (ρ_2) follows normal distribution with mean 0 and standard deviation $1/\sqrt{N-2} = 0.236$. Hence, the 95% confidence interval for ρ_2 is given by:

$$[-0.236Z_{0.025}, 0.236Z_{0.025}] = [-0.46, 0.46]$$

Example 9.7.2

The autocorrelation coefficients for a monthly streamflow time series at a gauging station at lags 0, 1, and 2 are 1.0, 0.79, and 0.52 respectively. Estimate the partial autocorrelation at these lags. Also, check whether the partial autocorrelation at lag 2 is significant or not at 5% significance level. Assume the data length to be 60.

Solution Partial autocorrelation function at lag 0 is 1. Hence,

$$\varphi_0 = 1$$

According to Yule–Walker equation (9.40), the partial autocorrelation at lag 1 (φ_1) is given by

$$\varphi_1 = \rho_1 = 0.79$$

Similarly partial autocorrelation at lag 2 (φ_2) can be calculated using Yule–Walker equation as

$$\begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \Phi_1 \\ \Phi_2 \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}$$

$$\text{or, } \Phi_1 + 0.79\Phi_2 = 0.79 \text{ and } 0.79\Phi_1 + \Phi_2 = 0.52$$

$$\text{or } \Phi_1 = 1.00 \text{ and } \Phi_2 = -0.28$$

Hence, $\varphi_2 = \Phi_2 = (-0.28)$

Test for Significance of φ_2

Null Hypothesis: Partial autocorrelation is not significant, i.e., $\varphi_2 = 0$
Alternative Hypothesis: $\varphi_2 \neq 0$
Level of Significance: $\alpha = 5\%$

As partial autocorrelation at lag 2 is expected to follow normal distribution with mean 0 and standard deviation $1\big/\sqrt{N-k} = 0.131$ (as $N = 60$ and $k = 2$). Hence, at 5% significance level the critical value of autocorrelation at lag 2 is given by,

$$\pm\frac{Z_{(\alpha/2)}}{\sqrt{N-2}} = \pm 1.96 \times 0.131 = \pm 0.26$$

Since partial autocorrelation at lag 2 (i.e., -0.28) falls in critical zone $((-\infty, -0.26] \cup [0.26, \infty))$, the null hypothesis is rejected. Hence, the partial autocorrelation at lag 2 is significant.

9.7.2 Statistical Operators on Time Series

Backward Shift Operator

Backward shift operator or Backshift operator (represented as $B(\bullet)$) returns the immediate previous value of time series. For a time series $X(t)$, backshift operation is represented by:

$$\begin{aligned} BX(t) &= X(t-1) \\ B^2X(t) &= X(t-2) \\ B^nX(t) &= X(t-n) \end{aligned} \tag{9.46}$$

Forward Shift Operator

Forward shift operator (represented as $F(\bullet)$) returns the immediate next value of time series. It works opposite of backshift operator, and thus, also represented as B^{-1}. For a time series $X(t)$ it is represented by:

$$\begin{aligned} FX(t) &= B^{-1}X(t) = X(t+1) \\ F^2X(t) &= B^{-2}X(t) = X(t+2) \\ F^nX(t) &= B^{-n}X(t) = X(t+n) \end{aligned} \tag{9.47}$$

Difference Operator

Difference operator returns the difference of the current and previous time step value in a time series. It is expressed as:

$$\begin{aligned} \nabla(X(t)) = (1-B)X(t) \quad &= X(t) - X(t-1) \\ \nabla^2(X(t)) = (1-B)^2X(t) \quad &= (1-2B+B^2)X(t) \\ &= X(t) - 2X(t-1) + X(t-2) \\ \nabla^nX(t) = (1-B)^nX(t) \quad & \end{aligned} \tag{9.48}$$

Moving Average—Low Pass Filtering

Moving average (also known as rolling or running average) tries to reduce the short-term fluctuations in time series by taking the average of the neighboring (say n) values of the time series. Moving average works as a low pass filter and reduces the high-frequency oscillation in the time series. A n term or n window Moving average is expressed as:

$$\begin{aligned} Y(n) &= \frac{1}{n}(X(1) + X(2) + \cdots + X(n)) \\ Y(n+1) &= \frac{1}{n}(X(2) + X(3) + \cdots + X(n+1)) \\ Y(t+n-1) &= \frac{1}{n}(X(t) + X(t+1) + \cdots + X(n+t-1)) \end{aligned} \tag{9.49}$$

In the above equations, the moving average is assigned at the end of the window over which average is captured. Sometimes, the moving average values are assigned to the central value of the window selected. As the window of moving average increases, the time series $Y(t)$ gets smoother, i.e., low pass filtering effect of moving average increases with increase in terms being used for averaging (n).

Differencing—High Pass Filtering

Differencing is a high pass filtering method that removes low-frequency oscillation from the time series. The nth-order differencing is expressed as:

$$\begin{aligned} Y_1(t) &= X(t) - X(t-1) && \text{for } t = 2, 3, \ldots \\ Y_2(t) &= Y_1(t) - Y_1(t-1) && \text{for } t = 3, 4, \ldots \\ Y_n(t) &= Y_{n-1}(t) - Y_{n-1}(t-1) && \text{for } t = n+1, n+2, \ldots \end{aligned} \tag{9.50}$$

Differencing can be used transforming the time series into normal distribution and hence, differencing is also considered as 'whitening filter'.

Example 9.7.3
For the rainfall time series given in Example 9.6.1, calculate a moving average with window 2 and first order differencing. Check their respective behavior of being a low and high pass filter by visualizing the results.

Solution Let us represent rainfall series as X. The moving average series with window 2 (Y) is expressed as

$$Y(t) = \frac{1}{2}(X(t) + X(t+1)) \qquad \text{for } t = 1, 2, \ldots, (n-1)$$

Similarly, the 1st-order differencing series (Z) is given by,

$$Z(t) = X(t) - X(t-1) \qquad \text{for } t = 2, 3, \ldots, n$$

The moving average and differencing series are assigned at the end of window. The calculation is shown in Table 9.3. From Fig. 9.6, the moving average series is smoother than rainfall (the peaks have reduced), thus moving average acts as a low pass filter. However, the differencing operator shows higher values corresponding to peak and hence acts as high pass filter.

Table 9.3 Rainfall series and its moving average and differencing series

Rainfall	Moving average window 2	1st-order differencing
2.89		
7.39	5.14	4.5
23.88	15.63	16.49
10.59	17.23	−13.29
5.91	8.25	−4.68
1.53	3.72	−4.38
3.48	2.505	1.95
56.54	30.01	53.06
26.19	41.36	−30.35
6.35	16.27	−19.84
38.09	22.22	31.74
0.01	19.05	−38.08
3.03	1.52	3.02
41.57	22.3	38.54
44.73	43.15	3.16
21.39	33.06	−23.34
15.87	18.63	−5.52
1.22	8.545	−14.65
21.75	11.48	20.53
0.21	10.98	−21.54

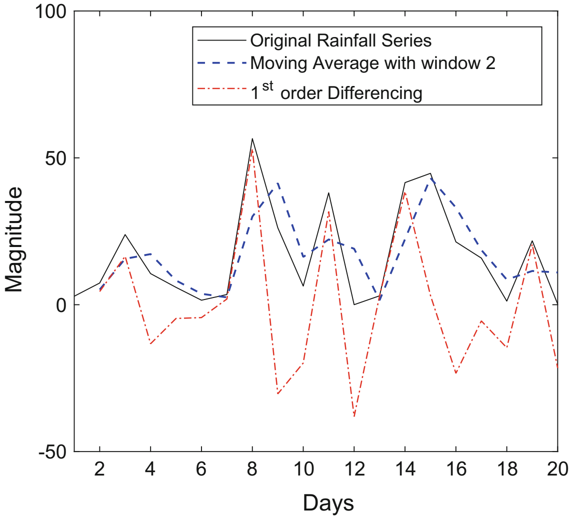


Fig. 9.6 Rainfall along with its moving average with window 2 and 1st-order differencing

9.7.3 Properties of Time Series Models

Stationary and Non-stationary Time Series Models

A stationary model assumes that the process remains in equilibrium in terms of its statistical properties over time. Hence, stationary time series models have finite variance. For using the stationary model on any time series, the time series is required to be stationary (i.e., statistical properties remain same over time). On the other hand, non-stationary model do not assume that the process is in equilibrium with respect to its statistical properties over time. Suppose that there is a mathematical model that takes white noise (normally distributed uncorrelated series with zero mean, $\varepsilon(t)$) as input and model the time series $X(t)$. This type of model is called linear filter and represented as:

$$X(t) = \mu + \varepsilon(t) - \theta_1 \varepsilon(t-1) - \theta_2 \varepsilon(t-2) - \cdots = \mu + \theta(B)\varepsilon(t) \qquad (9.51)$$

where μ is the mean of $X(t)$, θ_i is ith parameter of model, and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots = \theta_0 - \theta_1 B - \theta_2 B^2 - \cdots$ with ($\theta_0 = 1$) is a function of backshift operator also called transfer function of the linear filter. If the absolute summation of sequence of parameters is finite $\left(\sum_{i=0}^{\infty} |\theta_i| < \infty \right)$, then the model is stationary and model is in equilibrium around the mean μ. It should be noted that the condition $\sum_{i=0}^{\infty} |\theta_i| < \infty$ also employs that all roots of $\theta(B) = 0$ fall outside the unit circle, i.e., $|B| > 1$.

Invertibility

Invertibility is another property of the time series model. Non-stationary models can also be invertible or vice versa. A time series model is called invertible if error can be expressed as function of backshift operator over the time series with finite variance. Hence, a model that can be expressed in the form of,

$$\begin{aligned} (1 - \Phi_1 B - \Phi_2 B^2 + \ldots)X(t) &= \varepsilon(t) \\ \Phi(B)X(t) &= \varepsilon(t) \end{aligned} \qquad (9.52)$$

is invertible, if the absolute sum of its parameters converges $\left(\sum_{j=0}^{\infty} |\Phi_j| < \infty \right)$. Invertibility is also ensured if all roots of $\Phi(B) = 0$ falls outside the unit circle, i.e., $|B| > 1$.

9.7.4 Auto-Regressive (AR) Model

Autoregressive model tries to estimate the current value of time series using linear combination of weighted sum of previous values of the same time series. AR models are extensively used in hydroclimatic time series as current values of the time series are expected to be affected by the previous values. This characteristic of hydroclimatic variables is also referred as memory component. The number of lagged values being considered (say p) is called order of AR model. pth-order AR model (AR(p))is given by

$$X(t) = \sum_{i=1}^{p} \Phi_i X(t-i) + \varepsilon(t) \tag{9.53}$$

where Φ_i(for $i \in \{1, 2, \ldots, p\}$) are called autoregressive coefficients and $\varepsilon(t)$ is uncorrelated identically distributed error with mean zero, also known as white noise. Time series $X(t)$ is obtained after removing the deterministic components like trend and periodicity. Using the backshift operator ARMA(p) can also be written as,

$$\begin{aligned} &X(t) - \Phi_1 B(X(t)) - \Phi_2 B^2(X(t)) - \cdots - \Phi_p B^p(X(t)) = \varepsilon(t) \\ &\text{or, } \Phi(B)X(t) = \varepsilon(t) \end{aligned} \tag{9.54}$$

where $\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^p$ for AR(p) model.

As an initial guess, the order p is decided from partial autocorrelation function. Number of lags for which partial autocorrelation is significant is considered as p. Hence, for a AR(p) model, all partial autocorrelation with lag more than p should be zero and autocorrelation decays exponentially to zero. Different AR models are fitted using the slight variation in initial guess of AR order, the best model out of all fitted models is chosen on the basis of their parsimony (Sect. 9.7.10).

Following assumptions are made while developing an AR model.

$$E(\varepsilon(t)) = 0 \tag{9.55}$$

$$E(\varepsilon(t)\varepsilon(t-k)) = E(\varepsilon(t)X(t-k)) = 0 \qquad \text{for } k = 1, 2, \ldots, p \tag{9.56}$$

For an AR model, the coefficient of determination is given by

$$R^2 = \sum_{i=1}^{p} \Phi_i \rho_i = 1 - \frac{Var(\varepsilon)}{Var(X)} \tag{9.57}$$

The parameters of a pth-order AR model are obtained by Yule–Walker equations. Yule–Walker equations are derived by taking expectation of p different equations obtained by multiplying lagged values of time series, i.e., $X(t-1)$, $X(t-2)$, $\ldots$, $X(t-p)$ with the general form of AR model given in Eq. 9.53. The Yule–Walker equations are given by

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_2 & \cdots & \rho_{p-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \Phi_1 \\ \Phi_2 \\ \Phi_3 \\ \vdots \\ \Phi_p \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \rho_p \end{bmatrix} \tag{9.58}$$

where ρ_i is autocorrelation coefficient at lag i. It should be noted that $\rho_0 = 1$, hence, the above Yule–Walker equation can also be written as:

$$\begin{bmatrix} \rho_0 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & \rho_0 & \rho_2 & \cdots & \rho_{p-2} \\ \rho_2 & \rho_1 & \rho_0 & \cdots & \rho_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & \rho_0 \end{bmatrix} \begin{bmatrix} \Phi_1 \\ \Phi_2 \\ \Phi_3 \\ \vdots \\ \Phi_p \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \rho_p \end{bmatrix}$$

$$\text{or, } \sum_{i=1}^{p} \rho_{k-i}\Phi_i = \rho_k \tag{9.59}$$

It should be noted that Φ_p is the partial autocorrelation at lag p (φ_p).

Properties of AR Model

Stationarity: The developed AR model is required to be a stationary model. For a stationary AR(p) model, the autocorrelation matrix for order p should be positive-definite, i.e., determinant of all minors of the correlation matrix is positive. Hence, for an AR(2) model

$$\begin{bmatrix} \rho_0 & \rho_1 & \rho_2 \\ \rho_1 & \rho_0 & \rho_2 \\ \rho_2 & \rho_1 & \rho_0 \end{bmatrix} \text{ should be positive-definite.}$$

$$\rho_0 > 0; \begin{vmatrix} \rho_0 & \rho_1 \\ \rho_1 & \rho_0 \end{vmatrix} > 0; \begin{vmatrix} \rho_0 & \rho_1 & \rho_2 \\ \rho_1 & \rho_0 & \rho_2 \\ \rho_2 & \rho_1 & \rho_0 \end{vmatrix} > 0$$

$$\rho_0 > 0; -1 < \rho_1 < 1; -1 < \rho_2 < 1; \text{ and } -1 < \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} < 1 \tag{9.60}$$

Using the above relationship, for an AR(1), to ensure stationarity $\Phi_1 = \rho_1$ and hence, $|\Phi_1| < 1$. AR(1) model can be expressed using backshift operator as:

$$
\begin{aligned}
X(t) &= \Phi_1 X(t-1) + \varepsilon(t)\\
X(t) &= \Phi_1 B(X(t)) + \varepsilon(t)\\
(1-\Phi_1 B)X(t) &= \varepsilon(t)\\
\Phi(B)X(t) &= \varepsilon(t)
\end{aligned}
\tag{9.61}
$$

The root of $\Phi(B) = 0$ is $B = 1/\Phi_1$. For staionarity, the root of $\Phi(B) = 0$ should fall outside the unit circle (Sect. 9.7.3). This result is valid for AR model of higher orders also. The equation $\Phi(B) = 0$ is called **characteristic equation** of AR model.

Invertibility: A model is called invertible if error can be expressed as function of backshift operator over the time series with finite variance (Sect. 9.7.3). In an AR model, the error can be related to time series by using a function of backshift operator (Eq. 9.61 with $\sum\limits_{i=0}^{p} |\Phi_i| < \infty$), hence, all stationary AR models are invertible.

Example 9.7.4

Derive the nature of autocorrelation function for AR(1) and AR(2) models. Also find the error variance and stationarity condition.

Solution First-order AR model

For time series $X(t)$, the first-order autoregressive model AR(1) is given by,

$$X(t) = \Phi_1 X(t-1) + \varepsilon(t)$$

From the Yule–Walker equation we can write,

$$
\begin{bmatrix}
1 & \rho_1 & \rho_2 & \dots & \rho_{n-1}\\
\rho_1 & 1 & \rho_2 & \cdots & \rho_{n-2}\\
\rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3}\\
\vdots & \vdots & \vdots & \ddots & \vdots\\
\rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1
\end{bmatrix}
\begin{bmatrix}
\Phi_1\\ \Phi_2\\ \Phi_3\\ \vdots\\ \Phi_n
\end{bmatrix}
=
\begin{bmatrix}
\rho_1\\ \rho_2\\ \rho_3\\ \vdots\\ \rho_n
\end{bmatrix}
$$

For AR(1) model $\Phi_2 = \Phi_3 = \cdots = \Phi_n = 0$. Hence, the autocorrelation function is given by:

$$
\begin{aligned}
\rho_1 &= \Phi_1,\\
\rho_2 &= \Phi_1\rho_1 = \rho_1^2\\
&\vdots\\
\rho_n &= \Phi_1\rho_{p-1} = \rho_1^p
\end{aligned}
$$

Hence, the autocorrelation function of AR(1) model decays exponentially to zero for positive values of ρ_1. For negative values of the autoregressive coefficient, the autocorrelation function is damped and oscillates around zero.

The variance of the error series (σ_ϵ) is given by,

$$\begin{aligned}
&\varepsilon(t) = X(t) - \Phi_1 X(t-1)\\
&E\left(\varepsilon(t)^2\right) = E\left[(X(t) - \rho_1 X(t-1))^2\right]\\
&\sigma_\epsilon^2 = \sigma_X^2\left(1 - \rho_1^2\right)
\end{aligned}$$

For stationary condition the roots of equation $\Phi(B)$ should fall outside the unit circle.

$$\begin{aligned}
1 - \Phi_1 B &= 0\\
B &= \frac{1}{\Phi_1}
\end{aligned}$$

For $|B| > 1$, the $\Phi_1 < 1$, hence, for AR(1) to be stationary the autoregressive parameter should be less than 1.

Second-order AR model

The second-order autoregressive model AR(2) has the form,

$$X(t) = \Phi_1 X(t-1) + \Phi_2 X(t-2) + \varepsilon(t)$$

From the Yule–Walker equations,

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1}\\ \rho_1 & 1 & \rho_2 & \cdots & \rho_{n-2}\\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \Phi_1\\ \Phi_2\\ \Phi_3\\ \vdots\\ \Phi_n \end{bmatrix} = \begin{bmatrix} \rho_1\\ \rho_2\\ \rho_3\\ \vdots\\ \rho_n \end{bmatrix}$$

For AR(2) model $\Phi_3 = \Phi_4 = \cdots = \Phi_n = 0$. Hence, the autocorrelation function is given by:

$$\begin{aligned}
\rho_1 &= \Phi_1 + \Phi_2\rho_1\\
\rho_2 &= \Phi_1\rho_1 + \Phi_2\\
\rho_3 &= \Phi_1\rho_2 + \Phi_2\rho_1\\
&\ \vdots\\
\rho_n &= \Phi_1\rho_{n-1} + \Phi_2\rho_{n-2}
\end{aligned}$$

By solving the first two equations simultaneously, we get the following results,

$$\Phi_1 = \frac{\rho_1(1-\rho_2)}{\left(1-\rho_1^2\right)}$$
$$\Phi_2 = \frac{\left(\rho_2-\rho_1^2\right)}{\left(1-\rho_1^2\right)}$$

For $k > 2$ the nature of autocorrelation function depends upon the values of Φ_1 and Φ_2. For instance, if $\Phi_1^2 + 4\Phi_2 \geq 0$ and $\Phi_1 > 0$ then the autocorrelation function decays exponentially to zero. However, if $\Phi_1^2 + 4\Phi_2 \geq 0$ and $\Phi_1 < 0$ then the autocorrelation function oscillates around zero. On the other hand, if $\Phi_1^2 + 4\Phi_2 < 0$ then the autocorrelation function is damped.

The variance of the error series $\varepsilon(t)$ can be calculated by taking expectation of its square as done in the case of AR(1) model.

$$\sigma_\varepsilon^2 = \sigma_X^2\left(1 - \rho_1\Phi_1 - \rho_2\Phi_2\right)$$

For stationarity, the roots of equation $\Phi(B)$ should fall outside unit circle.

$$1 - \Phi_1 B - \Phi_2 B^2 = 0$$
$$B = \frac{\Phi_1 \pm \sqrt{\Phi_1^2 + 4\Phi_2}}{2\Phi_2}$$

Hence, for stationary AR(2) model $|B| > 1$ or $\left|\frac{\Phi_1 \pm \sqrt{\Phi_1^2+4\Phi_2}}{2\Phi_2}\right| > 1$. Alternatively, for an AR(2) model to be stationary, the autocorrelation matrix should be positive-definite. Hence, the stationary criteria is also given as (Eq. 9.60).

$$-1 < \rho_1 < 1; -1 < \rho_2 < 1;\ \text{and}\ -1 < \frac{\rho_2 - \rho_1^2}{1-\rho_1^2} < 1$$

Example 9.7.5
Check the stationarity for an AR(2) model, the parameters have been estimated as $\Phi_1 = 0.1$ and $\Phi_2 = 0.2$.

Solution In order to satisfy the stationarity condition, the roots of characteristics equation should lie beyond unit circle,

$$\Phi(B) = 0$$
$$\text{or, } 1 - \Phi_1 B - \Phi_2 B^2 = 0$$
$$\text{or, } 1 - 0.1B - 0.2B^2 = 0$$
$$\text{or, } B = 2 \text{ or } (-2.5)$$

The roots of the equation lie outside the unit circle, hence, AR(2) model with model parameters $\Phi_1 = 0.1$ and $\Phi_2 = 0.2$ is stationary.

Example 9.7.6
Find out the model parameters for an AR(1) and AR(2) model if the estimated values of the autocorrelation at first and second lags are 0.6 and 0.2, respectively. Compare the AR models using the coefficient of determination. Also find the variance of residual series by considering the variance of the time series to be 20.

Solution **Case 1**: AR(1) Model
Using the Yule–Walker Equation we can write,

$$\Phi_1 = \rho_1 = 0.6$$

The coefficient of determination is given by (Eq. 9.57)

$$R^2 = \rho_1 \Phi_1 = 0.36$$

The variance of the residual series as,

$$\sigma_\varepsilon^2 = \sigma_X^2 \left(1 - \rho_1 \Phi_1\right) = 20 \left(1 - (0.6)^2\right) = 12.8$$

Case 2: AR(2) Model
Using the Yule–Walker equation (Eq. 9.58),

$$\begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \Phi_1 \\ \Phi_2 \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}$$
$$\text{or,} \quad \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix} \begin{bmatrix} \Phi_1 \\ \Phi_2 \end{bmatrix} = \begin{bmatrix} 0.6 \\ 0.2 \end{bmatrix}$$

Therefore, $\Phi_1 = 0.75$ and $\Phi_2 = -0.25$.

The coefficient of determination is given by

$$R^2 = \rho_1 \Phi_1 + \rho_2 \Phi_2 = 0.31$$

The variance of the residual series is given as,

$$\sigma_\varepsilon^2 = \sigma_X^2 \left(1 - \rho_1 \Phi_1 - \rho_2 \Phi_2\right) = 20(1 - 0.31) = 13.8$$

Hence, AR(2) model is marginally better than AR(1) model.

Example 9.7.7
For the sample inflow data given in the following table, determine the parameters of an AR(2) model (assume that data is free from trend, periodicity, or jumps). Also, find the variance of residual series.

Year	Flow ($\times$100 m^3/s)	Year	Flow ($\times$100 m^3/s)	Year	Flow ($\times$100 m^3/s)
1	560	11	850	21	250
2	630	12	870	22	360
3	590	13	340	23	1200
4	660	14	560	24	950
5	580	15	190	25	880
6	490	16	250	26	560
7	300	17	380	27	450
8	350	18	670	28	320
9	470	19	990	29	170
10	900	20	840	30	580

Solution Following the methodology discussed in Example 9.7.1, the autocorrelation function at lags 1 and 2 is

$$\rho_1 = 0.408 \text{ and } \rho_2 = -0.108$$

The parameters of AR(2) model can be calculated by using Eq. 9.58.

$$\begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \Phi_1 \\ \Phi_2 \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}$$

$$\text{or, } \begin{bmatrix} 1 & 0.408 \\ 0.408 & 1 \end{bmatrix} \begin{bmatrix} \Phi_1 \\ \Phi_2 \end{bmatrix} = \begin{bmatrix} 0.408 \\ -0.108 \end{bmatrix}$$

Solving the equation simultaneously, $\Phi_1 = 0.54$ and $\Phi_2 = -0.33$.

The variance of the inflow time series is calculated as,

$$\sigma_X^2 = \frac{\sum_{t=1}^{30} (X(t) - \overline{X})^2}{30 - 1} = 70456 \times (100 \text{ m}^3/\text{s})^2$$

The variance of the residual series is given by,

$$\begin{aligned} \sigma_\varepsilon^2 &= \sigma_X^2 (1 - \rho_1\Phi_1 - \rho_2\Phi_2) \\ &= 70456 (1 - 0.408 \times 0.542 - (-0.108) \times (-0.329)) \\ &= 52372 \times (100 \text{ m}^3/\text{s})^2 \end{aligned}$$

9.7.5 *Moving Average (MA) Model*

In MA model, the current time series values are modeled using linear association with the lagged residual values. The MA model of order q considers q lagged residual for developing the model. In general, the qth-order moving average model is expressed as:

$$X(t) = \varepsilon(t) - \sum_{i=1}^{q} \theta_i \varepsilon(t-i) \tag{9.62}$$

where θ_i and $\varepsilon(t-i)$ are the MA parameter and residual at lag i respectively. Time series $X(t)$ is obtained after removing the deterministic components like trend and periodicity. The above expression for MA(q) model can be expressed in terms of function of backshift operator as,

$$\begin{aligned} X(t) &= \varepsilon(t) - \theta_1 B(\varepsilon(t)) - \theta_2 B^2(\varepsilon(t)) - \cdots - \theta_q B^q(\varepsilon(t)) \\ \text{or, } X(t) &= \theta(B)X(t) \end{aligned} \tag{9.63}$$

where $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ for MA(q) model.

The assumptions for AR model (Eqs. 9.55 and 9.56) also hold for MA model. Under these assumptions, the relationship between the variance of residual and parameters of the MA model can be obtained by calculating expectation after squaring Eq. 9.62. The relationship is expressed as,

$$\sigma_\varepsilon^2 = \frac{\sigma_X^2}{(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)} \tag{9.64}$$

where σ_ε^2 and σ_X^2 are variance of residual series and time series, respectively. Hence, the coefficient of determination (R^2) for a MA(q) model is expressed as,

$$R^2 = 1 - \frac{\sigma_\varepsilon^2}{\sigma_X^2} = 1 - \frac{1}{(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)} \tag{9.65}$$

Using the MA model given in Eq. 9.62, time series value at k lag can be written as:

$$X(t-k) = \varepsilon(t-k) - \sum_{i=1}^{q} \theta_i \varepsilon(t-k-i) \tag{9.66}$$

The expectation of product of Eqs. 9.66 and 9.62 and using the assumptions (Eqs. 9.55 and 9.56) gives:

$$\rho_k = \begin{cases} \dfrac{-\theta_k + \sum_{i=1}^{q-k} \theta_i \theta_{k+i}}{1 + \sum_{i=1}^{q} \theta_i^2} & \text{for } k = 1, 2, \ldots, q \\ 0 & \text{for } k > q \end{cases} \tag{9.67}$$

Parameters of an MA model can be estimated by solution of the above equation. The order of the MA model is estimated on the basis of autocorrelation function. The number of lag for which the autocorrelation function is significant is taken as order of

MA model as an initial guess. With a little modification in the initial guess, number of different MA models with different model orders are fitted and then checked for parsimony of model (Sect. 9.7.10). The model that is most suitable on the basis of parsimony is selected.

Properties of MA Model

Stationarity: The variance of MA model is given by Eq. 9.64, which is finite for finite number of parameters. Hence, a finite MA process is always a stationary model.

Invertibility: The roots of **characteristic equation** of MA ($\theta(B) = 0$) should lie outside the unit circle, i.e., $|B| > 1$. Sometimes, the parameter estimation for MA model may result in more than one solution for a single parameter (due to nonlinear nature of equations). In such cases, invertibility criteria should be checked for selecting the appropriate parameter values.

Example 9.7.8
Derive the nature of partial autocorrelation function for MA(1) and MA(2) models. Also find the error variance and invertibility condition.

Solution First-Order Moving Average (MA(1)) model
An MA(1) model for a time series $X(t)$ is given by,

$$X(t) = \varepsilon(t) - \theta_1 \varepsilon(t-1)$$

The autocorrelation function for MA(1) model is given as (Eq. 9.67),

$$\begin{aligned} \rho_1 &= -\frac{\theta_1}{\left(1+{\theta_1}^2\right)} \\ \text{or, } \theta_1 &= \frac{-1 \pm \sqrt{1-4\rho_1^2}}{2\rho_1} \end{aligned}$$

This equation gives two estimates of MA(1) model coefficient θ_1. However, the value that will conserve the invertibility condition will be used as estimate of θ_1. For invertibility, the roots of characteristic equation ($\theta(B)$) should lie outside the unit circle.

$$\begin{aligned} \theta(B) &= 0 \\ \text{or, } 1 - \theta_1 B &= 0 \\ \text{or, } B &= \frac{1}{\theta_1} \end{aligned}$$

As $|B| > 1$ so $|\theta_1| < 1$. For an MA(1) model, the partial autocorrelation function (PACF) is given by,

$$\varphi_k = -\frac{\theta_1^k \left(1 - \theta_1^2\right)}{1 - \theta_1^{2(k+1)}}$$

Hence, partial autocorrelation function for MA(1) model decays exponentially for positive θ_1. However, for negative θ_1, the partial autocorrelation function oscillates and damps around zero.

Second-Order Moving Average (MA(2)) model

An MA(2) model for a time series $X(t)$ is given by,

$$X(t) = \varepsilon(t) - \theta_1 \varepsilon(t-1) - \theta_2 \varepsilon(t-2)$$

The autocorrelation function for the MA model with order 2 is given as (Eq. 9.67),

$$\begin{aligned} \rho_1 =& \frac{-\theta_1 (1 - \theta_2)}{1 + \theta_1^2 + \theta_2^2} \\ \rho_2 =& \frac{-\theta_2}{1 + \theta_1^2 + \theta_2^2} \end{aligned}$$

These equations give two estimates of MA(2) model coefficients (θ_1 and θ_2). However, the value pair that will conserve the invertibility condition will be used as estimate of θ_1 and θ_2. Further, the partial autocorrelation function is given by (Eq. 9.58):

$$\begin{aligned} \varphi_1 =& \rho_1 \\ \varphi_2 =& \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \\ \varphi_3 =& \frac{\rho_1^3 - \rho_1 \rho_2 (2 - \rho_2)}{1 - \rho_2^2 - 2\rho_1^2 (1 - \rho_2)} \end{aligned}$$

The nature of PACF for MA(2) is similar to the nature of ACF for AR(2) model. The PACF for MA(2) decays exponentially or damps with oscillation depending on the sign and magnitude of the MA model parameters.

For invertibility, the roots of characteristic equation ($\theta(B)$) should lie outside the unit circle.

$$\begin{aligned} \theta(B) &= 0 \\ \text{or, } 1 - \theta_1 B - \theta_2 B^2 &= 0 \\ \text{or, } B &= \frac{\theta_1 \pm \sqrt{\theta_1^2 + 4\theta_2}}{-2\theta_2} \end{aligned}$$

As $|B| > 1$, Hence $\left|\frac{\theta_1 \pm \sqrt{\theta_1^2 + 4\theta_2}}{-2\theta_2}\right| > 1$, or

$$\theta_2 + \theta_1 < 1$$

$$\theta_2 - \theta_1 < 1$$

$$-1 < \theta_2 < 1$$

Example 9.7.9
Prove that MA(1) model is equivalent to AR(∞) model.

Solution A MA(1) model for a time series $X(t)$ can be expressed as:

$$\begin{aligned}
\varepsilon(t) =& X(t) + \theta_1 \varepsilon(t-1) \\
=& X(t) + \theta_1(X(t-1) + \theta_1 \varepsilon(t-2)) \\
=& X(t) + \theta_1(X(t-1) + \theta_1(X(t-3) + \theta_1 \varepsilon(t-3))) \\
& \ldots \\
=& \sum_{i=0}^{\infty} \theta_1^i X(t-i)
\end{aligned}$$

Hence, a MA(1) model is equivalent to AR(∞) model.

Example 9.7.10
Prove that AR(1) model is equivalent to MA(∞) model.

Solution For a time series $X(t)$, an AR(1) model can be expressed using backshift operator as following

$$\begin{aligned}
X(t) =& \Phi_1 B(X(t)) + \varepsilon(t) \\
=& \Phi_1(\Phi_1 B^2(X(t)) + B(\varepsilon(t))) + \varepsilon(t) \\
=& \Phi_1(\Phi_1(\Phi_1 B^3(X(t)) + B^2(\varepsilon(t))) + B(\varepsilon(t))) + \varepsilon(t) \\
& \ldots \\
=& \sum_{i=0}^{\infty} \Phi_1^i B^i(\varepsilon(t))
\end{aligned}$$

Hence, an AR(1) model is equivalent to MA(∞) model.

Example 9.7.11
Check the invertibility condition for a MA(2) model, the parameters have been estimated as $\theta_1 = 0.2$ and $\theta_2 = 0.5$.

Solution In order to satisfy the stationarity condition, the roots of following equation should lie outside the unit circle,

$$\begin{aligned}
\theta(B) &= 0\\
\text{or, } 1 - \theta_1 B - \theta_2 B^2 &= 0\\
\text{or, } 1 - 0.2B - 0.5B^2 &= 0\\
\text{or, } B &= (-1.628) \text{ or } 1.228
\end{aligned}$$

Both roots are lying outside the unit triangle ($|B| > 1$), so the MA(2) model with parameters $\theta_1 = 0.2$ and $\theta_2 = 0.5$ is invertible.

Example 9.7.12
The first and second parameters of a MA(2) model are 0.65 and 0.3, respectively. Calculate the values of the ACFs and PACFs.

Solution The values of the ACFs can be evaluated using Eq. 9.67.

$$\begin{aligned}
\rho_1 =& \frac{-0.65\,(1 - 0.3)}{1 + 0.65^2 + 0.3^2} = -0.3\\
\rho_2 =& \frac{-0.3}{1 + 0.65^2 + 0.3^2} = -0.198
\end{aligned}$$

The values of the PACFs can be evaluated using Eq. 9.58.

$$\begin{aligned}
\varphi_1 =& \rho_1 = (-0.3)\\
\varphi_2 =& \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} = (-0.316)\\
\varphi_3 =& \frac{\rho_1^3 - \rho_1\rho_2(2 - \rho_2)}{1 - \rho_2^2 - 2\rho_1^2\,(1 - \rho_2)} = (-0.211)
\end{aligned}$$

9.7.6 Auto-Regressive Moving Average (ARMA) Model

Auto-Regressive Moving Average (ARMA) Model is a linear regression model in which current value of time series is estimated using lagged values of time series and the lagged values of residuals. ARMA model is a combination of autoregressive (AR) and moving average (MA) models. In general, ARMA model with pth-order AR model and qth-order MA model (also represented as ARMA(p, q)) is expressed as:

$$X(t) = \sum_{i=1}^{p} \Phi_i X(t - i) + \varepsilon(t) - \sum_{i=1}^{q} \theta_i \varepsilon(t - i) \tag{9.68}$$

where Φ_i and θ_i represent the autoregressive and moving average parameters. ARMA(p, q) is also represented as,

$$\Phi(B)X(t) = \theta(B)\varepsilon(t) \tag{9.69}$$

where $\Phi(B)$ is characteristic function of AR(p) model (i.e., $\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^p$) and $\theta(B)$ is characteristic function of MA(q) model (i.e., $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$).

Properties of ARMA Model

An ARMA model is composed of AR and MA model, so it inherits the properties of these models. Any ARMA model of order (p, q) is stationary if AR(p) is stationary (Sect. 9.7.3), i.e., corresponding characteristic equation $\Phi(B) = 0$ has roots outside the unit circle ($|B| > 1$). Similarly, ARMA (p, q) is invertible (Sect. 9.7.3), if the characteristic equation for MA(q), i.e., $\theta(B) = 0$ has roots outside the unit circle.

Selection of Order of ARMA Model

The estimation of order of ARMA model is done using various methods. Two of those methods are discussed below:

(a) Order selection based on ACF and PACF: The order of the autoregressive component (p) is decided (initial guess) by using PACF. For an AR model, if first p partial autocorrelation coefficients are significant at given level of significance and the autocorrelation function is exponentially decaying, then order is taken as p. The confidence interval of partial autocorrelation function is given by Eq. 9.44. Similarly, order of moving average component (q) depends upon the number of significant ACF of the time series. If the first q partial autocorrelation functions are significant and autocorrelation function is exponentially decreasing for a time series, then the order of MA model is taken as q. The significance of autocorrelation function at any lag (say k) can be judged using Eq. 9.36.

(b) Order selection using canonical correlation analysis: For estimating the ARMA model order for time series $X(t)$ using canonical correlation analysis, two data sets $Y_{m,t} = [X(t)X(t-1)\ldots X(t-m)]^T$ and $Y_{m,t-j-1} = [X(t-j-1)X(t-j-2)\ldots X(t-j-m-1)]^T$ for various combinations for $m = 0, 1, \ldots$ and $j = 0, 1, \ldots$ are considered. Using canonical correlation analysis, different linear combination (loading vectors) of the two data set can be calculated such that it maximizes the correlation coefficients for similar loading vector pairs. Hence, if a_i^T and b_k^T are loading vectors for $Y_{m,t}$ and $Y_{m,t-j-1}$, respectively, then correlation between $a_i^T Y_{m,t}$ and $b_k^T Y_{m,t-j-1}$ is maximized if $i = k$, otherwise they are uncorrelated. Thus, for $m \geq p$ there exists one linear combination of $Y_{m,t}$

$$X(t) - \sum_{i=1}^{p} \psi_i X(t-i) = [\,1\ \psi_1\ \psi_2\ \dots\ \psi_p\ 0\ \dots\ 0\,]Y_{m,t} = a_i^T Y_{m,t} \tag{9.70}$$

such that,

$$a_i^T Y_{m,t} = \varepsilon(t) - \sum_{i=1}^{q} \theta_i \varepsilon(t-i) \tag{9.71}$$

which is uncorrelated with other linear combination of $b_k^T Y_{m,t-j-1}$ ($b_k^T = [1\,\theta_1\ \dots\ \theta_q\ \ 0\ \dots\ 0]$ for $k \neq i$ being the loading vector for $Y_{m,t-j-1}$) for $j \geq q$. Hence, the presence of zero or insignificant canonical correlation loading ($p \leq m$ and $q \leq j$) between $Y_{m,t}$ and $Y_{m,j-t-1}$ for various values of m and j helps in determining the order (p, q) of ARMA model.

It should be noted that the above two methods can be used for initial guess for the order of ARMA model. One needs to generate different ARMA models considering the some variation in the guessed order. The final selection of the most appropriate model order is done based on parsimony of the developed model. A parsimonious model aims to utilize a minimum number of parameters and adequately reproduce the statistics with the least variance. Parsimony of the model is measured using either Akaike Information Criteria (AIC) or Bayesian Information Criteria (BIC), which is discussed in Sect. 9.7.10.

Parameter Estimation of ARMA(p, q) Model

Parameters of ARMA(p, q) model (as expressed in Eq. 9.68) can be estimated either by principle of least square or maximum likelihood. These methods are discussed below

Principle of least square: In this method, the Sum of squared residuals is minimized to get an estimate of ARMA model parameters. In terms of residual ARMA(p, q) is expressed as:

$$\varepsilon(t) = X(t) - \sum_{i=1}^{p} \Phi_i X(t-i) + \sum_{i=1}^{q} \theta_i \varepsilon(t-i) \tag{9.72}$$

Parameter estimation via maximum likelihood: Maximum-likelihood relation of expectation of different moments can be used for parameter estimation. For instance, some of the Maximum-likelihood relationships are expressed as:
AR(1)

$$Var(\Phi_1) \simeq \frac{1 - \Phi_1^2}{n} \tag{9.73}$$

AR(2)

$$Var(\Phi_1) \simeq Var(\Phi_2) \simeq \frac{1-\Phi_2^2}{n} \tag{9.74}$$

$$Cov(\Phi_1, \Phi_2) \simeq -\frac{\Phi_1(1+\Phi_2)}{n} \tag{9.75}$$

$$\text{Hence, correlation between } \Phi_1 \text{ and } \Phi_2 \simeq -\frac{\Phi_1}{1-\Phi_2} = -\rho_1 \tag{9.76}$$

MA(1)

$$Var(\theta_1) \simeq \frac{1-\theta_1^2}{n} \tag{9.77}$$

MA(2)

$$Var(\theta_1) \simeq Var(\theta_2) \simeq \frac{1-\theta_2^2}{n} \tag{9.78}$$

$$Cor(\theta_1, \theta_2) \simeq -\frac{\theta_1(1+\theta_2)}{n} \tag{9.79}$$

$$\text{Hence, correlation between } \theta_1 \text{ and } \theta_2 \simeq -\frac{\theta_1}{1-\theta_2} = -\rho_1 \tag{9.80}$$

ARMA(1, 1)

$$Var(\Phi_1) \simeq \frac{1-\Phi_1^2}{n}\left(\frac{1-\Phi_1\theta_1}{\Phi_1-\theta_1}\right)^2 \tag{9.81}$$

$$Var(\theta_1) \simeq \frac{1-\theta_1^2}{n}\left(\frac{1-\Phi_1\theta_1}{\Phi_1-\theta_1}\right)^2 \tag{9.82}$$

$$Cov(\Phi_1, \theta_1) \simeq \frac{(1-\theta_1^2)(1-\Phi_1^2)(1-\theta_1\Phi_1)}{n(\Phi_1-\theta_1)^2} \tag{9.83}$$

$$\text{Hence, correlation between } \Phi_1 \text{ and } \theta_1 \simeq \frac{\sqrt{(1-\theta_1^2)(1-\Phi_1^2)}}{1-\theta_1\Phi_1} \tag{9.84}$$

Example 9.7.13

Check for stationarity and invertibility conditions for an ARMA(2, 2) model, if the model parameters are $\Phi_1 = 0.3$, $\Phi_2 = 0.5$, $\theta_1 = 0.3$, and $\theta_2 = -0.5$.

Solution Check for Stationarity Condition

$$\begin{aligned}
&\Phi(B) = 0\\
&\text{or, } 1 - \Phi_1 B - \Phi_2 B^2 = 0\\
&\text{or, } 1 - 0.3B - 0.5B^2 = 0
\end{aligned}$$

The roots of the equation are (-1.746) and 1.146. As both the roots lie outside the unit circle ($|B| > 1$) therefore, the parameters satisfy the stationarity condition.

Check for Invertibility Condition

$$\begin{aligned}
&\theta(B) = 0\\
&\text{or, } 1 - \theta_1 B - \theta_2 B^2 = 0\\
&\text{or, } 1 - 0.3B + 0.5B^2 = 0
\end{aligned}$$

The roots of the equation are $0.3 \pm 1.382i$. As $|B| > 1$ therefore, the parameters satisfy the invertibility condition.

9.7.7 Autoregressive Integrated Moving Average (ARIMA) Model

Autoregressive Integrated Moving Average (ARIMA) model is used for modeling non-stationary time series. The time series is transformed to a stationary time series by using a series of differencing operator. ARMA is then applied on the resulting time series. If p is the order of autoregressive model, d is the order of differencing operator and q is the order of moving average model then the ARIMA is represented as ARIMA(p, d, q). In general ARIMA(p, d, q) is expressed as:

$$\Phi(B)\nabla^d X(t) = \theta(B)\varepsilon(t) \tag{9.85}$$

where $X(t)$ is non-stationary time series. $\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ are characteristic functions for AR(p) and MA(q) model, respectively. ∇ represents the differencing operation. The order of differencing (d) is decided based on the stationarity of resulting time series. The autoregressive and moving average orders and parameters of the ARIMA are decided in the same way as in ARMA. As after differencing in ARIMA the time series is stationary and ARMA is then used, so stationarity and invertibility criteria for ARIMA are same as that of ARMA. For stationarity and invertibility of ARIMA model, the root of equations $\Phi_p(B) = 0$ and $\theta_q(B) = 0$ should lie outside unit circle (Sects. 9.7.3 and 9.7.3).

Example 9.7.14

Fit an ARIMA(2,1,0) model over the data provided in Example 9.6.1.

Solution For fitting ARIMA(2,1,0), 1st differencing of the precipitation time series is needed (Table 9.3). The ARMA(2,0) or AR(2) model is fitted on the time series obtained after differencing.

Covariance matrix at lag 1 for the differencing series (Table 9.3) is given by:

$$\text{cov}_1 = \text{cov}\left(\begin{bmatrix} 4.50 & 16.49 \\ 16.49 & -13.29 \\ -13.29 & -4.68 \\ -4.68 & -4.38 \\ -4.38 & 1.95 \\ 1.95 & 53.06 \\ 53.06 & -30.35 \\ -30.35 & -19.84 \\ -19.84 & 31.74 \\ 31.74 & -38.08 \\ -38.08 & 3.02 \\ 3.02 & 38.54 \\ 38.54 & 3.16 \\ 3.16 & -23.34 \\ -23.34 & -5.52 \\ -5.52 & -14.65 \\ -14.65 & 20.53 \\ 20.53 & -21.54 \end{bmatrix}\right) = \begin{bmatrix} 576.32 & -193.55 \\ -193.55 & 603.41 \end{bmatrix}$$

and the corresponding autocorrelation at lag 1 $= -193.55\big/\sqrt{(576.32 \times 603.41)} = -0.328$.

Hence, $\rho_1 = -0.328$. Similarly,

$$\text{cov}_2 = \begin{bmatrix} 587.21 & -211.08 \\ -211.08 & 622.24 \end{bmatrix}$$

and the corresponding autocorrelation matrix at lag 2, $\rho_2 = -0.349$. For AR(2) model, the parameters are given by Eq. 9.58:

$$\begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix}\begin{bmatrix} \Phi_1 \\ \Phi_2 \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}$$

$$\text{or,} \begin{bmatrix} 1 & -0.328 \\ -0.328 & 1 \end{bmatrix}\begin{bmatrix} \Phi_1 \\ \Phi_2 \end{bmatrix} = \begin{bmatrix} -0.328 \\ -0.349 \end{bmatrix}$$

Hence, $\Phi_1 = -0.239$ and $\Phi_2 = 0.27$. Thus, the model is expressed as

$$Y(t) = -0.239Y(t-1) + 0.27Y(t-2) + \varepsilon(t)$$

where $Y(t) = X(t) - X(t-1)$.

9.7.8 Autoregressive Moving Average Model with Exogenous Inputs (ARMAX)

The models discussed above, i.e., AR, MA, ARMA, and ARIMA are developed using the information from the same time series, and these models do not consider any other variables/time series. However, in many cases in hydroclimatology, the time series under study (say precipitation) associated with other influencing time series (like air temperature, pressure), etc. Hence, for modeling these kind of interrelationships, the model should be able to use the information from the causal variable/time series known as exogenous input. Autoregressive Moving Average Model with Exogenous Inputs (ARMAX) consists of an ARMA model and weighted sum of lagged values of exogenous time series. For an ARMAX model, if the r lagged value of exogenous time series is used and the ARMA part is of order (p, q), then the ARMAX model is said to be of the order of (p, q, r). In general, ARMAX model with order (p, q, r) is expressed as:

$$X(t) = \sum_{i=1}^{p} \Phi_i X(t-i) + \varepsilon(t) - \sum_{j=1}^{q} \theta_j \varepsilon(t-j) + \sum_{k=1}^{r} \psi_k I(t-k) \tag{9.86}$$

where $X(t)$ is stationary time series. ψ_k $(k = 1, 2, \ldots, r)$ is the weighting coefficients associated with lagged values of exogenous stationary time series $I(t)$. Φ_i $(i = 1, 2, \ldots, p)$ and θ_j $(j = 1, 2, \ldots, q)$ are autoregressive and moving average parameters, respectively.

Estimation of ARMAX Parameters

The parameters of the ARMAX model are estimated by minimizing the sum of square of prediction errors. Sum of square of prediction errors for Eq. 9.86 is given by:

$$\begin{aligned}\sum(\varepsilon(t))^2 &= \sum(X(t)-\hat{X}(t))^2\\ &= \sum\left(X(t)-\sum_{i=1}^{p}\Phi_i X(t-i)+\sum_{j=1}^{q}\theta_j\varepsilon(t-j)-\sum_{k=1}^{r}\psi_k I(t-k)\right)^2\end{aligned} \tag{9.87}$$

For minimizing the Sum of square error, the above equation is partially differentiated with respect to each parameter and is equated to zero. Hence, a total of following $(p+q+r)$ equations is obtained.

$$\left.\begin{aligned}\frac{\partial\sum(\varepsilon(t))^2}{\partial\Phi_i}&=0 \qquad \text{for } i=1,2,\ldots,p\\ \frac{\partial\sum(\varepsilon(t))^2}{\partial\theta_j}&=0 \qquad \text{for } j=1,2,\ldots,q\\ \frac{\partial\sum(\varepsilon(t))^2}{\partial\psi_k}&=0 \qquad \text{for } k=1,2,\ldots,r\end{aligned}\right\} \tag{9.88}$$

Simultaneous solution of all above equations will provide the estimate of parameters.

Identification of ARMAX Orders

With a initial guess of p, q, and r and its variation, a number of ARMAX model can be estimated. The identification of most suited ARMAX model is done on the basis of following three criteria:

(a) The prediction focus or the model fit (MF)
(b) The Mean Square Error (MSE) function
(c) The Akaike's Final Prediction Error (FPE).

Let $\hat{X}(t)$ represents the estimated time series using ARMAX model for observed time series $X(t)$. The MF and MSE are expressed as,

$$\text{MF} = 100\left(1-\frac{\sqrt{\sum_{i=1}^{n}\left(\hat{X}(t)-X(t)\right)^2}}{\sqrt{\sum_{i=1}^{n}\left(\hat{X}(t)-\overline{X}\right)^2}}\right) \tag{9.89}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{X}(t)-X(t)\right)^2 \tag{9.90}$$

where $\overline{X}$ represents the mean of observed time series $X(t)$ and n is number of observations in $X(t)$. Higher value of MF is considered favorable. However, the lower value of MSE indicates better model performance. The range of MF is 0 to 100%, whereas MAE may vary form 0 to ∞.

Akaike's Final Prediction Error (FPE) compares both the error or residual of the model and effect of number of model parameters. FPE is given by,

$$\text{FPE} = V\left(\frac{1 + m/n}{1 - m/n}\right) \tag{9.91}$$

where m is the number of estimated parameters, i.e., $p + q + r$, n is number of observation in time series $X(t)$ and V is loss function. Mathematically, the loss function is the determinant of error or residual series ($\varepsilon(t)$). Hence,

$$V = \det(\text{cov}(\varepsilon(t))) \tag{9.92}$$

If $m \ll n$, the FPE is approximated as,

$$\text{FPE} = V\left(1 + \frac{2m}{n}\right) \tag{9.93}$$

The range of FPE is 0 to ∞. The smaller the value of FPE is, the better is the fitted model.

9.7.9 Forecasting with ARMA/ARMAX

Forecasting is the process of estimating future values of a time series, often using the past (or lagged) values of the same or other causal time series. Forecasting of hydroclimatic variables is important for making future plans/policies or preparedness for future extremes, if any. For instance, flood prediction system can be used for as early warning system and hence helps in evacuation. The procedure of forecasting can be used to estimate the past values of time series, this process is called hindcasting.

The forecast depends on the time step till which the information is being used (also known as origin of forecast). The difference in time step for which a forecast is made and the origin of forecast is called lead period. With the increase in lead period, the utility of forecast increases. However, the uncertainty in forecast also increases with increase in lead period. Hence, a suitable lead period can be used as compromise between two contrasting requirements. Further, a forecasting model can be static or dynamic. For static forecasting model the parameters once estimated do not change with time. However, for dynamic forecasting model the parameters change with time. The change in parameters of dynamic forecasting model tries to incorporate the information available in new observation(s) if any to enhance the prediction performance.

ARMA (or similar models) can be used for forecasting of hydroclimatic time series. In general, ARMA model is given by (Eq. 9.69):

$$\Phi(B)X(t) = \theta(B)\varepsilon(t) \tag{9.94}$$

where $\Phi(B)$ is characteristic function of AR(p) model (i.e., $\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^p$) and $\theta(B)$ is characteristic function of MA(q) model (i.e., $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$).

The forecast at a lead period of l with origin of t can be obtained using following relationships.

$$\begin{aligned} X_t(t+l) =& \Phi_1 X(t+l-1) + \Phi_2 X(t+l-2) \cdots + \Phi_p X(t+l-p) \\ & - \theta_1 \varepsilon(t+l-1) - \theta_2 \varepsilon(t+l-2) \cdots - \theta_q \varepsilon(t+l-q) + \varepsilon(t+l) \\ \hat{X}_t(t+l) =& E(X_t(t+l)) = \Phi_1 X(t+l-1) + \Phi_2 X(t+l-2) \cdots + \Phi_p X(t+l-p) \\ & - \theta_1 \varepsilon(t+l-1) - \theta_2 \varepsilon(t+l-2) \cdots - \theta_q \varepsilon(t+l-q) \end{aligned} \tag{9.95}$$

where t is considered current time step or origin for forecast and l is lead period. Forecast depends on origin, hence, with new observations available (shift in origin) the forecast needs to be updated. The updated forecast can be obtained using the Eq. 9.94. One alternate method of correcting the forecast is to utilize the difference (or error) in new observation and its earlier forecast. This process is described as follows. Suppose that $X_t(t+l)$ and $X_{t+1}(t+l)$ are two different forecasts for time step $(t+l)$ using the information till tth and $(t+1)$th time steps. These forecasts can be represented in the form of linear function of deviations or residuals as following:

$$\begin{aligned} \hat{X}_{t+1}(t+l) &= \lambda_j \varepsilon(t+1) + \lambda_{j+1}\varepsilon(t) + \lambda_{j+2}\varepsilon(t-1) + \cdots \\ \hat{X}_t(t+l) &= \qquad\qquad \lambda_{j+1}\varepsilon(t) + \lambda_{j+2}\varepsilon(t-1) + \cdots \\ \text{or, } \hat{X}_{t+1}(t+l) &= \lambda_j \varepsilon(t+1) + X_t(t+l) \end{aligned} \tag{9.96}$$

where $j = l - 1$. The parameter λ_l can be obtained using following equation:

$$\lambda_j = \Phi_1 \lambda_{j-1} + \Phi_2 \lambda_{j-2} + \cdots + \Phi_p \lambda_{j-p} - \theta_j \tag{9.97}$$

where Φ_i and θ_i are ith autocorrelation and moving average parameters from Eq. 9.94. $\lambda_0 = 1$, $\lambda_l = 0$ for $l < 0$ and $\theta_l = 0$ for $l > q$. Hence, the correction parameters (λ_i) are given as:

$$\lambda_0 = 1 \tag{9.98}$$

$$\lambda_1 = \Phi_1 - \theta_1 \tag{9.99}$$

$$\lambda_2 = \Phi_1 \lambda_1 + \Phi_2 - \theta_2 \tag{9.100}$$

$$\lambda_3 = \sum_{i=1}^{3} \Phi_i \lambda_{3-i} - \theta_3 \tag{9.101}$$

$$\lambda_k = \sum_{i=1}^{k} \Phi_i \lambda_{k-i} - \theta_k \tag{9.102}$$

These parameters λ_j for $j \in \{0, 1, 2, \ldots\}$ depend upon the fitted model, not on the values of time series and do not change with new observations.

Confidence Interval of Forecast

Forecast at future time step $t + l$ using the information of time series till t time step has some uncertainty associated with it. The mean of time series ($X(t)$) can also be taken as the forecasted value, i.e., $\hat{X}_t(t + l)$. The forecasted values are assumed to follow normal distribution. The standard deviation of the forecast can be estimated by calculating expectation of square of Eq. 9.96. The variance of forecast with lead period (l) is given by $\left(\sum\limits_{j=0}^{l-1} \lambda_j^2\right) \sigma_\varepsilon^2$. Hence, the confidence interval of the forecast with lead period of l at α level of significance is given by:

$$\left[\hat{X}_t(t+l) - Z_{(\alpha/2)}\sigma_\varepsilon \sqrt{1 + \sum_{j=1}^{l-1} \lambda_j^2},\ \hat{X}_t(t+l) + Z_{(\alpha/2)}\sigma_\varepsilon \sqrt{1 + \sum_{j=1}^{l-1} \lambda_j^2}\right] \tag{9.103}$$

where $Z_{(\alpha/2)}$ is standard normal variate at $(1 - \alpha/2) \times 100\%$ probability. It can be observed that with increase in forecast step the variance and hence the confidence interval of the forecast increases very fast. So, forecast with longer lead period has more uncertainty involved with them compared to shorter lead period forecast.

Analysis of Forecast Errors

Forecast errors are the measure of the deviation of forecast from the observation. Suppose that for a time series $X(t)$ if $X(t + l)$ is observation and corresponding forecast value using the time series information till time step t is $X_t(t + l)$ then the forecast error can be expressed as one of following statistics:

(i) **Mean Square Error**: This statistics represents the mean square deviation of forecasted values from the observed values of time series.

$$\text{MSE} = \frac{1}{N} \sum_{l=1}^{N} (X(t+l) - \hat{X}_t(t+l))^2 \tag{9.104}$$

where N is the number of elements in forecasted series $\hat{X}_t(t+1)$.

(ii) **Mean Absolute Percentage Error**: This statistics represents the mean percentage deviation of forecasted values with respect to the observed values of time series.

$$\text{MAPE} = \frac{100}{N} \sum_{l=1}^{N} \left| \frac{X(t+l) - \hat{X}_t(t+l)}{X(t+1)} \right| \tag{9.105}$$

where N is the number of elements in forecasted series $\hat{X}_t(t+1)$.

(iii) **Mean Absolute Error**: This statistics represents the mean absolute deviation of forecasted values from the observed values of time series.

$$\text{MAE} = \frac{1}{N} \sum_{l=1}^{N} \left| X(t+l) - \hat{X}_t(t+l) \right| \tag{9.106}$$

where N is the number of elements in forecasted series $\hat{X}_t(t+1)$.

Example 9.7.15

Daily rainfall depth ($X(t)$ in mm/day) at a location is found to follow an ARMA(3,1) model given by:

$$X(t) = 0.9X(t-1) + 0.5X(t-2) - 0.3X(t-3) + \varepsilon(t) - 0.3\varepsilon(t-1)$$

If daily observed values of time series $X(t)$ for a week are 0, 8.4, 11.84, 16.52, 17.12, 21.20, and 16.85, then

(a) Forecast daily rainfall depth for next week.

(b) If next observed value in $X(t)$, i.e., $X(8)$ is 14.70 mm then update the forecast.

(c) Assuming the variance of residual to be 5, calculate the variance of the forecast with lead periods of 1, 3 and 5.

(d) Find the mean absolute deviation of forecast (with origin at $X(7)$), if the observed rainfall for 8th–14th days are 14.7, 10.5, 6.7, 13.0, 0, 0, 2 respectively. Compare it with the mean absolute deviation of forecast made at origin 8.

Solution

(a) From the observed values of series $X(t)$, the error series can be calculated with assumption that first three errors are assumed to be zero. Hence, $\varepsilon(i) = 0$ for $i \in \{1, 2, 3\}$

$$\begin{aligned}
\varepsilon(4) =& X(4) - 0.9X(4-1) - 0.5X(4-2) + 0.3X(4-3) + 0.3\varepsilon(3) \\
=& 16.52 - 0.9 \times 11.84 - 0.5 \times 8.40 = 1.664 \\
\varepsilon(5) =& X(5) - 0.9X(4) - 0.5X(3) + 0.3X(2) + 0.3\varepsilon(4) \\
=& 17.12 - 0.9 \times 16.52 - 0.5 \times 11.84 + 0.3 \times 8.40 + 0.3 \times 1.664 = -0.649
\end{aligned}$$

Similarly, $\varepsilon(6) = 0.889$ and $\varepsilon(7) = -5.567$. The forecasts for next 7 days can be obtained using Eq. 9.95.

$$\begin{aligned}\hat{X}_7(8) =& 0.9X(7) + 0.5X(6) - 0.3X(5) - 0.3\varepsilon(7)\\ =& 0.9 \times 16.85 + 0.5 \times 21.20 - 0.3 \times 17.12 + 0.3 \times 5.567 = 22.30\\ \hat{X}_7(9) =& 0.9\hat{X}_7(8) + 0.5X(7) - 0.3X(6)\\ =& 0.9 \times 22.30 + 0.5 \times 16.85 - 0.3 \times 21.20 = 22.135\end{aligned}$$

Similarly, $\hat{X}_7(10) = 26.016$, $\hat{X}_7(11) = 27.792$, $\hat{X}_7(12) = 31.380$, $\hat{X}_7(13) = 34.333$ and $\hat{X}_7(14) = 38.252$.

(b) Now, the observed value of $X(8)$ is 14.70, then residual for 8th day

$$\begin{aligned}\varepsilon =& X(8) - 0.9X(7) - 0.5X(6) + 0.3X(5) + 0.3\varepsilon(7)\\ =& X(8) - \hat{X}_7(8) = 14.70 - 22.30 = -7.6\end{aligned}$$

The forecast can be updated using the new information provided as observed value of daily rainfall for 8th day, i.e., $X(8)$ by using Eq. 9.96. We need to calculate the correction parameters (λ_i for $i \in \{0, 1, \ldots, 6\}$) by using Eq. 9.97.

$$\begin{aligned}&\lambda_0 = 1\\ &\lambda_1 = \Phi_1 - \theta_1 = 0.9 - 0.3 = 0.6\\ &\lambda_2 = \Phi_1\lambda_1 + \Phi_2 - \theta_2 = 0.9 \times 0.6 + 0.5 - 0 = 1.04\\ &\lambda_3 = \sum_{i=1}^{3} \Phi_i\lambda_{3-i} - \theta_3 = 0.9 \times 1.04 + 0.5 \times 0.6 - 0.3 \times 1 = 0.936\\ &\lambda_4 = \sum_{i=1}^{4} \Phi_i\lambda_{4-i} - \theta_4 = 0.9 \times 0.936 + 0.5 \times 1.04 - 0.3 \times 0.6 = 1.182\\ &\lambda_5 = \sum_{i=1}^{5} \Phi_i\lambda_{5-i} - \theta_5 = 0.9 \times 1.182 + 0.5 \times 0.936 - 0.3 \times 1.04 = 1.220\\ &\lambda_6 = \sum_{i=1}^{6} \Phi_i\lambda_{6-i} - \theta_6 = 0.9 \times 1.220 + 0.5 \times 1.182 - 0.3 \times 0.936 = 1.408\end{aligned}$$

As $\lambda_0 = 1$, so $\hat{X}_8(8) = \hat{X}_7(8) + \lambda_1\varepsilon(8) = 14.70 = X(8)$. The updated forecasts are

$$\begin{aligned}\hat{X}_8(9) =& \hat{X}_7(9) + \lambda_1\varepsilon(8) = 22.135 + 0.6(-7.6) = 17.57\\ \hat{X}_8(10) =& \hat{X}_7(10) + \lambda_2\varepsilon(8) = 26.016 + 1.04(-7.6) = 18.11\\ \hat{X}_8(11) =& \hat{X}_7(11) + \lambda_3\varepsilon(8) = 27.792 + 0.936(-7.6) = 20.68\end{aligned}$$

$$\hat{X}_8(12) = \hat{X}_7(12) + \lambda_4\varepsilon(8) = 31.380 + 1.182(-7.6) = 22.40$$
$$\hat{X}_8(13) = \hat{X}_7(13) + \lambda_5\varepsilon(8) = 34.333 + 1.220(-7.6) = 25.06$$
$$\hat{X}_8(14) = \hat{X}_7(14) + \lambda_6\varepsilon(8) = 38.252 + 2.425(-7.6) = 27.55$$

The same forecast can be obtained by utilization of given ARMA model; however, this method has two advantages first being λ_i's do not change with new observation and correction to the forecast is done by adding some factor of difference of new observation and its old forecast. In ARMA, one needs to calculate the error series repeatedly for $t = 1$ for updating any forecast, which makes it cumbersome.

(c) The variance of forecast with lead 1 is given by:

$$\text{Var}_1 = \sigma_\varepsilon^2 = 5$$

Similarly,

$$\text{Var}_3 = \left(\sum_{j=0}^{2} \lambda_j^2\right) \sigma_\varepsilon^2 = 5(1 + 0.6^2 + 1.04^2) = 12.21$$
$$\text{Var}_5 = \left(\sum_{j=0}^{4} \lambda_j^2\right) \sigma_\varepsilon^2 = 5(1 + 0.6^2 + 1.04^2 + 0.936^2 + 1.182^2) = 23.57$$

Hence, with increase in lead period the variance of forecast increases.

(d) Mean Absolute error for the forecast made at origin 7 is,

$$\text{MAE}_7 = \frac{1}{7}\sum_{l=1}^{7} \left| X(7+l) - \hat{X}_7(7+l) \right| = 22.186$$

Mean Absolute error for the forecast made at origin 8 is,

$$\text{MAE}_8 = \frac{1}{6}\sum_{l=1}^{6} \left| X(8+l) - \hat{X}_8(8+l) \right| = 16.53$$

Hence, it can be observed that inclusion of new observations leads to decrease in forecast error.

9.7.10 *Parsimony of Time Series Models*

A parsimonious model should utilize minimum number of parameters and adequately reproduce the statistics with least variance. Parsimony of the model can be used as selection criteria for the model if they are reasonably close in prediction performance. Parsimony can be measured using following two criteria:

(i) **Akaike Information Criterion**: For an ARMA(p, q) model, the Akaike Information Criterion (AIC) is defined as:

$$\text{AIC}(p, q) = N \ln(\sigma_\varepsilon^2) + 2(p + q) \tag{9.107}$$

where σ_ε^2 is Maximum-likelihood estimate of variance of the residual series with N elements. The model with least AIC is selected.

(ii) **Bayesian Information Criterion**: For an ARMA(p, q) model, the Bayesian Information Criterion (BIC) is defined as:

$$\text{BIC}(p, q) = N \ln\left[\frac{\sigma_\varepsilon^2 M}{N}\right] + (p + q) \ln\left[\frac{M\left(\frac{\sigma_X^2}{\sigma_\varepsilon^2} - 1\right)}{p + q}\right] \tag{9.108}$$

where σ_ε^2 is Maximum-likelihood estimate of variance of the residual series, σ_X^2 is variance of time series $X(t)$ with N elements and $M = N - (p + q)$. The model with least BIC is selected.

For selecting the best-suited ARMA model from a pool of feasible ARMA models (with different orders), AIC should be preferred over BIC.

Example 9.7.16

Calculate the Akaike Information Criteria for two AR models developed in Example 9.7.6. Assume that the length of time series is 40.

Solution For AR(1) model, length of residual series $(N) = 40 - 1 = 39$, $p = 1$ and $q = 0$

$$\text{AIC}(1, 0) = N \ln(\sigma_\varepsilon^2) + 2(p + q) = 39 \ln(12.8) + 2 = 101.42$$

For AR(2) model, length of residual series $(N) = 40 - 2 = 38$, $p = 2$ and $q = 0$

$$\text{AIC}(2, 0) = N \ln(\sigma_\varepsilon^2) + 2(p + q) = 38 \ln(12) + 4 = 98.43$$

Hence, as per lower AIC criteria, AR(2) is a better model.

Example 9.7.17

Three MA models are developed for a time series having unit variance. The length of time series is 50. The parameter for MA(1) model is 0.7. The parameters of MA(2) model are $\theta_1 = 0.3$ and $\theta_2 = 0.45$. The parameters of MA(3) model are $\theta_1 = 0.2$, $\theta_2 = 0.3$, and $\theta_3 = 0.37$. Based on AIC criteria, select the best order for MA model.

Solution For MA(1) model, length of residual series $(N) = 50 - 1 = 49$, $p = 0$ and $q = 1$. The variance of error is given by (Eq. 9.64)

$$\sigma_\varepsilon^2 = \frac{1}{1+\theta_1^2} = \frac{1}{1+0.7^2} = 0.671$$

$$\text{AIC}(0, 1) = N\ln(\sigma_\varepsilon^2) + 2(p+q) = 49\ln(0.671) + 2 = -17.55$$

For MA(2) model, length of residual series $(N) = 50 - 2 = 48$, $p = 0$ and $q = 2$

$$\sigma_\varepsilon^2 = \frac{1}{1+\theta_1^2+\theta_2^2} = \frac{1}{1+0.3^2+0.45^2} = 0.774$$

$$\text{AIC}(0, 2) = N\ln(\sigma_\varepsilon^2) + 2(p+q) = 48\ln(0.774) + 4 = -8.29$$

For MA(3) model, length of residual series $(N) = 50 - 3 = 47$, $p = 0$ and $q = 3$

$$\sigma_\varepsilon^2 = \frac{1}{1+\theta_1^2+\theta_2^2+\theta_3^2} = \frac{1}{1+0.2^2+0.3^2+0.37^2} = 0.789$$

$$\text{AIC}(0, 3) = N\ln(\sigma_\varepsilon^2) + 2(p+q) = 47\ln(0.789) + 6 = -5.13$$

Hence, as per lower AIC criteria, best order for MA model is 1.

9.7.11 Diagnostic Check for ARMA Models

Adequacy of an ARMA model can be checked by analysis of residual series $(\varepsilon(t))$. The residual are the difference between observed and modeled time series. Most of the models discussed in previous sections (AR, MA, ARMA, or ARMAX) are linear regression models, so the residual series is assumed to be aperiodic, independent, and identically distributed with zero mean. These assumptions about the residual series are required to be checked for accessing the adequacy of model.

Test for Independence

The residual series is considered independent when the autocorrelation function at nonzero lag is zero. This criteria can be checked using autocorrelogram or by statistical test like Portmanteau lack of fit test. Portmanteau test statistic (Q) is given by:

$$Q = N\sum_{i=1}^{k}(\rho_i(\varepsilon))^2 \tag{9.109}$$

where N is the length of residual series, $\rho_i(\varepsilon)$ represent the autocorrelation in the residual series, and k is highest lag considered, which is generally more than $N/5$. The test statistic (Q) approximately follows χ^2 distribution with $(k-q-p)$ degree of freedom. If $Q < \chi^2_\alpha(k-p-q)$ at α level of significance then the residual series can be considered independent. Some researchers have proposed a modified statistics (modified Ljung–Box–Pierce statistics denoted as $\overline{Q}$) for this test. The modified Ljung–Box–Pierce statistics is given by:

$$\overline{Q} = N(N+2)\sum_{i=1}^{k}\frac{(\rho_i(\varepsilon))^2}{N-i} \tag{9.110}$$

The modified Ljung–Box–Pierce statistic is also recommended to be used in Portmanteau test as it follows $\chi^2(k-p-q)$ better as compared to Q (Eq. 9.109).

Test for Normal Distribution for Residual Series

The residual series from an ideal model should be independent and identically distributed. The residual from ARMA model should follow normal distribution. For checking that the residual series follows normal distribution, normal probability paper can be used. Some statistical tests like chi-square (χ^2) test, Kolmogorov–Smirnov test, Anderson–Darling test, skewness test (discussed in Sects. 6.4.4 and 9.6.1) can be used for checking that the residual series is normally distributed or not. If the residual series does not follow normal distribution (i.e., null hypothesis is rejected for above tests), then the original time series can be transformed using the data transformation techniques given in Sect. 9.6.

Test for Periodicity

The periodicity in residual series (if any) can be observed in cumulative periodogram (Sect. 9.5.1). The periodicity of frequency ν_i can be tested for statistical significance using the following statistic:

$$F(\varepsilon(t)) = \frac{\gamma^2(N-2)}{4\beta} \tag{9.111}$$

where

$$\gamma^2 = a^2 + b^2 \tag{9.112}$$

$$\beta = \frac{\sum_{t=1}^{N}[\varepsilon(t) - a\cos(2\pi\nu_i t) - b\sin(2\pi\nu_i t)]^2}{N} \tag{9.113}$$

where a and b are Fourier transform parameters for frequency ν_i (Eqs. 9.17 and 9.18). The statistic $F(\varepsilon(t))$ follows F-distribution with 2 and $(N-2)$ degree of freedom. Hence, for α level of significance if $F(\varepsilon(t)) \leq F_\alpha(2, N-2)$, then the periodicity corresponding to frequency ν_i is considered not significant.

Test for Zero Mean

To test whether the residual series has zero mean or not, the $T(\varepsilon)$ can be calculated using:

$$T(\varepsilon) = \frac{\overline{\varepsilon}(t) - \mu_\varepsilon}{SE_\varepsilon} \tag{9.114}$$

where $\overline{\varepsilon}(t)$ and SE_ε are the mean and standard deviation of residual series. μ_ε is expected value of residual mean, hence, in case of checking whether the mean is zero, $\mu_\varepsilon = 0$. The statistics $T(\varepsilon)$ approximately follows Student-t distribution with $N-1$ degree of freedom. At α level of significance, if $|T(\varepsilon)| \leq t_{\alpha/2}(N-1)$, then the mean is considered not to differ from μ_ε.

Example 9.7.18
The residual for the ARMA(2,2) model are 1.32, −1.97, −10.88, −5.98, 1.83, 12.06, 3.70, 1.55, −2.71, 0.61, 4.81, −1.27, 9.46, −6.10, −1.88, 0.260, −9.77, 2.83, 0.390, 0.400, 3.97, 5.22, 4.01, −2.34, −0.230, 1.77, −6.28, −5.18, −2.13, and 0.390, respectively. Check that residual can be considered as white noise at a 5% level of significance.

Solution
Test of Independence

Null Hypothesis: Residuals are independent.
Alternative Hypothesis: Residuals are not independent.
Level of Significance: $\alpha = 5\%$

For checking the independence, the 6 lagged autocorrelations are considered. The autocorrelation function for lags 1 to 6 is 0.156, −0.049, −0.175, −0.311, 0.069, and −0.150.

The test statistics $\overline{Q}$ is given by

$$\overline{Q} = N(N+2)\sum_{i=1}^{k}\frac{(\rho_i(\varepsilon))^2}{N-i} = 6.614$$

For $k-2 = 4$ degree of freedom $\chi^2_\alpha(4) = 9.488$. As $6.614 < 9.488$, so the null hypothesis of data being independent cannot be rejected.

Test for Normal distribution using skewness

Null Hypothesis: Data is normally distributed.
Alternative Hypothesis: Data is not normally distributed.
Level of Significance: $\alpha = 5\%$

The calculation of skewness of residual series is given in Table 9.4. The skewness of residual series is given by (Eq. 9.31)

Table 9.4 Calculation for skewness test of residual series

S. No.	Residual series $Y(t)$	Residual series deviation $Y_d(t)$	$Y_d(t)^2$	$Y_d(t)^3$
1	1.32	1.39	1.94	2.69
2	−1.97	−1.9	3.6	−6.84
3	−10.88	−10.81	116.83	−1262.75
4	−5.98	−5.91	34.91	−206.29
5	1.83	1.9	3.62	6.87
6	12.06	12.13	147.17	1785.36
7	3.7	3.77	14.22	53.64
8	1.55	1.62	2.63	4.26
9	−2.71	−2.64	6.96	−18.37
10	0.61	0.68	0.46	0.32
11	4.81	4.88	23.83	116.31
12	−1.27	−1.2	1.44	−1.72
13	9.46	9.53	90.85	865.89
14	−6.1	−6.03	36.34	−219.11
15	−1.88	−1.81	3.27	−5.92
16	0.26	0.33	0.11	0.04
17	−9.77	−9.7	94.06	−912.3
18	2.83	2.9	8.42	24.42
19	0.39	0.46	0.21	0.1
20	0.4	0.47	0.22	0.1
21	3.97	4.04	16.33	66
22	5.22	5.29	28	148.15
23	4.01	4.08	16.66	67.98
24	−2.34	−2.27	5.15	−11.68
25	−0.23	−0.16	0.03	0
26	1.77	1.84	3.39	6.24
27	−6.28	−6.21	38.55	−239.33
28	−5.18	−5.11	26.1	−133.33
29	−2.13	−2.06	4.24	−8.72
30	0.39	0.46	0.21	0.1
Total	−2.14	0	729.74	122.12

$$S = \frac{\frac{1}{N}\sum_{i=1}^{N}(X(t)-\overline{X})^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(X(t)-\overline{X})^2\right]^{3/2}} = \frac{\frac{122.12}{30}}{\left(\frac{729.74}{30}\right)^{1.5}} = 0.0339$$

Null hypothesis is acceptable for $|S| < Z_{(\alpha/2)}\sqrt{6/N}$ (Eq. 9.32)
As, $Z_{(\alpha/2)}\sqrt{6/N} = Z_{0.025}\sqrt{6/20} = 1.074$
and $|S| < 1.074$, i.e., $0.033 < 1.074$, so the null hypothesis of data being normally distributed cannot be rejected.

Test for zero mean

Null Hypothesis: Population mean of residual series is zero, i.e., $\mu_\varepsilon = 0$
Alternative Hypothesis: Population mean of residual series is zero, i.e., $\mu_\varepsilon \neq 0$
Level of Significance: $\alpha = 5\%$

The test statistics $T_\varepsilon = \overline{\varepsilon}/SE_\varepsilon = -0.0713/5.016 = -0.014$.
T_ε follows Student-t distribution with $N - 1 = 29$ degree of freedom.

$$t_{\alpha/2}(29) = 2.045$$

As $|-0.01| < 2.045$, so the null hypothesis of data having zero mean cannot be rejected.

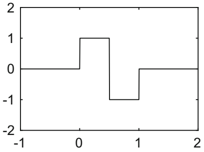
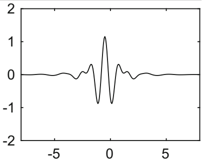
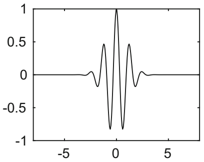
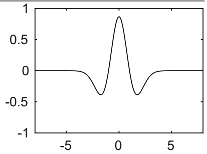
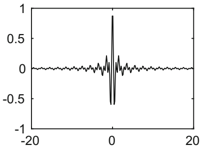
Hence, the residual series is independent, normally distributed with zero mean at 5% level of significance and thus, can be considered as white noise.

9.8 Wavelet Analysis

Time series are represented in the time domain with their amplitude varying with time. This representation is also known as amplitude–time representation. However, often the frequency information is required to extract important information. Mathematical tools like Fourier transform (FT) and wavelet transform (WT) aim to represent the time series in frequency domain so that the information about constituent frequencies are revealed. Whereas both FT and WT are potential in separating the frequencies of a time series (also referred as signal), time information associated with different frequencies can only be revealed by WT. This is the reason of popularity of WT over FT in case of the non-stationary time series with respect to its frequency. In other words, if the constituting frequencies of the time series do not change over time, both FT and WT are equally useful but if it is not, WT is the essential to extract the time information of constituting frequencies.

The WT is a mathematical tool that separates a time series into different constituting components, each corresponding to a particular frequency bands. Separated components are called wavelet components of the original series. The WT utilizes a specific function with zero mean and finite length having unit energy (variance),

Table 9.5 Details of some mother wavelets

Name	Mother wavelet function	Graphical representation
Haar or Daubechies 1	$\Psi(t) = \begin{cases} 1 & 0 \le t \le 0.5 \\ -1 & 0.5 \le t \le 1 \\ 0 & \text{otherwise.} \end{cases}$	
Meyer	In frequency domain $\Psi(\omega) =$ $\begin{cases} \frac{1}{\sqrt{2\pi}} \sin\left(\frac{\pi}{2}\nu\left(\frac{3\lvert\omega\rvert}{2\pi}-1\right)\right)e^{j\omega/2} & \text{if } \frac{2\pi}{3} < \lvert\omega\rvert < \frac{4\pi}{3} \\ \frac{1}{\sqrt{2\pi}} \cos\left(\frac{\pi}{2}\nu\left(\frac{3\lvert\omega\rvert}{2\pi}-1\right)\right)e^{j\omega/2} & \text{if } \frac{4\pi}{3} < \lvert\omega\rvert < \frac{8\pi}{3} \\ 0 & \text{otherwise} \end{cases}$ where $\nu(x) = \begin{cases} 0 & x \le 0 \\ x & 0 < x < 1 \\ 1 & x \ge 1 \end{cases}$	
Morlet	$\Psi(t) = c_\sigma \pi^{(-1/4)} e^{(-1/2)t^2}(e^{i\sigma t} - k_\sigma)$ where $k_\sigma = e^{-1/2\sigma^2}$ $c_\sigma = \left(1 + e^{-\sigma^2} - 2e^{-3/4\sigma^2}\right)^{-1/2}$	
Ricker or Mexican hat	$\Psi(t) = \frac{2}{\sqrt{3\sigma}\pi^{1/4}}\left(1 - \left(\frac{t}{\sigma}\right)^2\right)e^{-\frac{t^2}{2\sigma^2}}$	
Complex Shannon 1–1	$\Psi(t) = \sqrt{F_b}\text{sinc}(F_b x)e^{(2i\pi F_c x)}$ The wavelet is named as $F_b - F_c$ For figure $F_b = F_c = 1$	

known as '*mother wavelet*'. There are several mother wavelets with different mathematical forms, such as Haar, Meyer, Morlet, Mexican Hat. Details of some of these wavelets are provided in Table 9.5. Apart from the mother wavelet function given in Table 9.5, many families of wavelet functions exist like Daubechies, Bi-orthogonal, Gaussian, Shannon.

Any mother wavelet ($\Psi(t)$) can be scaled and/or shifted (known as '*daughter wavelets*'). For a particular mother wavelet, the daughter wavelets ($\Psi_{a,b}(t)$) are mathematically represented as:

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\Psi\left(\frac{t-b}{a}\right) \tag{9.115}$$

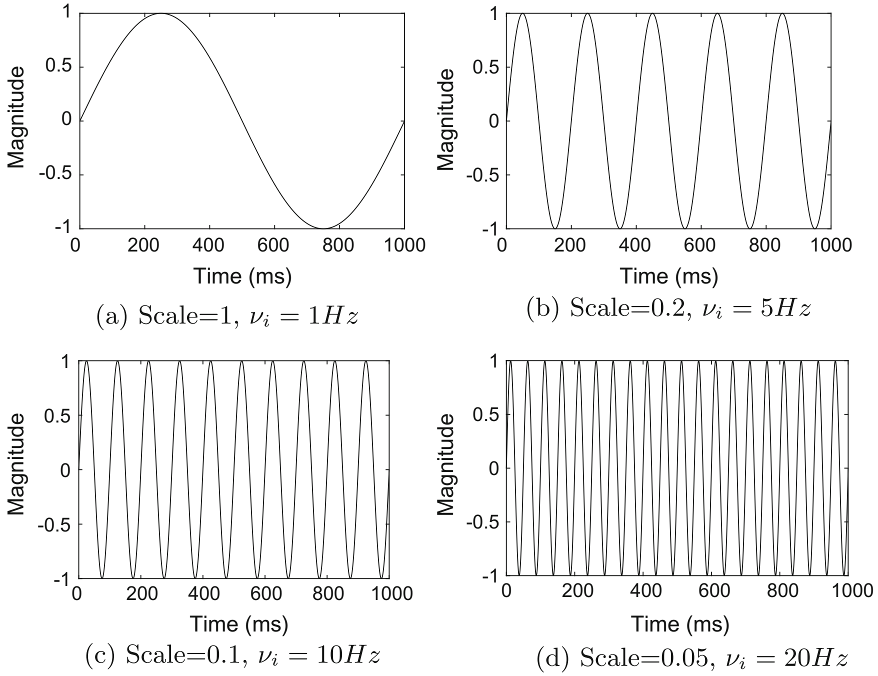
where a and b are shifting and scaling parameters, respectively. The shifting parameter (a) shows the location of wavelet, as the wavelet window is gradually shifted through the time series. Inverse of the scale parameter (b) provides the information of frequency (ν_i). Due to scaling, WT is able to recognize the frequencies in the time series and due to shifting, the WT is able to extract the time varying feature (amplitude) of those frequencies. It should be noted that the scaling as a mathematical operation either dilates or compresses a signal, i.e., larger scales (thus lower frequency) correspond to the dilated (or stretched out) signals and small scales correspond to the compressed signals. For instance, in Fig. 9.7 different scale of sine wave with unit amplitude is shown. It can be observed that from Fig. 9.7a and 9.7d that decrease in scale leads to contraction in signal and vice versa. Hence, by using higher scale, WT extracts the slow moving changes or global information in signal and by using lower scale, WT extracts the detailed information about local disturbances. It should be noted that the WT components at a particular frequency band is obtained by convolution of shifted version of correspondingly scaled daughter wavelet. Depending upon selection of the scaling and shifting parameters and transformation procedure, many WT exist. Three of the most popular WT are discussed here:

(a) **Continuous Wavelet Transform** (CWT): If shifting and scaling parameters are considered to be continuous real number while applying wavelet transform, the WT is called continuous wavelet transform (CWT). The CWT is computed by changing the scale of the analysis window, shifting the window over time, multiplying with the signal, and integrating it over the times. In CWT, the wavelet transform is mathematically expressed as:

$$W_f(a, b) = \frac{1}{\sqrt{C_\Psi}} \int X(t)\Psi^*_{a,b}(t)dt \tag{9.116}$$

where $\Psi^*(t)$ denotes complex conjugate, $C_\Psi = 2\int |F(\Psi(\omega))|^2/\omega d\omega$ and $F(\bullet)$ denote the Fourier transform (Eq. 9.16). If the basis wavelet or mother wavelet ($\Psi(t)$) is orthogonal, then the inverse of wavelet transformation is given by:

$$X(t) = \frac{1}{\sqrt{C_\Psi}} \iint \frac{W_f(a, b)\Psi_{(a,b)}(t)}{a^2}\, da\, db \tag{9.117}$$



(a) Scale=1, $\nu_i = 1Hz$

(b) Scale=0.2, $\nu_i = 5Hz$

(c) Scale=0.1, $\nu_i = 10Hz$

(d) Scale=0.05, $\nu_i = 20Hz$

Fig. 9.7 Different scale/frequencies of unit amplitude sign wave

(b) **Discrete Wavelet Transform** (DWT): Discrete class of wavelets is formed when shifting and scaling parameters are considered discrete instead of continuous variables while applying wavelet transform. If the discrete wavelet is sampled over dyadic space, time grid, the resulting wavelets are called dyadic discrete wavelets. These wavelets are denoted by:

$$\Psi_{j,b}(t) = \frac{1}{\sqrt{2^j}}\Psi\left(\frac{t}{2^j} - b\right) \tag{9.118}$$

The wavelet transform is given by:

$$W_f(a, b) = \frac{1}{\sqrt{C_\Psi}}\sum X(t)\Psi^*_{a,b}(t)dt \tag{9.119}$$

where $\Psi^*(t)$ denotes complex conjugate. Discrete wavelet component is down-sampled or subband coded according to Nyquist–Shannon theorem. The Nyquist–Shannon sampling theorem is a fundamental connection between continuous and discrete representation of time series or signal. This theorem is applicable to any signal having finite range of frequencies or in other words, signal having zero Fourier transform coefficient outside some finite range of frequencies. According to this theorem, if the signal is sampled two times, first with a sampling rate of N_1 at scale a_1, second at a sampling rate of N_2 at scale a_2, then the information contained in these two sampling procedures is equivalent, given

$$N_2 = \frac{a_1}{a_2} N_1 \tag{9.120}$$

As the frequency range of wavelet components (generated by Eq. 9.119) is decreased by half, hence, the components can therefore be subsampled by 2, by discarding every alternate sample or sample falling at even places from beginning. As a result, each of the components has half the length that original time series or signal had. Hence, DWT halves the time resolution, but doubles the frequency resolution. Since, the frequency band of the signal now spans only half the previous frequency band; it effectively reduces the uncertainty in the frequency by half. This procedure is also known as subband coding (or downsampling). Subband coding, however, results in wavelet coefficients depending on their location. As a result, a small change in input signal causes large changes in wavelet coefficients. This is termed as transition-invariance of DWT and is considered a major drawback which limits its application in signal analysis.
It should be noted that a discrete mother wavelet acts as a band-pass filter and scaling it for each level (for dyadic space) effectively halves its bandwidth. This creates the problem that in order to cover the entire spectrum (till the frequency limiting to zero), an infinite number of scaling is required. Hence, to cover the complete spectrum another function associated with the mother wavelet called scaling function or '*father wavelet*' is used. Scaling function is also having finite domain and unit energy. Further, dyadic wavelet functions are orthogonal so the inverse of wavelet transform is given by:

$$X(t) = \frac{1}{\sqrt{C_\Psi}} \sum_{j,k \in \mathbb{Z}} X(t) \Psi_{a,b}(t) \tag{9.121}$$

Alternatively, DWT can also be carried out by using a pair of filters – a high pass and a low pass filter. In DWT, signal convolution with low pass filter followed by dyadic down-sampling gives an approximate coefficient, and one obtained by using high pass filter and dyadic down-sampling is called detailed coefficients. These filters are made using the mother wavelet and scaling function. The DWT filter for Haar mother wavelet is discussed in Sect. 9.8.1.

(c) **Stationary Wavelet Transform**(SWT): Stationary Wavelet Transform (SWT) is specially designed to avoid the transition-invariance of DWT. For avoiding time-invariance, SWT components are not down-sampled (as per Nyquist–Shannon sampling theorem) and the filter coefficients are up-sampled by a factor of $2^{(j-1)}$

> in the jth level of algorithm. Hence, the SWT unlike DWT does not change the time resolution at any stage. But lack of subband coding, results in redundancies in components as SWT components have twice the number of elements needed as per Nyquist–Shannon Theorem. However, SWT reduces the complexity of signal analysis as both input signal and its components have equal length.

Wavelet transform depends upon the selected mother wavelet function. In the next subsection, Haar mother wavelet is discussed in detail to show how the wavelet transformation is carried out using filters.

9.8.1 Haar Wavelet

Haar wavelet is one of the 'square-shaped' wavelet, proposed by Alfréd Haar in 1909. A special case of Daubechies family of mother wavelet functions, the Haar wavelet is considered first member of Daubechies family of wavelet and also regarded as Db1. The Haar mother wavelet function $\mathbb{H}(t)$ is expressed as

$$\mathbb{H}(t) = \begin{cases} 1 & 0 \le t \le 0.5 \\ -1 & 0.5 \le t \le 1 \\ 0 & \text{otherwise.} \end{cases} \tag{9.122}$$

The associated scaling function is given by:

$$\mathbb{S}(t) = \begin{cases} 1 & 0 \le t \le 1 \\ 0 & \text{otherwise.} \end{cases} \tag{9.123}$$

The Haar wavelet and scaling function can be expressed as linear combination of scaling function of different scales.

$$\mathbb{S}(t) = \mathbb{S}(2t) + \mathbb{S}(2t-1) \tag{9.124}$$

$$\mathbb{H}(t) = \mathbb{S}(2t) - \mathbb{S}(2t-1) \tag{9.125}$$

Any continuous real function on [0, 1] can be approximated by linear combinations of dyadic Haar wavelet with different scales and shifts $(1, \mathbb{H}(t+b_1), \mathbb{H}(2t+b_2), \mathbb{H}(4t+b_3), \ldots, \mathbb{H}(2^n t - b_n), \ldots)$. Similarly, any continuous real function with compact support can be approximated by a linear combination of scale functions with different scale and shifts $(\mathbb{S}(t+b_1), \mathbb{S}(2t+b_2), \mathbb{S}(4t+b_3), \ldots, \mathbb{S}(2^n t - b_n), \ldots)$. As stated earlier, the wavelet transform can be carried out using a set of two filter matrices – low pass and high pass filter. High pass filter is formulated on the basis of mother wavelet function and separates the high frequencies from the data. Low pass filter matrix is formulated on the basis of scaling function and allow low-frequency information in data to pass. For stationary wavelet transform of time series of length 'n', Haar wavelet high pass (G) and low pass (H) filters are of ($n \times n$) size. These filters are constructed using following rules:

$$h_{r,c} = \begin{cases} 1/\sqrt{2} & c \in \{r, (r+1) \mod n\} \\ 0 & \text{otherwise.} \end{cases} \tag{9.126}$$

$$g_{r,c} = \begin{cases} (-1)^{r-c}/\sqrt{2} & c \in \{r, (r+1) \mod n\} \\ 0 & \text{otherwise.} \end{cases} \tag{9.127}$$

where $h_{r,c}$ and $g_{r,c}$ are the elements of matrix H and G, respectively, r and c represent the row and column of filter matrix. Here, 'mod' represents a module function. $k \mod n = n$ if $k = n$, otherwise $k \mod n =$ remainder of k divided by n. On closer observation, the low pass filter is 2 term moving average operation and the high pass filter is 1st-order differencing operation normalized with a factor of $1/\sqrt{2}$. When the time series is multiplied with these filters, two components are obtained. The component obtained after multiplication with high pass filter is called detailed SWT component (denoted by d) and component obtained after multiplication with low pass filter is termed approximate SWT component (denote by a). For obtaining DWT components, subband coding of SWT components is done by neglecting every second component value or component values falling on even positions. These components can further be separated into lower frequency bands by applying wavelet transformation on them. This approach of applying wavelet transformation multiple time to get wavelet components at even lower frequency bands is called multiresolution analysis (MRA).

9.8.2 Multiresolution Analysis

Multiresolution analysis or multiresolution wavelet transform (MRWT)can be performed by using low pass filter component (approximate component) as input to wavelet transform at each subsequent level. Hence, MRA helps in analysis of time series or signal at smaller frequency bands. Multiresolution analysis of signal can be carried out with both SWT or DWT. Depending on the wavelet transform used, it is called multiresolution stationary wavelet transform (MRSWT) or multiresolution discrete wavelet transform (MRDWT). By using MRA, a time series $X(t)$ can be represented as

$$X(t) = \sum_{k} a_{0,k} \mathbb{S}_{0,k}(t) + \sum_{j=0}^{\infty} \sum_{k} d_{j,k} \mathbb{H}_{j,k}(t) \tag{9.128}$$

where $\mathbb{S}$ and $\mathbb{H}$ represent scaling function and mother wavelet function, respectively. The subscript pair j, k shows scale and shift parameters of mother wavelet or scaling function. The approximate component ($a_{0,k}$) and detailed component ($d_{j,k}$) are expressed as:

$$a_{0,k} = \sum X(t)\mathbb{S}(t-k) \tag{9.129}$$

$$d_{j,k} = \sum X(t)2^{-j}\mathbb{H}\left(2^{-j}t-k\right) \tag{9.130}$$

If the level of decomposition is L then $a_{0,k}$ series is also represented as a_L and $d_{j,k}$ series are also represented as d_j, where $j \in 1, 2, \ldots L$. In the form of filters, the components a_L and d_j are expressed as:

$$a_L = G_L G_{L-1} \ldots G_1 X \tag{9.131}$$

$$d_j = H_j G_{j-1} G_{j-2} \ldots G_1 X = H_j a_{j-1} \qquad \text{for } j \in \{1, 2, \ldots, L\} \tag{9.132}$$

The low and high pass filters for Haar mother wavelet at any level l are given by:

$$h_{l,r,c} = \begin{cases} 1/\sqrt{2} & c \in \{r, (r + 2^{(l-1)}) \mod n\} \\ 0 & \text{otherwise.} \end{cases} \tag{9.133}$$

$$g_{l,r,c} = \begin{cases} (-1)^{r-c}/\sqrt{2} & c \in \{r, (r + 2^{(l-1)}) \mod n\} \\ 0 & \text{otherwise.} \end{cases} \tag{9.134}$$

where $h_{l,r,c} \in H_l$, $g_{l,r,c} \in G_l$, H_l and G_l are low pass and high pass filters at level l. r and c represent row and column, respectively. It should be noted that for $l = 1$ the above equations are same as Eqs. 9.126 and 9.127. The steps of MRA will be more clear with an example as provided below.

Example 9.8.1
Monthly sea surface temperature (in °C) at a location for last 8 months are 24.8, 23.6, 26.1, 28.4, 24, 22.8, 21.5, and 23. Decompose the time series into its MRSWT components using Haar as mother wavelet upto level 2.

Note: This is just an illustrative problem to facilitate the reader to understand the steps involved in MRSWT. In reality, the length of time series is sufficiently long. However, once the basic steps are understood, the computer codes can be written for longer hydroclimatic data set.

Solution Let the time series of sea surface temperature is represented as matrix X having a size of 8×1. First-level Haar filters (high pass (Eq. 9.127) and low pass (Eq. 9.126)) for time series of length 8 are given by:

$$G_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1\,1\,0\,0\,0\,0\,0\,0 \\ 0\,1\,1\,0\,0\,0\,0\,0 \\ 0\,0\,1\,1\,0\,0\,0\,0 \\ 0\,0\,0\,1\,1\,0\,0\,0 \\ 0\,0\,0\,0\,1\,1\,0\,0 \\ 0\,0\,0\,0\,0\,1\,1\,0 \\ 0\,0\,0\,0\,0\,0\,1\,1 \\ 1\,0\,0\,0\,0\,0\,0\,1 \end{bmatrix}$$

$$H_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Similarly, the second-level Haar filters (high pass (Eq. 9.134) and low pass (Eq. 9.133)) for time series of length 8 are given by :

$$G_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1\,0\,1\,0\,0\,0\,0\,0 \\ 0\,1\,0\,1\,0\,0\,0\,0 \\ 0\,0\,1\,0\,1\,0\,0\,0 \\ 0\,0\,0\,1\,0\,1\,0\,0 \\ 0\,0\,0\,0\,1\,0\,1\,0 \\ 0\,0\,0\,0\,0\,1\,0\,1 \\ 1\,0\,0\,0\,0\,0\,1\,0 \\ 0\,1\,0\,0\,0\,0\,0\,1 \end{bmatrix}$$

$$H_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Hence, the second-level Haar MRSWT components are given by (Eqs. 9.131 and 9.132):

$$\begin{aligned} a_2 &= G_2 G_1 X = \begin{bmatrix} 51.45\ 51.05\ 50.65\ 48.35\ 45.65\ 46.05\ 46.45\ 48.75 \end{bmatrix} \\ d_2 &= H_2 G_1 X = \begin{bmatrix} -3.05\ -1.35\ 3.85\ 4.05\ 1.15\ -1.75\ -1.95\ -0.95 \end{bmatrix} \\ d_1 &= G_1 X = \begin{bmatrix} 0.85\ -1.77\ -1.63\ 3.11\ 0.85\ 0.92\ -1.06\ -1.27 \end{bmatrix} \end{aligned}$$

9.9 MATLAB Examples

MATLAB scripts can be written for solving various examples in this chapter. Following MATLAB built-in functions are helpful.

- Autocorrelation and partial autocorrelation functions for a time series can be calculated by using following two functions:

  ```
  [acf, lags, bounds] = autocorr(y, numLags)
  [pacf, lags, bounds] = parcorr(y, numLags)
  ```

  where `acf` and `pacf` are autocorrelation and partial autocorrelation functions, `lags` and `bounds` are corresponding lag values and 95% confidence intervals. `y` is the time series and `numLags` is number of lags till which the autocorrelation or partial autocorrelation is calculated.
- The skewness of the data can be calculated using the '`skewness`' built-in function.
- For fitting AR, MA, or ARMA model over time series, following functions can be used:

  – `advice(data)`
    This built-in function suggests about the requirement of detrending, suitable model structure, and its order.
  – `m = ar(y,n)`
    The function estimates an AR model on time series `y` with order `n`.
  – `sys = armax(data,[na nb nc nk])`
    This function can be used for estimating AR, MA, ARMA, ARX, MAX, or ARMAX on time series `data` depending upon the second parameter. Different components of the second parameter `[na nb nc nk]` specify different parameters of the generalized ARMAX model.
    `na`: order of autoregressive part
    `nb`: number of term considered from exogenous input.
    `nc`: order of moving average part
    `nk`: lag (if any) in exogenous components.

- For estimation of AR model, Yule–Walker equation can also be solved easily using built-in function '`solve`' or by matrix division operation.
- For MRSWT decomposition the function '`swt`' can be used.

  ```
  SWC = swt(X,N,'wname')
  ```

  where `X` is the one-dimensional matrix, `N` is level of MRSWT, and `'wname'` is name of mother wavelet function. `SWC` is component matrix having `N`+1 rows.

Some of the sample scripts to solve the example are provided in this section. These scripts make use of some of above discussed built-in functions. For instance, the Example 9.4.2 can be solved by script given in Box 9.1. For solving the associated Example 9.4.1, Sects. 7.7 and 7.8 can be referred.

Box 9.1 Test for Trend (Example 9.4.2)

```
clc;clear;close all;
alpha=0.1;
streamflow=[1.10;0.50;2.70;1.30;1.50;2.20;2.10;3;2.90;4.40;4.60;...
    3.10;4.70;4;4.60;5.10;6.10;5.30;6.70;5.60];

data_length=length(streamflow);

%% Mann - Kendall Test
compare_mat=zeros(data_length);
for t=1:data_length
    compare_mat(:,t)=sign(streamflow-streamflow(t));
end

compare_mat=tril(compare_mat);
mann_kenall_stat=sum(sum(compare_mat));
var_man_kendall_stat=data_length*(data_length-1)*(2*data_length+5)/18;
u_c=(mann_kenall_stat-sign(mann_kenall_stat))/sqrt(var_man_kendall_stat);

%%% Display the results
output_file=['output' filesep() 'code_1_result.txt'];
delete(output_file);diary(output_file);diary on;
fprintf('Results for Mann-Kendall Test:\n');
fprintf(' The Mann-Kendall Statistics is %2.0f.\n', mann_kenall_stat);
fprintf(' The test statistics (u_c) is %2.2f.\n',u_c);
if abs(u_c)>norminv(1-alpha/2,0,1)
  fprintf(' As |u_c| > %1.3f (Z_%0.3f), so the null hypothesis of no trend is rejected.\n', ...
        norminv(1-alpha/2,0,1), 1-alpha/2);
else
    fprintf('%s As |u_c| < %1.3f (Z_%0.3f), so the null hypothesis of no trend can not be rejected.\n', ...
        norminv(1-alpha/2,0,1), 1-alpha/2);
end

%% Kendall Tau Test
compare_mat=zeros(data_length);
for t=1:data_length
    compare_mat(:,t)=streamflow>streamflow(t);
end
compare_mat=tril(compare_mat);

p=sum(sum(compare_mat));
tau=4*p/(data_length*(data_length-1))-1;

var_tau=2*(2*data_length+5)/(9*data_length*(data_length-1));
test_stat_z=(tau)/sqrt(var_tau);

%%% Display the results
fprintf('Results for Kendall Tau Test:\n');
fprintf(' The Kendall tau is %2.2f.\n',tau);
fprintf(' The test statistics is %2.2f.\n',test_stat_z);
if abs(test_stat_z)>norminv(1-alpha/2,0,1)
    fprintf(' As |z| > %1.3f (Z_%0.3f), so the null hypothesis of no trend is rejected.\n', ...
        norminv(1-alpha/2,0,1), 1-alpha/2);
else
```

```
    fprintf(' As |z| < %1.3f (Z_%0.3f), so the null hypothesis of no trend can not be rejected.', ...
        norminv(1-alpha/2,0,1), 1-alpha/2);
end
diary off
```

The result for the script given in Box 9.1 is given in Box 9.2. The result matches with inference drawn in Solution 9.4.2, i.e., the null hypothesis of no trend in data is rejected.

Box 9.2 Results for Box 9.1

```
Results for Mann-Kendall Test:
 The Mann-Kendall Statistics is 157.
 The test statistics (u_c) is 5.06.
 As |u_c| > 1.645 (Z_0.950), so the null hypothesis of no trend is rejected.
Results for Kendall Tau Test:
 The Kendall tau is 0.82.
 The test statistics is 5.06.
 As |z| > 1.645 (Z_0.950), so the null hypothesis of no trend is rejected.
```

Sample script for solving Examples 9.6.1, 9.7.3, and 9.7.14 is provided in Box 9.3. In this example, script autocorrelation, partial autocorrelation, and skewness is calculated using MATLAB built-in functions.

Box 9.3 Sample MATLAB script for solving Example 9.6.1 and associated examples

```
clear;clc;close all;

alpha=0.05;
rainfall =[2.89;7.39;23.88;10.59;5.91;1.53;3.48;56.54;26.19;6.35;...
38.09;0.01;3.03;41.57;44.73;21.39;15.87;1.22;21.75;0.21];

y_skewness=skewness(rainfall);

% Calculate Moving average and differencing
[~,m]=movavg(rainfall,1,2);
mov_avg_rainfall=m(2:end);
diff_rainfall=diff(rainfall);
figure('Position',[0 0 500 450]);
plot(1:length(rainfall),rainfall,'k');hold on;
plot((1:length(mov_avg_rainfall))+1,mov_avg_rainfall,'--b','LineWidth',1.5);
plot((1:length(diff_rainfall))+1,diff_rainfall,'-.r','LineWidth',1);
axis([1 20 -50 100]);xlabel('Days','FontSize',14);
ylabel('Magnitude','FontSize',14);
h_l=legend('Original Rainfall Series','Moving Average with window 2',...
'1^{st} order Differencing');set(h_l,'FontSize',11)

% Calculate autocorrelation and partial autocorrelation of rainfall
% till lag 5
[rain_autocorr,autocorr_lags,autocorr_bounds]=autocorr(rainfall,5);
[rain_parcorr,parcorr_lags,parcorr_bounds]=parcorr(rainfall,5);

%%% Display the results
output_file=['output' filesep() 'code_2_result.txt'];
delete(output_file);diary(output_file);diary on;
fprintf(' The skewness is %2.2f.\n',y_skewness);
if abs(y_skewness)>norminv(1-alpha/2,0,1)*sqrt(6/length(rainfall))
  fprintf('As |S| > %1.3f , so the null hypothesis of data being normal is rejected.\n', ...
  norminv(1-alpha/2,0,1)*sqrt(6/length(rainfall)));
else
  fprintf(' As |S| < %1.3f , so the null hypothesis of data being normal can not be rejected.\n', ...
  norminv(1-alpha/2,0,1)*sqrt(6/length(rainfall)));
end
fprintf('\n ACF and PACF function for rainfall upto lag 5 is given by:\n');
fprintf('\n lag\t\t ACF \t\t PACF\n');
for i=1:size(autocorr_lags)
  fprintf('%d \t\t %0.2f \t\t %0.2f\n', autocorr_lags(i),...
          rain_autocorr(i),rain_parcorr(i));
end
fprintf('\n The 95%% confidence interval for ACF and PACF are\n');
fprintf(' ACF  \t (%0.2f,%0.2f)\n PACF \t (%.2f,%0.2f)\n', autocorr_bounds(2),...
    autocorr_bounds(1),parcorr_bounds(2),parcorr_bounds(1));
diary off;
```

The result for the script given in Box 9.3 is given in Box 9.4. The results match with Solution 9.6.1, i.e., according to skewness test the transformed rainfall depth follow normal distribution.

Box 9.4 Results for Box 9.3

```
 The skewness is 0.93.
 As |S| < 1.074 , so the null hypothesis of data being normal can not be rejected.

 ACF and PACF function for rainfall upto lag 5 is given by:

 lag      ACF      PACF
0       1.00       1.00
1       0.03       0.03
2       -0.26      -0.28
3       0.03       0.06
4       -0.20      -0.39
5       -0.01      0.02

 The 95% confidence interval for ACF and PACF are
 ACF      (-0.45,0.45)
 PACF     (-0.46,0.46)
```

Solution of Example 9.7.7 can be obtained by matrix division as shown in Box 9.5.

Box 9.5 Sample MATLAB script for solving Example 9.7.7

```
clear;clc;close all;

flow=[560;630;590;660;580;490;300;350;470;900;850;870;340;...
  560;190;250;380;670;990;840;250;360;1200;950;880;...
  560;450;320;170;580];

flow_auto_corr=autocorr(flow,2);
flow_auto_corr=flow_auto_corr(2:end); % Remove lag 0 ACF

% Solution of Yule Walker Equation
AR_2_params=flow_auto_corr'/[1, flow_auto_corr(1);flow_auto_corr(1),1];

% Display Results
output_file=['output' filesep() 'code_3_result.txt'];
delete(output_file);diary(output_file);diary on;
fprintf('The AR(2) parameters are: %2.3f\t%2.3f\n',...
      AR_2_params(1),AR_2_params(2));
diary off;
```

The results obtained by solution of Yule–Walker equation (Box 9.6) match with Solution 9.7.7.

Box 9.6 Result of script provided in Box 9.5

```
The AR(2) parameters are: 0.543 -0.331
```

The white noise testing of residual (Example 9.7.18) can be carried out by using the script presented in Box 9.7. Three tests are carried out on residual series in the script, namely – test of independence, test for normality, and test for zero mean.

Box 9.7 Sample MATLAB script for solving Example 9.7.18

```
clear;clc;close all;

residual=[1.32;-1.97;-10.88;-5.98;1.83;12.06;3.70;1.55;...
    -2.71;0.61;4.81;-1.27;9.46;-6.10;-1.88;0.260;-9.77;...
    2.83;0.390;0.400;3.97;5.22;4.01;-2.34;-0.230;1.77;...
    -6.28;-5.18;-2.13;0.390];

alpha=0.05;
% Test of independence
N=length(residual);
k=ceil(N/5);
res_autocorr=autocorr(residual,k);
res_autocorr=res_autocorr(2:end);
sq_autocorr=res_autocorr.^2;
weighted_sum_sq_autocorr=0;
for i=1:length(res_autocorr)
weighted_sum_sq_autocorr=weighted_sum_sq_autocorr+...
    sq_autocorr(i)/(N-i);
end
```

```
Q_bar=N*(N+2)*weighted_sum_sq_autocorr;

% Calculate skewness
y=residual;
y_d=y-mean(y);
y_d_squared=y_d.^2;
y_d_cubic=y_d.^3;
y_skewness=mean(y_d_cubic)/(mean(y_d_squared))^1.5;
table_skew_calc=[y,y_d,y_d_squared,y_d_cubic];
table_skew_calc(end+1,:)=sum(table_skew_calc);

% Test for zero mean
T_e=mean(residual)/std(residual);

% Display results

%%% Display the results
output_file=['output' filesep() 'code_4_result.txt'];
delete(output_file);diary(output_file);diary on;
if Q_bar<chi2inv(1-alpha,k-2)
fprintf('As %1.3f < %1.3f , so the null hypothesis of data being independent can not be rejected.\n', ...
  Q_bar, chi2inv(1-alpha,k-2));
else
fprintf('As %1.3f > %1.3f , so the null hypothesis of data being independent is rejected.\n', ...
  Q_bar, chi2inv(1-alpha,k-2));
end
if abs(y_skewness)>norminv(1-alpha/2,0,1)*sqrt(6/length(y))
 fprintf('As |%3.2f| > %1.3f , so the null hypothesis of data being normal is rejected.\n', ...
 y_skewness, norminv(1-alpha/2,0,1)*sqrt(6/length(y)));
else
 fprintf('As |%3.2f| < %1.3f , so the null hypothesis of data being normal can not be rejected.\n', ...
 y_skewness, norminv(1-alpha/2,0,1)*sqrt(6/length(y)));
end

if abs(T_e)<=tinv(1-alpha/2,length(residual)-1)
 fprintf('As |%2.2f| < %1.3f , so the null hypothesis of data having zero mean can not be rejected.\n', ...
 T_e, tinv(1-alpha/2,length(residual)-1));
else
 fprintf('%s As |%2.2f| > %1.3f , so the null hypothesis of data having zero mean is rejected.\n', ...
 T_e, tinv(1-alpha/2,length(residual)-1));
end
diary off
```

The result of white noise test script (Box 9.7) is provided in Box 9.8. Like Solution 9.7.18, the result suggest that residual can be considered white noise as it passes all three tests.

Box 9.8 Result of script provided in Box 9.7

```
As 6.614 < 9.488 , so the null hypothesis of data being independent can not be rejected.
As |0.03| < 0.877 , so the null hypothesis of data being normal can not be rejected.
As |-0.01| < 2.045 , so the null hypothesis of data having zero mean can not be rejected.
```

A sample script for solving Example 9.8.1 using '`swt`' built-in function of MATLAB is presented in Box 9.9.

Box 9.9 Sample MATLAB script for solving Example 9.8.1

```
clear all;close all;clc;

%% Input
X=[24.8, 23.6, 26.1, 28.4, 24, 22.8, 21.5, 23];
comp=swt(X,2,'haar');

%%% Display and save the output
output_file=['output' filesep() 'code_5_result.txt'];
delete(output_file);diary(output_file);diary on;
fprintf('The Haar MRSWT Components are \n');
a_2=[];d_2=[];d_1=[];
for i=1:length(comp)
  a_2=[a_2 sprintf('%2.2f',comp(3,i)) ', '];
  d_2=[d_2 sprintf('%2.2f',comp(2,i)) ', '];
  d_1=[d_1 sprintf('%2.2f',comp(1,i)) ', '];
end
fprintf(' a_2=%s \n d_2=%s \n d_1=%s\n', a_2, d_2, d_1);
diary off
```

The result of the script provided in Box 9.9 is provided in Box 9.10. The result of the script matches with our Solution 9.8.1.

Box 9.10 Result of script provided in Box 9.9

```
The Haar MRSWT Components are
 a_2=51.45, 51.05, 50.65, 48.35, 45.65, 46.05, 46.45, 48.75,
 d_2=-3.05, -1.35, 3.85, 4.05, 1.15, -1.75, -1.95, -0.95,
 d_1=0.85, -1.77, -1.63, 3.11, 0.85, 0.92, -1.06, -1.27,
```

Exercise

9.1 The annual evapotranspiration (in cm/year) for a basin in last 20 years are 61.04, 58.71, 60.02, 60.36, 62.65, 64.17, 62.82, 64.41, 64.6, 63.45, 65.35, 64.65, 67.37, 66.27, 68.39, 66.77, 68.24, 68.04, 66.53, and 68.02.
Check the evapotranspiration data for any trend using (a) Mann–Kendall test and (b) Kendall tau test. Use 5% level of significance.

(Ans. At 5% significance level null hypothesis of no trend is rejected for Mann–Kendall. However, in the Kendall tau test null hypothesis of no trend cannot be rejected at 5% significance level.)

9.2 The monthly average atmospheric pressure (in mb) measured at surface level for 24 consecutive months are
963.65, 965.03, 961.18, 959.43, 957.68, 953.42, 950.11, 952.44, 952.25, 956.88, 963.66, 963.36, 965.56, 964.5, 963.66, 960.91, 956.9, 952.18, 950.71, 952.54, 951.43, 955.06, 959.01, and 962.60. Find the autocorrelation and partial autocorrelation functions at lags 0, 1, 2, and 3.

Ans. Autocorrelation function at lag 0, 1, 2, and 3 are 1, 0.782, 0.414, and 0.008 respectively.
Partial autocorrelation function at lag 0, 1, 2, and 3 are 1, 0.807, −0.617, and −0.481 respectively.

9.3 For the data, provided in Exercise 9.1, find the autocorrelation and partial autocorrelation coefficient at lags 0, 1 and 2. Find the 95% confidence limit for the ACF and PACF at lag 2.

Ans. Autocorrelation function at lag 0, 1, and 2 are 1, 0.784, and 0.678 respectively.
Partial autocorrelation function at lag 0, 1, and 2 are 1, 0.852 and 0.489 respectively.
95% confidence interval of ACF and PACF is [−0.462, 0.462].

9.4 Streamflow at a section for 30 consecutive days is shown in the following table

Day	Flow ($\times$1000 m^3/s)	Day	Flow ($\times$1000 m^3/s)	Day	Flow ($\times$1000 m^3/s)
1	14.12	11	14.45	21	12.50
2	22.05	12	12.72	22	16.10
3	22.34	13	13.67	23	17.40
4	20.07	14	12.58	24	9.48
5	21.15	15	9.33	25	8.41
6	19.82	16	9.67	26	9.33
7	20.65	17	10.65	27	10.40
8	23.57	18	14.47	28	12.62
9	22.19	19	11.02	29	15.30
10	18.32	20	9.82	30	13.84

From historical records, streamflow is found to follow gamma distribution. The rating curve for the section is given by:

$$Q = 59.5(G - 5)^2$$

where Q is streamflow in m^3/s and G is river stage at section in meters. Calculate the river stage at the section and check whether river stage follows normal distribution

at 5% level of significance or not.
(Ans. The river stage for the section (in m) are 20.4, 24.3, 24.4, 23.4, 23.9, 23.3, 23.6, 24.9, 24.3, 22.5, 20.6, 19.6, 20.2, 19.5, 17.5, 17.7, 18.4, 20.6, 18.6, 17.8, 19.5, 21.4, 22.1, 17.6, 16.9, 17.5, 18.2, 19.6, 21.0, and 20.3. At 5% significance level, the null hypothesis of river stage follows normal distribution cannot be rejected.)

9.5 Fit AR(1) and AR(2) model on the river stage data given in Exercise 9.4. What percentage of variance in the river stage time series is explained by these two models? Calculate Akaike Information Criteria for the models and suggest the best model.

Ans. For AR(1) model $\Phi_1 = 0.79$, $R^2 = 0.626$, AIC $= 187.77$
For AR(2) model $\Phi_1 = 0.98$ and $\Phi_2 = -0.24$, $R^2 = 0.647$, AIC $= 189.21$
Hence, out of AR(1) and AR(2), AR(1) is better model.

9.6 Soil moisture is usually found to have high memory component. Using a sensor the surface soil moisture was recorded daily at a location for 60 days. For this time series, PACF at successive lags from 0 to 4 are 1, 0.56, 0.41, 0.15, and 0.11 and corresponding ACF are 1, 0.85, 0.62, 0.25, and 0.12. Suggest the appropriate order of AR model and find the parameters of selected AR model. Check the AR parameters for model stationarity.
(Ans. On the basis of significance of PACF function highest order of AR model is 2. The AR(2) parameters are $\Phi_1 = 1.164$ and $\Phi_2 = -0.370$. The AR(2) model is stationary.)

9.7 For a location, monthly average zonal wind is found to follow a moving average model. From a monthly average zonal wind time series record of length 35, the ACF function at lags 0 to 5 are found as 1, 0.45, 0.35, 0.25, 0.15, and 0.08. Suggest an appropriate order for MA model and find corresponding parameters. Check the invertibility of the selected model.
(Ans. On the basis of significance of ACF, MA(1) is an appropriate model. The parameter for MA(1) model is −0.627. The MA(1) is invertible.)

9.8 The parameters of AR(2) model are $\Phi_1 = 0.77$ and $\Phi_2 = -0.25$. Calculate the ACF till lag 2 for the corresponding time series.
(Ans. $\rho_1 = 0.616$ and $\rho_2 = 0.224$)

9.9 For a MA(2) model fitted on time series $X(t)$, if parameters are $\theta_1 = 0.57$ and $\theta_2 = 0.36$, calculate the PACF and ACF up to lag 2 for the time series $X(t)$.
(Ans. $\rho_1 = -0.251$, $\rho_2 = -0.247$, $\varphi_1 = -0.251$, and $\varphi_2 = -0.331$)

9.10 Considering the following ARMA model,

$$X(t) = 0.63X(t-1) - 0.45X(t-2) + \varepsilon(t) - 0.58\varepsilon(t-1) + 0.21\varepsilon(t-2)$$

Check the invertibility and stationarity of the model.
(Ans. The model is stationary but not invertible.)

9.11 At a location, the daily air temperature follows the ARMA(2,1) model given below,

$$X(t) = 0.7X(t-1) + 0.2X(t-2) + \varepsilon(t) + 0.7\varepsilon(t-1)$$

If the air temperature recorded in the last week (in °C) was 16.5, 15.2, 18.2, 16.3, 19.4, 17.8, and 15.7, then forecast air temperature and their 95% confidence limit for next three days. Assume that the variance of residual is unity. Further, update the forecast for remaining two days, if the temperature on eighth day is recorded as 14.5°C.
(Ans. Forecasted temperatures (in °C) for next three days are 15.7, 14.1, and 13.0 respectively. Their confidence intervals are (13.7, 17.7), (10.8, 17.5), and (9.0, 17.1) respectively. The update forecasts for next two days (in °C) are 12.4 and 11.6, respectively.)

9.12 For the monthly average atmospheric pressure at surface data provided in Exercise 9.2, check the data for any seasonality (periodicity of 12 months) at 5% level of significance.
(Ans. Data is seasonal at 5% level of significance.)

9.13 For AR(2) model developed in Exercise 9.5, check that the residual series is white noise at 5% level of significance. A series is called white noise when it is independent and normally distributed with zero mean.
(Ans. The residual is white noise at 5% level of significance.)

9.14 Decompose the annual evapotranspiration time series provided in Exercise 9.1 into its Haar MRSWT components up to level 2. [Hint: Code written in Box 1.9 may be used]
Ans. The decomposed series are

a_2 [120.1, 120.9, 123.6, 125.0, 127.0, 128.0, 127.6, 128.9, 129.0, 130.4, 131.8, 133.3, 134.4, 134.8, 135.7, 134.8, 135.4, 131.8, 127.2, 123.9]

d_2 [−0.3, −2.1, −3.2, −2.0, −0.2, −1.0, −0.4, 0.1, −1.0, −1.6, −1.8, −1.3, −0.8, −0.2, −0.6, 0.2, 0.9, 2.8, 7.4, 5.2]

d_1 [1.6, −0.9, −0.2, −1.6, −1.1, 1.0, −1.1, −0.1, 0.8, −1.3, 0.5, −1.9, 0.8, −1.5, 1.1, −1.0, 0.1, 1.1, −1.1, 4.9]

Chapter 10
Theory of Copula in Hydrology and Hydroclimatology

This chapter deals with an introduction to copula theory and its applications in hydrology and hydroclimatology. The copula theory is relatively new to this field but has already established itself to be highly potential in frequency analysis, multivariate modeling, simulation and prediction. Development of joint distribution between multiple variables is the key to analyze utilizing the potential of copulas. The chapter starts with the mathematical theory of copulas and gradually move on to the application. If the readers are already aware of the background theory and look for application of copula theory, they can directly proceed to Sect. 10.8. *Basic mathematical formulations for most commonly used copulas are discussed, and illustrative examples are provided. It will enable the readers to carry out applications to other problems. All the illustrative examples are designed with very few data points. This helps to show the calculation steps explicitly. Please note that any statistical analysis should be done with sufficiently long data. Once the readers understand the steps, computer codes can be written easily for large data sets. Example of MATLAB codes is also provided at the end.*

10.1 Introduction

Theory of copula itself may need an entire book (Nelsen 1999). Focus of this chapter is to introduce this theory for hydrologic and hydroclimatologic applications. The word *copula* originates from a Latin word 'copulare,' which means 'to join together.' In many cases of statistical modeling, it is essential to obtain the joint probability distribution function between two or more random variables. Even though the marginal distributions of each of the random variables are known, their joint distributions may not be easy to derive from these marginal distributions. However, copula can be used to obtain their joint distribution, if the information on scale-free measures of dependence between random variables is available.

R. Maity, *Statistical Methods in Hydrology and Hydroclimatology*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-10-8779-0_10

10.2 Preliminary Concepts

10.2.1 Definition of Copula

Let X and Y be a pair of random variables with cumulative distribution function (CDF) $F_X(x)$ and $F_Y(y)$, respectively. Also, let their joint CDF be $H_{X,Y}(x, y)$. Hence, each pair (x, y) of real numbers leads to a point $(F_X(x), F_Y(y))$, in the unit square, i.e., I^2 or $[0, 1] \times [0, 1]$. The ordered pair in turn corresponds to a number $H_{X,Y}(x, y)$, in $[0, 1]$. This correspondence is a function, which is known as ***copula***. Thus, copula (designated by C) is a function that joins or couples one-dimensional marginal distributions of multiple random variables to their joint distribution function.

It is worthwhile to note here that such correspondence is irrespective of the marginal distributions of the random variables. In other words, any form of marginal distributions can be coupled to get their joint distribution, which is the reason for the popularity of copula theory in many areas of research. Moreover, the theory of copula can be extended to higher dimensions also.

10.2.2 Graphical Representation of Copula

An n-dimensional copula is represented in a I^n dimensional space. Graphically, only two-dimensional copula can be shown as a surface in I^3 space or as a contour plot in I^2. For example, a two-dimensional independent copula, given by $C(u, v) = uv$ for $0 \leq u, v \leq 1$, is graphically represented in Fig. 10.1. This copula function is called independent copula, as it defines the joint distribution for two independent random variables.

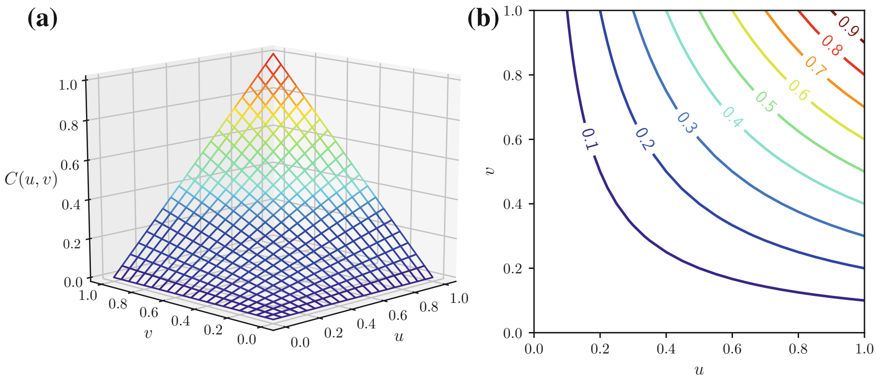


Fig. 10.1 Independent copula function represented as **a** three-dimensional surface plot, and **b** contour plot

10.3 Sklar's Theorem

Application of copula to probability and statistics is achieved through Sklar's theorem (Sklar 1959). It states that if $H_{X,Y}(x, y)$ is a joint distribution function, then there exists a copula $C(u, v)$, such that, for all $x, y \in \mathbb{R}$

$$H_{X,Y}(x, y) = C(F_X(x), F_Y(y)) \tag{10.1}$$

where $F_X(x)$ and $F_Y(y)$ are the marginal distribution of X and Y, respectively. Hence, if $F_X(x)$ and $F_Y(y)$ are continuous, then copula function is unique, otherwise copula is uniquely determined on $\mathrm{Ran}(F_X(x)) \times \mathrm{Ran}(F_Y(y))$, where $\mathrm{Ran}(\bullet)$ represents range for $\bullet$. Sklar's theorem is used for coupling two marginal distributions to obtain their joint distribution.

Equation 10.1 can be inverted as,

$$C'(u, v) = H(F_X^{(-1)}(u), F_Y^{(-1)}(v)) \tag{10.2}$$

where $F_X^{(-1)}(u)$ and $F_Y^{(-1)}(v)$ are called '*quasi-inverse*' of $F_X(x)$ and $F_Y(y)$, respectively. If any marginal ($F_X(x)$ or $F_Y(y)$) is strictly increasing function, then the *quasi-inverse* is same as its inverse (denoted by F_X^{-1} or F_Y^{-1}). However, if any marginal distribution $F_X(x)$ is not strictly increasing, then the *quasi-inverse* is given by,

(a) For any $t \in \mathrm{Ran}(F_X(x))$, then $F_X^{(-1)}(t)$ is any number $x \in \mathbb{R}$ such that $F_X(x) = t$, i.e., for all t in $\mathrm{Ran}(F_X(x))$,

$$F_X\left(F_X^{(-1)}(t)\right) = t \tag{10.3}$$

(b) If t is not in $\mathrm{Ran}(F_X(x))$, then,

$$F_X^{(-1)}(t) = \inf\{x | F_X(x) \geq t\} = \sup\{x | F_X(x) \leq t\} \tag{10.4}$$

where inf stands for infimum and sup stands for supremum. In Eq. 10.2, if $F_X(x)$ and $F_Y(y)$ are continuous function, then $C'(u, v)$ is a valid copula function. Hence, copula can be constructed using the information of marginal distributions and joint distribution. This method of constructing copula is called inversion technique to construct copula.

Example 10.3.1
If a joint distribution function is given by,

$$H_{X,Y}(x, y) = \begin{cases} \frac{(x+1)(e^y-1)}{x+2e^y-1} & (x, y) \in [-1, 1] \times [0, \infty), \\ 1 - e^{-y} & (x, y) \in [1, \infty) \times [0, \infty), \\ 0 & \text{elsewhere.} \end{cases}$$

with the marginal distribution functions as,

$$F_X(x) = \begin{cases} 0 & x < -1, \\ (x+1)/2 & x \in [-1, 1], \\ 1 & x > 1. \end{cases} \quad \text{and} \quad F_Y(y) = \begin{cases} 0 & y < 0, \\ 1 - e^{-y} & y \geq 0. \end{cases}$$

Find the corresponding copula function.

Solution If the reduced variate of X is $u \in [0, 1]$ (i.e., $F_X(x) = u$ or $F_X^{(-1)}(u) = x$), then the *quasi-inverse* of $F_X(x)$ is given by,

$$u = \frac{F_X^{(-1)}(u) + 1}{2} \quad or \quad F_X^{(-1)}(u) = 2u - 1$$

Similarly, if v is the reduced variate of Y, then the *quasi-inverse* of $F_Y(y)$ is given by,

$$F_Y^{(-1)}(v) = -\ln(1 - v)$$

Both $F_X(x)$ and $F_Y(y)$ are both continuous functions with range I. Hence, the corresponding copula C is given by (Eq. 10.2),

$$C = H_{X,Y}(F_X^{(-1)}(u), F_Y^{(-1)}(v)) = \frac{(2u - 1 + 1)(e^{-\ln(1-v)} - 1)}{2u - 1 + 2e^{-\ln(1-v)} - 1} = \frac{uv}{u + v - uv}$$

10.4 Basic Properties of a Copula Function

Before discussing the basic properties of a copula function, some basic terminologies are discussed in the following section.

10.4.1 Basic Terminologies

H − volume

Let us assume that S_1 and S_2 are two non-empty subsets of set of real numbers, $\mathbb{R}$, and $H(x, y)$ is a function defined on $S_1 \times S_2$. For any two points (x_1, y_1) and (x_2, y_2) in $S_1 \times S_2$, the corresponding rectangle B can be defined as $[x_1, x_2] \times [y_1, y_2]$. The H − volume for rectangle B is the volume enclosed by H function and XY plane. Mathematically, H − volume for rectangle B is expressed as

$$V_H(B) = H(x_2, y_2) - H(x_1, y_2) - H(x_2, y_1) + H(x_1, y_1) \qquad (10.5)$$

$H-$ volume can also be expressed as *second-order differencing* of H on B.

$$V_H(B) = \Delta_{y_1}^{y_2}\Delta_{x_1}^{x_2}H(x, y) \tag{10.6}$$

where $\Delta_{y_1}^{y_2}H(x, y)$ represents first-order differencing of function $H(x, y)$ keeping X constant, i.e., $\Delta_{y_1}^{y_2}H(x, y) = H(x, y_2) - H(x, y_1)$. Similarly, $\Delta_{x_1}^{x_2}H(x, y) = H(x_2, y) - H(x_1, y)$.

2-Increasing Function

The concept of 2-increasing function in two-dimensional case is analogous to non-decreasing functions in one dimension. A two-dimensional real function is 2-increasing, if H-volume of any rectangle B is nonnegative, i.e., $V_H(B) \geq 0$ for all rectangles B whose vertices lie in $Dom(H)$. This is graphically represented in Fig. 10.2.

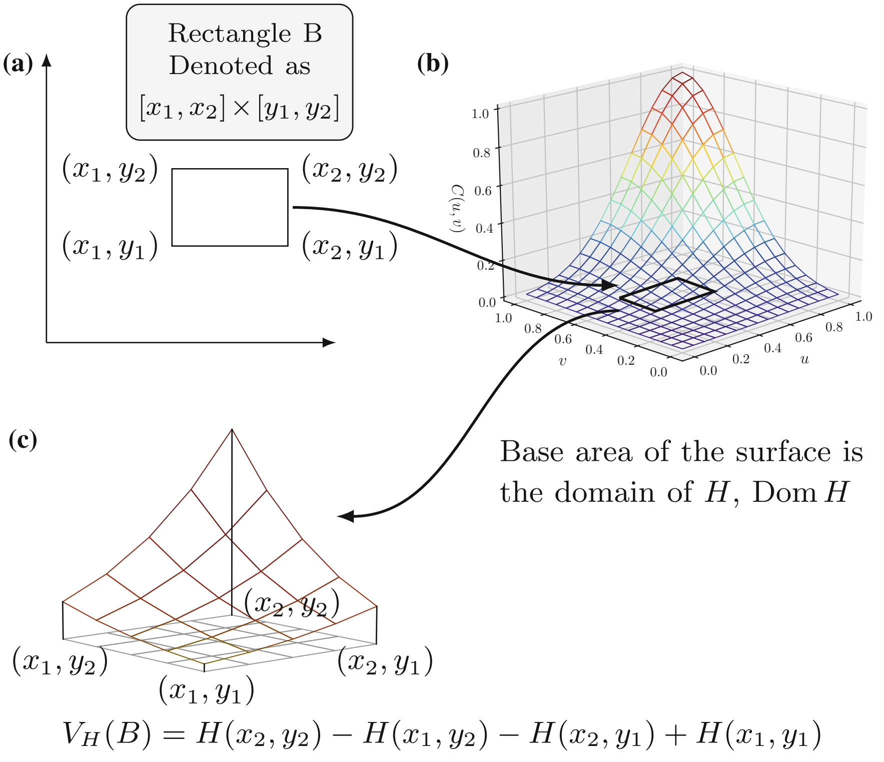


Fig. 10.2 **a** Rectangle B denoted as $[x_1, y_1] \times [x_2, y_2]$; **b** Joint *CDF* $H_{X,Y}(x, y)$; **c** $H-$volume of B denoted as $V_H(B)$

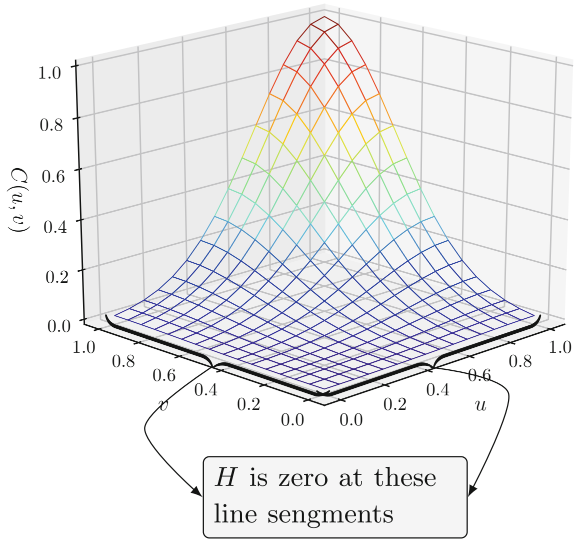


Fig. 10.3 Pictorial representation of a two-dimensional grounded function

Grounded Function

A two-dimensional function $G(x, y)$ having domain $S_1 \times S_2$ is called grounded, if $G(a_1, y) = 0 = G(x, a_2)$ for all $(x, y) \in \mathrm{Dom}(G(x, y))$, where, a_1 and a_2 are the least elements of S_1 and S_2, respectively. Copula functions are required to be grounded. As domain of copula function is I^2, so $a_1 = a_2 = 0$. The copula function shown in Figs. 10.1 and 10.3 is grounded as the value of copula function at u and v axes is zero, i.e., $C(0, v) = C(u, 0) = 0$.

Properties of Copula Function ($C(u, v)$)

A copula having a domain of I^2 has following properties:

(i) For every u_1, u_2, v_1, v_2 in I, if $u_1 \leq u_2$ and $v_1 \leq v_2$, then

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0 \tag{10.7}$$

This property indicates that the copula functions are 2-increasing.

(ii) For every u, v in I

$$C(u, 0) = C(0, v) = 0 \tag{10.8}$$

This property indicates that the copula functions are grounded.

(iii) For every u, v in I

$$C(u, 1) = u \tag{10.9}$$

$$C(1, v) = v \tag{10.10}$$

Example 10.4.1

Check whether the following functions of $u, v \in \text{I}$ can be considered a valid copula function or not.

(a) $C(u, v) = uv$

(b) $C(u, v) = 1$

(c) $C(u, v) = \max(u + v - 1, 0)$

(d) $C(u, v) = \frac{u^2+v^2}{2}$

(e) $C(u, v) = \left[\max\left(u^{-1} + v^{-1} - 1, 0\right)\right]^{-1}$

(f) $C(u, v) = |u + v - 2|$

Solution A valid copula function should follow all the properties listed in Sect. 10.4.1.

(a) For the function $C(u, v) = uv$, for all $u_1, u_2, v_1, v_2 \in \text{I}$, such that $u_1 \le u_2$ and $v_1 \le v_2$,

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) = u_2v_2 - u_2v_1 - u_1v_2 + u_1v_1$$

Assuming $u_2 = u_1 + d$, then $u_2v_2 + u_1v_1 = (u_1 + d)v_2 + (u_2 - d)v_1 = u_1v_2 + u_2v_1 + d(v_2 - v_1)$. Hence, the above expression reduces to $d(v_2 - v_1)$. As $d \ge 0$ and $(v_2 - v_1) \ge 0$, so

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \ge 0$$

Then, for copula $C(u, v) = uv$,

$$C(u, 0) = u \times 0 = 0$$

$$C(0, v) = 0 \times v = 0$$

and

$$C(u, 1) = u \text{ and } C(1, v) = v$$

Hence, the function given by $C(u, v) = uv$ is both grounded, 2-increasing in I. $C(u, v) = uv$ is a valid copula function.

(b) For the function $C(u, v) = 1$, for any $u_1, u_2, v_1, v_2 \in \text{I}$

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) = 1 - 1 - 1 + 1 = 0$$

However, the function is not grounded on u and v axes as $C(u, 0) = C(0, v) = 1 \neq 0$. Moreover, $C(u, 1) = 1 \neq u$ and $C(1, v) = 1 \neq v$, and hence, the function given by $C(u, v) = 1$ is not a valid copula function. Similarly, any constant function cannot be a copula function.

(c) The function given by $C(u, v) = [\max(u + v - 1, 0)]$ should be 2-increasing in I. For all $u_1, u_2, v_1, v_2 \in$ I, such that $u_1 \leq u_2$ and $v_1 \leq v_2$, the function (by definition) is lower bounded by 0 and upper bounded by 1.

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$$
$$C(u_2, v_2) + C(u_1, v_1) \geq C(u_2, v_1) + C(u_1, v_2)$$

By definition of $C(u, v)$, no negative value of $C(u, v)$ is possible, so there are following four possibilities

(i) $u_2 + v_2 \leq 1$, then $C(u_2, v_2) = C(u_1, v_1) = C(u_2, v_1) = C(u_1, v_2) = 0$ and hence, $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) = 0$

(ii) $u_2+v_1 \geq 1$ and $u_1+v_2 \leq 1$, then $C(u_2, v_2) = u_2+v_2-1 \geq 0$, $C(u_2, v_1) = u_2 + v_1 - 1 \geq 0$ and $C(u_1, v_1) = C(u_1, v_2) = 0$ and hence, $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) = u_2 + v_2 - 1 - u_2 - v_1 + 1 = v_2 - v_1 \geq 0$

(iii) $u_1+v_2 \geq 1$ and $u_2+v_1 \leq 1$, then $C(u_2, v_2) = u_2+v_2-1 \geq 0$, $C(u_1, v_2) = u_1 + v_2 - 1 \geq 0$ and $C(u_1, v_1) = C(u_2, v_1) = 0$ and hence, $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) = u_2 + v_2 - 1 - u_1 - v_2 + 1 = u_2 - u_1 \geq 0$

(iv) $u_1 + v_1 \leq 1$ then $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) = u_2 + v_2 - 1 - u_2 - v_1 + 1 - u_1 - v_2 + 1 + u_1 + v_1 - 1 = 0$

Hence, the function given by $C(u, v) = \max(u + v - 1, 0)$ is 2-increasing. The function is also grounded as $C(u, 0) = C(0, v) = 0$. Further, $C(u, 1) = \max(u + 1 - 1, 0) = u$, and similarly, $C(1, v) = v$. Hence, the $C(u, v)$ is a valid copula function.

(d) The function given by $C(u, v) = (u^2 + v^2)/2$ is not grounded as $C(u, 0) = u^2/2 \neq 0$ for all $u \in$ I, and hence, it cannot be a valid copula function. It should be noted that violation of even single property listed in Sect. 10.4.1 is enough to declare the function unfit for being a copula function.

(e) The function $C(u, v) = \left[\max\left(u^{-1} + v^{-1} - 1, 0\right)\right]^{-1}$ can be proved to be 2-increasing using the different cases as done in Example 10.4.1d above. The function is grounded as $C(u, 0) = C(0, v) = 0$. Further, $C(u, 1) = u$ and $C(1, v) = v$, hence, the function $C(u, v) = \left[\max\left(u^{-1} + v^{-1} - 1, 0\right)\right]^{-1}$ is a valid copula function. It should be noted that this copula function and function discussed in Example 10.4.1d above are derived from same class of copula known as Clayton copula. The details of Clayton copula are discussed in Table 10.1.

(f) The function $C(u, v) = |u + v - 2|$ is not grounded as $C(u, 0) = |u - 2|$, and hence, it is not a valid copula function.

Frechet–Hoeffding Bounds

Let C be a copula, then for every (u, v) in $Dom(C)$,

$$\max(u + v - 1, 0) \leq C(u, v) \leq \min(u, v) \tag{10.11}$$

$$\text{or, } W(u, v) \leq C(u, v) \leq M(u, v) \tag{10.12}$$

where $W(u, v) = \max(u + v - 1, 0)$ and $M(u, v) = \min(u, v)$. The functions $W(u, v)$ and $M(u, v)$ are called lower and upper Frechet–Hoeffding bounds. The graphical representation for these bounds is given in Figs. 10.4 and 10.5, respectively. As copula couples two marginal distributions to obtain joint distribution, hence,

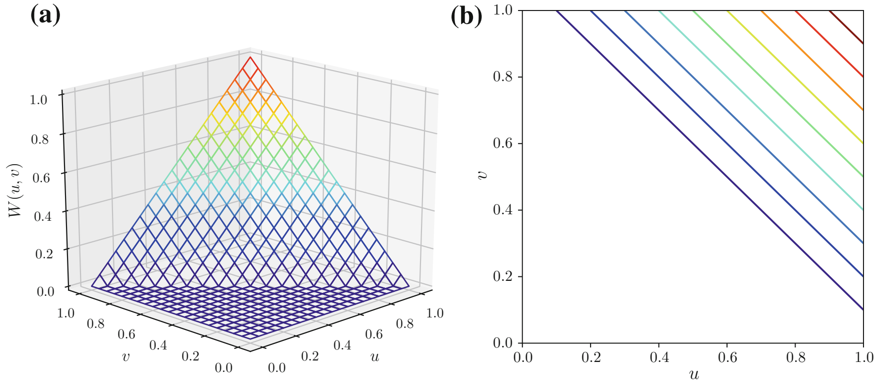


Fig. 10.4 Graphical representation of $W(u, v)$ – **a** 3-d representation **b** Contour Plot

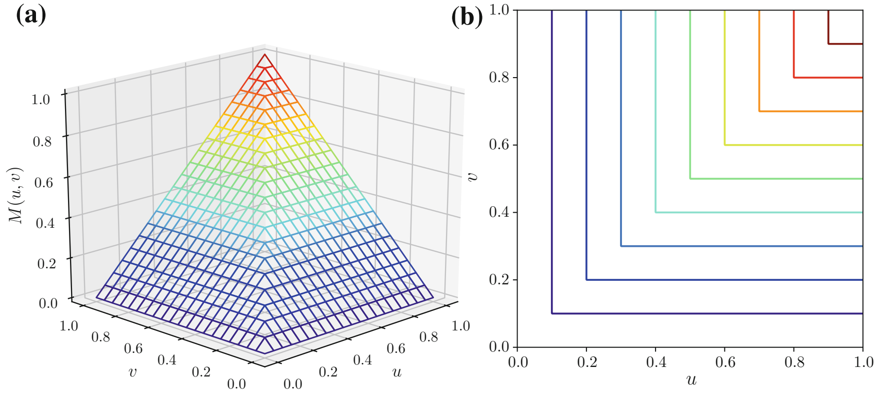


Fig. 10.5 Graphical representation of $M(u, v)$ – **a** 3-d representation **b** Contour Plot

Frechet–Hoeffding bounds also apply as bounds to joint probability distribution. Interestingly, for any $t \in \mathrm{I}$, these bounds define the region in I^2, for which copula/joint distribution function can have a value equal to t.

10.5 Nonparametric Measures of Association

Correlation coefficient is used to quantify the linear association between two random variables. However, if the variables are not linearly associated, then correlation coefficient will be low despite the existence of association (nonlinear) between variables. For example, the data pairs (x, y), where, $y = mx + \varepsilon$ will show high correlation coefficient (given ε is iid with zero mean). However, if $z = \ln(y)$ then y and z will have correlation coefficient close to zero, though they are associated. Hence, other measures of associations are needed that can suggest the existence of association between two variables irrespective of the nature of their interrelationship. One way to achieve this is by measuring the association between the ranks of the variable, instead of their values. The nonparametric or scale-free measures of association make use of the ranks of the variables, rather than their values (as done in correlation coefficient). Two such nonparametric measures of association are as follows:

(a) **Kendall rank correlation coefficient or Kendall's Tau** (τ)
Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be the paired sample of two random variables, X and Y. The ordered pair (x_k, y_k) can be transformed into their respective ranks (R_k^x, R_k^y). Two pairs (x_i, y_i) and (x_j, y_j) are known to be concordant if $(x_i - x_j)(y_i - y_j) > 0$ or $(R_i^x - R_j^x)(R_i^y - R_j^y) > 0$ and discordant if $(x_i - x_j)(y_i - y_j) < 0$ or $(R_i^x - R_j^x)(R_i^y - R_j^y) < 0$. Sample estimate of Kendall's tau is obtained as the difference between the probability of concordance and the probability of discordance. Out of n paired samples, there are ${}^nC_2$ different ways to select two pairs. If there are c number of concordant pairs and d number of discordant pairs, sample estimate of Kendall's tau is expressed as:

$$\hat{\tau} = P[(x_i - x_j)(y_i - y_j) > 0] - P[(x_i - x_j)(y_i - y_j) < 0] = \frac{c - d}{{}^nC_2} \tag{10.13}$$

The Kendall's tau is expected to follow normal distribution with mean 0 and variance $2(2n+5)\big/9n(n-1)$ under the assumption that X and Y are independent. This information can be utilized to check the significance of Kendall's tau. The Kendall's Tau (τ) for u, v is related to copula function C, as given in the following expression:

$$\tau = 4 \int C(u, v)\, dC(u, v) - 1 \tag{10.14}$$

(b) **Spearman's rank correlation coefficient or Spearman's rho** (ρ_s)
Spearman's rank correlation coefficient (ρ_s) is analogous to correlation coef-

ficient and calculated in similar way with one difference that instead of using the values of variables, their ranks are used. For a sample (x_i, y_i) of size n, to compute ρ_s the samples are first transformed to their respected ranks (R_i^x, R_i^y). Spearman's rho (ρ_s) is expressed as:

$$\rho_s = \frac{\sum_{i=1}^{n}(R_i^x - \overline{R}^x)(R_i^y - \overline{R}^y)}{\sqrt{\sum_{i=1}^{n}(R_i^x - \overline{R}^x)^2 \sum_{i=1}^{n}(R_i^y - \overline{R}^y)^2}} \tag{10.15}$$

If there is tie between two or more observations, an average of tie ranks is assigned to all those ties. If all the ranks are distinct, ρ_s can be computed as

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i}{n(n^2 - 1)} \tag{10.16}$$

where $d_i = (R_i^x - R_i^y)$ and n is the number of data. R_i^x and R_i^y stand for rank of x_i and y_i in X and Y, respectively. For large samples, under the assumption that X and Y are independent, Spearman's rho follows normal distribution with mean 0 and variance $1/(n-1)$, like correlation coefficient. Hence, the confidence interval of Spearman's rho can be calculated in similar fashion as correlation coefficient. The Spearman's rho between u and v is related to the copula function, as given in the following expression

$$\rho_s = 12 \iint uv\, dC(u, v) - 3 \tag{10.17}$$

Equations 10.14 and 10.17 link the scale-free measures of association with copula functions and hence can be used to derive relationship between copula parameter and the scale-free measures of association.

Example 10.5.1
The daily temperature is measured for two towns (A and B). Following paired observations are obtained: (18.1, 23.3), (22.3, 26.0), (18.7, 25.5), (17.5, 30.0), and (24.5, 28.2). Calculate following measure of association for the ordered series.

(a) Kendall's Tau (τ)
(b) Spearman's Rho (ρ_s)

Note For drawing meaningful statistical inference, the data length should be sufficiently large, which is not the case in this example. This example only illustrates the procedure for calculating the scale-free measures of association. In the real world, the data set can never be this small; however, the methodology for calculating the measures of association does not change and can easily be programmed using concepts from this example.

Solution Let us assume that random variables X and Y represent the temperature for town A and B, respectively. The given data set can be arranged in increasing order by considering the order of X as follows:

X	17.5	18.1	18.7	22.3	24.5
Y	30.0	23.3	25.5	26.0	28.2

The corresponding ranks are given by:

R^x	1	2	3	4	5
R^y	5	1	2	3	4

(a) Calculation of Kendall's Tau τ
For the first pair in table (i.e., (17.5, 30.0)), all other pairs are discordant. Similarly, for second observation, there are 3 concordant pair. Similarly, concordant and discordant pairs can be counted for each pair. The total number of concordant pairs is $3 + 2 + 1 = 6$, and total number of discordant pairs is 4. Kendall's Tau is given by

$$\tau = \frac{2(c-d)}{n(n-1)} = \frac{2(6-4)}{5 \times 4} = 0.2$$

(b) Calculation of Spearman's Rho (ρ_s)
Here, $\overline{R^x} = \frac{(1+2+3+4+5)}{5} = 3 = \overline{R^y}$. Hence,

$$\sum_{i=1}^{n}(R_i^x - \overline{R^x})^2 = (-2)^2 + (-1)^2 + 2^2 + 1^2 = 10 = \sum_{i=1}^{n}(R_i^y - \overline{R^y})^2$$

$$\sum_{i=1}^{n}(R_i^x - \overline{R^x})(R_i^y - \overline{R^y}) = (-2) \times 2 + (-1) \times (-2) + 0 + 0 + 2 \times 1 = 0$$

$$\rho_s = \frac{\sum_{i=1}^{n}(R_i^x - \overline{R^x})(R_i^y - \overline{R^y})}{\sqrt{\sum_{i=1}^{n}(R_i^x - \overline{R^x})^2 \sum_{i=1}^{n}(R_i^y - \overline{R^y})^2}} = 0$$

Example 10.5.2
The monthly anomaly of outgoing long-wave radiation (OLR; in W/m^2) and monthly precipitation (in cm) for last 6 months are recorded as (18, 1), (−15, 10), (14, 1), (−2, 7), (−23, 12), and (1, 2). Calculate the following measures of association for the ordered series.

(a) Kendall's Tau (τ)
(b) Spearman's Rho (ρ_s)

Solution Let us assume that random variables X and Y represent the OLR and precipitation, respectively. The given data set can be arranged in increasing order by considering the order of X as follows:

X	−23	−15	−2	1	14	18
Y	12	10	7	2	1	1

The corresponding rank in increasing order is given by:

R^x	1	2	3	4	5	6
R^y	5	4	3	2	1.5	1.5

(a) Calculation of Kendall's Tau τ
For the first pair in table (i.e., (−23, 12)), no pair is concordant and five pairs are discordant. Similarly, for second observation, there are 4 discordant pairs. Concordant and discordant pairs can be counted for each pair. Hence, the total number of concordant pairs is 0 and number of discordant pair is $5+4+3+2=14$. Hence, Kendall's Tau is given by

$$\tau = \frac{2(c-d)}{n(n-1)} = \frac{2(0-14)}{6\times 5} = -0.93$$

(b) Calculation of Spearman's Rho (ρ_s)
Here, $\overline{R}^x = \frac{(1+2+3+4+5+6)}{6} = 3.5$. Similarly, $\overline{R}^y = 2.67$. Hence,

$$\begin{aligned}
\textstyle\sum_{i=1}^{n}(R_i^x - \overline{R}^x)(R_i^y - \overline{R}^y) &= (-2.5)\times 2.33 + (-1.5)\times 1.33 + (-0.5)\times 0.33 \\
&\quad +0.5\times(-0.67) + 1.5\times(-1.67) + 2.5\times(-1.67) \\
&= -15 \\
\textstyle\sum_{i=1}^{n}(R_i^x - \overline{R}^x)^2 &= (-2.5)^2 + (-1.5)^2 + (-0.5)^2 + (0.5)^2 + 1.5^2 + 2.5^2 \\
&= 17.5 \\
\textstyle\sum_{i=1}^{n}(R_i^y - \overline{R}^y)^2 &= 2.33^2 + 1.33^2 + 0.33^2 + (-0.67)^2 + 2\times(-1.67)^2 \\
&= 13.33
\end{aligned}$$

Hence,

$$\rho_s = \frac{\sum_{i=1}^{n}(R_i^x - \overline{R}^x)(R_i^y - \overline{R}^y)}{\sqrt{\sum_{i=1}^{n}(R_i^x - \overline{R}^x)^2 \sum_{i=1}^{n}(R_i^y - \overline{R}^y)^2}} = \frac{-15}{\sqrt{17.5\times 13.33}} = -0.98$$

10.6 Copula and Function of Random Variables

Function/transformation of random variable may change their nature of association. From Eqs. 10.14 and 10.17, it is evident that copula is dependent upon the rank correlation structure of the random variables. Hence, for any transformation of random variable that does not change the rank dependence structure, copula function does not change.

Let us assume two random variables X and Y with reduced variates u and v, respectively, having a copula function $C_{X,Y}$ associated with them. Further, if α and β are strictly monotonic function on $\text{Ran}(X)$ and $\text{Ran}(Y)$ that transform X and Y, respectively, then

(i) If α and β are both strictly increasing (hence, they preserve the rank dependence structure)

$$C_{\alpha(X)\beta(Y)}(u, v) = C_{XY}(u, v) \tag{10.18}$$

(ii) If α is strictly decreasing and β is strictly increasing

$$C_{\alpha(X)\beta(Y)}(u, v) = v - C_{XY}(1 - u, v) \tag{10.19}$$

(iii) If α is strictly increasing and β is strictly decreasing

$$C_{\alpha(X)\beta(Y)}(u, v) = u - C_{XY}(u, 1 - v) \tag{10.20}$$

(iv) If α and β are both strictly decreasing

$$C_{\alpha(X)\beta(Y)}(u, v) = u + v - 1 + C_{XY}(1 - u, 1 - v) \tag{10.21}$$

10.7 Survival Copula

For studying extreme events, the probability of hydroclimatic variables being higher (or lower) than some threshold value, in other words, the tails of the distribution are of more interest. In such cases, a *reliability function* or *survival function* is defined as $\bar{F}_X(x) = P(X > x) = 1 - F_X(x)$, where $F_X(x)$ is *CDF* of random variable X.

The joint reliability function for a random variable pair (X, Y) is given by $\bar{H}_{X,Y}(x, y) = P(X > x, Y > y)$. The corresponding survival functions are given by $\bar{F}_X(x) = \bar{H}_{X,Y}(x, -\infty)$ and $\bar{F}_Y(y) = \bar{H}_{X,Y}(-\infty, y)$. The joint distribution function is related to joint reliability function as,

$$\bar{H}_{X,Y}(x, y) = 1 - F_X(x) - F_Y(y) + H_{X,Y}(x, y) \tag{10.22}$$

It should be noted that the $F_X(x)$ and $F_Y(y)$ are monotonically increasing function of X and Y, respectively; hence, $\bar{F}_X(x)$ and $\bar{F}_Y(y)$ are monotonic decreasing function of X and Y. If C is the copula function for $F_X(x)$ and $F_Y(y)$, then the copula function $\hat{C}$ for $\bar{F}_X(x)$ and $\bar{F}_Y(y)$ can be expressed by using Eq. 10.21 as,

$$\hat{C} = u + v - 1 + C(1 - u, 1 - v) \tag{10.23}$$

where u and v are reduced variates for X and Y. The copula function $\hat{C}$ is known as *survival copula* of X and Y. Hence, survival copula establishes a relationship between the joint survival function and marginals of X and Y in similar manner as done by copula C for joint distribution and marginal distribution. In hydroclimatology, the survival copula has been used for studying extreme events and its return period (Salvadori and De Michele 2007).

10.8 Most Commonly Used Copula Function

Many copula functions exist that follow the properties listed in Sect. 10.4.1 and lie between the bounds defined by Frechet–Hoeffding bounds (Sect. 10.4.1). Copula function can also be derived/constructed for different joint distributions on case by case basis. However, some of families of the copula functions are commonly used. Two popular classes of copula families are Elliptical copula and Archimedean copula.

10.8.1 Elliptical Copula

Elliptical copulas constitute a family of copula derived from elliptical distributions, such as normal distribution, Student's-t distribution. An elliptical copula tries to conserve the linear correlation between the data, and they use correlation coefficient ρ as parameter. The elliptical copulas do not have close formed expressions and are restricted to have a radial symmetry.

One of the popular elliptical copulas is Gaussian copula. A multidimensional Gaussian copula $C_R(u_1, u_2, \ldots, u_n)$ with correlation matrix R is given by:

$$C_R(u_1, u_2, \ldots, u_n) = \Phi_R^n\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_n)\right) \tag{10.24}$$

where $u_1, u_2, \ldots, u_n$ represent the reduced variates, Φ^{-1} is the inverse cumulative distribution function of a univariate standard normal distribution, and Φ_R^n is the joint cumulative distribution function of a multivariate normal distribution with zero mean vector and covariance matrix equal to correlation matrix between $\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_n)$, denoted as R. There is no analytical, closed-form solution for this copula function $C_R(u_1, u_2, \ldots, u_n)$.

The bivariate Gaussian copula can be written as,

$$C_R(u, v) = \Phi_R^2\left(\Phi^{-1}(u), \Phi^{-1}(v)\right) \tag{10.25}$$

In bivariate case, the correlation matrix R stands for $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, where ρ is correlation coefficient between $\Phi^{-1}(u)$, $\Phi^{-1}(v)$. The copula density for the same can be written as,

$$c_R(u, v) = \frac{1}{\sqrt{\rho^2 - 1}} \exp\left\{\frac{2\rho\Phi^{-1}(u)\Phi^{-1}(v) - \rho^2(\Phi^{-1}(u)^2 + \Phi^{-1}(v)^2)}{2(\rho^2 - 1)}\right\} \tag{10.26}$$

and the bivariate gaussian copula can be written as,

$$C_R(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{\sqrt{\rho^2 - 1}} \exp\left\{\frac{2\rho st - \rho^2(s^2 + t^2)}{2(\rho^2 - 1)}\right\} dsdt \tag{10.27}$$

Another commonly used elliptical copula is t-copula. For a matrix X having n different variables $(X_1, X_2, \ldots, X_n)$, if

$$X = \mu + \frac{\sqrt{\nu}}{\sqrt{S}} Z \tag{10.28}$$

where $\mu \in \mathbb{R}^n$, $S \sim \chi_\nu^2$ and $Z \sim \Phi_n(0, \Sigma)$ are independent, then X follows n-variate t$_\nu$-distribution with mean μ and covariance matrix $\frac{\nu}{\nu-2}\Sigma$ (for $\nu > 2$). If $\nu \leq 2$, then covariance of X is not defined. An n-dimensional t-copula $C_{\nu,R}(u_1, u_2, \ldots, u_n)$ for X is represented as,

$$C_{\nu,R}(u_1, u_2, \ldots, u_n) = t_{\nu,R}^n(t_\nu^{-1}(u_1), t_\nu^{-1}(u_2), \ldots, t_\nu^{-1}(u_n)) \tag{10.29}$$

where $u_1, u_2, \ldots, u_n$ represent the reduced variates for $X_1, X_2, \ldots, X_n$, respectively, t_ν^{-1} is the inverse cumulative distribution function of Student's t-distribution with ν degrees of freedom, and $t_{\nu,R}^n$ is the joint cumulative distribution function of a multivariate Student's t-distribution with ν degrees of freedom and covariance matrix equal to correlation matrix between $t_\nu^{-1}(u_1), t_\nu^{-1}(u_2), \ldots, t_\nu^{-1}(u_n)$, denoted as R. In bivariate case, the t-copula expression can be written as,

$$C_{\nu,R}(u, v) = \int_{-\infty}^{t_\nu^{-1}(u)} \int_{-\infty}^{t_\nu^{-1}(v)} \frac{1}{2\pi\sqrt{\det(R)}} \left\{1 + \frac{s^2 - 2\rho st + t^2}{\nu \det(R)}\right\}^{-(\nu+2)/\nu} dsdt \quad \text{for } \nu > 2 \tag{10.30}$$

where R is correlation matrix between $t_\nu^{-1}(u)$, $t_\nu^{-1}(v)$; thus, $R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, where ρ is correlation coefficient between $t_\nu^{-1}(u)$ and $t_\nu^{-1}(v)$. $\det(R)$ represents the determinant of matrix R.

10.8.2 Archimedean Copula

Archimedean copula is extensively used in hydrologic and hydroclimatic problems. Any copula that can be expressed in terms of $C(u, v) = \phi^{[-1]}(\phi(u) + \phi(v))$ is known as 'Archimedean copula,' where ϕ is known as generator function of copula. ϕ is a convex, strictly decreasing, continuous function from [0, 1] to [0, ∞), such that $\phi(1) = 0$ and $\phi^{[-1]}$ is its 'pseudo-inverse' $\phi^{[-1]} : [0, \infty) \rightarrow [0, 1]$. If $\phi(0)$ is ∞, then ϕ is called strict generator function, and corresponding Archimedean copula is called strict Archimedean copula. In case of strict generator function, the 'pseudo-inverse' ($\phi^{[-1]}$) is ϕ^{-1}, otherwise 'pseudo-inverse' ($\phi^{[-1]}$) is defined as:

$$\phi^{[-1]}(t) = \begin{cases} \phi^{-1}(t), & 0 \leq t \leq \phi(0) \\ 0, & \phi(0) < t \leq \infty \end{cases} \tag{10.31}$$

Any Archimedean copula C is symmetric and associative, i.e., if $u, v, w \in \mathrm{I}$, then $C(u, v) = C(v, u)$ and $C(C(u, v), w) = C(u, C(v, w))$. Further, Kendall's τ and the generator function are related in case of Archimedean copula. The relationship is expressed as,

$$\tau = 1 + 4 \int_0^1 \frac{\phi(u)}{\phi'(u)} du \tag{10.32}$$

This relationship is useful for obtaining joint distribution from sample measure of dependence in terms of the estimate of Kendall's tau. A list of few Archimedean copulas, commonly used in hydrology and hydroclimatology, is provided in Table 10.1. There are several advantages of Archimedean class of copulas that made them popular among researchers in the field of hydrology and hydroclimatology. Some such reasons are listed below.

(i) This class of copula can be easily constructed using generator function.
(ii) There are many different Archimedean copulas available that are applicable for a range of dependence parameters.
(iii) The different varieties of Archimedean copulas have useful properties, such as having an explicit expression based on generator and catering a high dimension using single parameter derived from measure of dependence.

Multivariate Archimedean Copula

Due to symmetry of two-dimensional Archimedean copula, these copula can be nested to get multivariate symmetrical Archimedean copula. For three-dimensional case,

$$C(u_1, u_2, u_3) = C(C(u_1, u_2), u_3) = \phi^{[-1]} \left(\phi(\phi^{[-1]}(\phi(u_1) + \phi(u_2))) + \phi(u_3)\right) \tag{10.33}$$

Table 10.1 Some Archimedean copulas

Copula type	Copula function $C_\theta(u, v)$	Generator function $\phi_\theta(t)$	$\theta \in$	Kendall's tau (τ) in terms of θ
Independent	uv	$\log(t)$		
Clayton	$\left[\max\left(u^{-\theta} + v^{-\theta} - 1, 0\right)\right]^{-1/\theta}$	$\frac{1}{\theta}\left(t^{-\theta} - 1\right)$	$[-1, \infty)\backslash\{0\}$	$\frac{\theta}{\theta+2}$
Frank	$-\frac{1}{\theta}\ln\left(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$	$-\ln\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$	$(-\infty, \infty)\backslash\{0\}$	$1 - \frac{4}{\theta}\left[D_1(\theta) - 1\right]$
Ali–Mikhail–Haq	$\frac{uv}{1-\theta(1-u)(1-v)}$	$\ln\left(\frac{1-\theta(1-t)}{t}\right)$	$[1, -1)$	$\left(\frac{3\theta-2}{\theta}\right) - \frac{2\ln(1-\theta)}{3}\left(1 - \frac{1}{\theta}\right)^2$
Gumbel–Hougaard	$\exp\left(-\left[(-\ln u)^\theta + (-\ln v)^\theta\right]^{1/\theta}\right)$	$(-\ln t)^\theta$	$[1, \infty)$	$\frac{\theta-1}{\theta}$
Joe	$1 - \left[(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta\right]^{1/\theta}$	$-\log\left(1 - (1-t)^\theta\right)$	$[1, \infty)$	

where, $D_1(\theta) = \frac{1}{\theta}\int_0^\theta \frac{t}{exp(t)-1}dt$ for $\theta > 0$ and $D_1(-\theta) = D_1(\theta) + \frac{\theta}{2}$

If generator function (ϕ) is strict and its inverse ϕ^{-1} is strictly monotonic on $[0, \infty)$, then n-dimensional Archimedean copula can be formed by nesting. It should also be noted that in Eq. 10.33, the generator functions are same. If generator function is not same while nesting two Archimedean copulas, then the nested copula is asymmetric. Some of the common nested asymmetric three-dimensional Archimedean copula is listed in Table 10.2 (after Joe 1997).

Example 10.8.1
Let $\phi_\theta(t) = (1-t)^\theta$ for $1 \leq \theta < \infty$ is the generator of an Archimedean copula. Formulate the corresponding Archimedean copula.

Solution The generator function is expressed as:

$$\phi_\theta(t) = (1-t)^\theta \text{ for } 1 \leq \theta < \infty$$

Thus,

$$\phi_\theta(0) = 1$$

The '*pseudo-inverse*' of the generator function is expressed as,

$$\phi^{[-1]} = \begin{cases} 1 - t^{\frac{1}{\theta}}, & 0 \leq t \leq 1 \\ 0, & t > 1 \end{cases}$$

and the corresponding Archimedean copula is expressed as,

$$\begin{aligned} C(u, v) &= \phi^{[-1]}(\phi(u) + \phi(v)) \\ &= \phi^{[-1]}\left((1-u)^\theta + (1-v)^\theta\right) \\ &= \left(1 - \left[(1-u)^\theta + (1-v)^\theta\right]^{\frac{1}{\theta}}\right) \qquad \text{for } u, v \in \mathrm{I} \end{aligned}$$

Here, it should be noted that $C(u, v)$ is not a valid copula function as it is not grounded, and there is chance that $\left(1 - \left[(1-u)^\theta + (1-v)^\theta\right]^{\frac{1}{\theta}}\right) < 0$; hence, to ensure that the function is bounded in I, the function $C(u, v)$ is modified as

$$\begin{aligned} C(u, v) &= \begin{cases} 1 - \left[(1-u)^\theta + (1-v)^\theta\right]^{\frac{1}{\theta}} & \text{for } 1 - \left[(1-u)^\theta + (1-v)^\theta\right]^{\frac{1}{\theta}} > 0 \\ 0 & \text{otherwise} \end{cases} \\ &= \max\left(1 - \left[(1-u)^\theta + (1-v)^\theta\right]^{\frac{1}{\theta}}, 0\right) \end{aligned}$$

Example 10.8.2
Formulate the Archimedean copula with generator function given as $\phi_\theta(t) = \left(1 - t^\theta\right)$ for $1 \leq \theta < \infty$?

Table 10.2 Some Asymmetric Archimedean Copula Family

Family	Nested copula $C_{\theta_1}(u_3, C_{\theta_2}(u_1, u_2))$	$\theta_2 \geq \theta_1 \in$	$\tau_{12}, \tau_{23}, \tau_{13} \in$
M3	$-\theta_1^{-1}\log\left[1-\left(1-e^{-\theta_1}\right)\left(1-\left(1-\left(1-e^{\theta_2}\right)^{-1}\left(1-e^{-\theta_2 u_1}\right)\left(1-e^{-\theta_2 u_2}\right)\right)^{\theta_1/\theta_2}\right)\left(1-e^{-\theta_1 u_3}\right)\right]$	$[0, \infty)$	$[0, 1]$
M4	$\left[\left(u_1^{-\theta_2}+u_2^{\theta_2}-1\right)^{\theta_1/\theta_2}+u_3^{-\theta_1}-1\right]^{-1/\theta_1}$	$[0, \infty)$	$[0, 1]$
M5	$1-\left[\left((1-u_1)^{\theta_2}+\left(1-(1-u_2)^{\theta_2}\right)+(1-u_2)^{\theta_2}\right)^{\theta_1/\theta_2}(1-(1-u_3)\,\theta_1)+(1-u_3)^{\theta_1}\right]^{1/\theta_1}$	$[1, \infty)$	$[0, 1]$
M6	$\exp\left[-\left(\left((-\log u_1)^{\theta_2}+(-\log u_2)^{\theta_2}\right)^{\theta_1/\theta_2}+(-\log u_3)^{\theta_1}\right)^{1/\theta_1}\right]$	$[1, \infty)$	$[0, 1]$
M12	$\left[\left(\left(\left(u_1^{-1}-1\right)^{\theta_2}+(u_2-1)^{\theta_2}\right)^{\theta_1/\theta_2}+\left(u_3^{-1}-1\right)^{\theta_1}\right)^{1/\theta_1}+1\right]^{(-1)}$	$[1, \infty)$	$[(1/3), 1]$

Solution The generator function is expressed as:

$$\phi_\theta(t) = (1 - t^\theta) \text{ for } 1 \leq \theta < \infty$$

At $t = 0$, $\phi_\theta(0) = 1$. Hence, the '*pseudoinverse*' of the generator function is expressed as,

$$\phi^{[-1]} = \begin{cases} (1-t)^{1/\theta}, & 0 \leq t \leq 1 \\ 0, & t > 1 \end{cases}$$

and the corresponding Archimedean copula is expressed as,

$$\begin{aligned} C(u, v) &= \phi^{[-1]}(\phi(u) + \phi(v)) \\ &= \phi^{[-1]}\left((1-u^\theta) + (1-v^\theta)\right) \\ &= \left(1 - \left[(1-u^\theta) + (1-v^\theta)\right]\right)^{1/\theta} \\ &= \left(u^\theta + v^\theta - 1\right)^{1/\theta} \qquad \text{for } u, v \in \mathrm{I} \end{aligned}$$

Here, it should be noted that $C(u, v)$ is not a valid copula function as it is not grounded, and there is chance that $\left(u^\theta + v^\theta - 1\right) < 0$; hence, to ensure that the function is bounded in I, the function $C(u, v)$ is modified as,

$$\begin{aligned} C(u, v) &= \begin{cases} \left(u^\theta + v^\theta - 1\right)^{1/\theta} & \text{for } \left(u^\theta + v^\theta - 1\right) > 0 \\ 0 & \text{otherwise} \end{cases} \\ &= \max\left(\left(u^\theta + v^\theta - 1\right)^{1/\theta}, 0\right) \end{aligned}$$

Example 10.8.3

In context of Example 10.5.1, the temperature for town A is found to follow normal distribution with mean of 17.5 °C and standard deviation of 2.7 °C. Similarly, the temperature for town B is distributed normally with mean 22 °C and standard deviation of 4.2 °C. Fit Clayton and Gumbel–Hougaard copula to the data.

Solution The observed temperature for both cities can be converted to reduced variate (u, v) through their respective marginal distributions.

u	0.588	0.962	0.672	0.500	0.995
v	0.622	0.830	0.798	0.972	0.930

From Example 10.5.1, $\tau = 0.2$.

(a) Clayton Copula Fitting

The parameter for Clayton copula can be calculated as (Table 10.1)

$$\begin{aligned}\tau &= \frac{\theta}{\theta+2}\\ \text{or, } 0.2(\theta+2) &= \theta\\ \text{or, } \theta &= 0.5\end{aligned}$$

The value of θ is valid for Clayton copula because the parameter for Clayton copula must be in the range of $[-1,\infty)\backslash\{0\}$. Hence,

$$\begin{aligned}u^{-\theta} = u^{-0.5} &= \begin{bmatrix}1.304 & 1.019 & 1.220 & 1.414 & 1.002\end{bmatrix}\\ v^{-\theta} = v^{-0.5} &= \begin{bmatrix}1.268 & 1.098 & 1.120 & 1.014 & 1.037\end{bmatrix}\\ \max(u^{-\theta}+v^{-\theta}-1,0) &= \begin{bmatrix}1.573 & 1.117 & 1.340 & 1.429 & 1.039\end{bmatrix}\\ \text{So, } C_\theta(u,v) = \left[\max(u^{-\theta}+v^{-\theta}-1,0)\right]^{-1/\theta} &= \begin{bmatrix}0.404 & 0.801 & 0.557 & 0.490 & 0.926\end{bmatrix}\end{aligned}$$

(b) Gumbel–Hougaard Copula Fitting

From Table 10.1, the parameter for Gumbel–Hougaard copula in terms of τ is given as

$$\begin{aligned}\tau &= \frac{\theta-1}{\theta}\\ \text{or, } \theta &= \frac{1}{0.8} = 1.25\end{aligned}$$

The value of parameter $\theta = 1.25$ is valid for Gumbel–Hougaard copula ($\theta \in [1,\infty)$)

$$\begin{aligned}(-\ln u)^\theta = (-\ln u)^{1.25} &= \begin{bmatrix}0.453 & 0.017 & 0.316 & 0.632 & 0.001\end{bmatrix}\\ (-\ln v)^\theta = (-\ln v)^{1.25} &= \begin{bmatrix}0.395 & 0.123 & 0.156 & 0.012 & 0.038\end{bmatrix}\\ \left[(-\ln u)^\theta + (-\ln v)^\theta\right]^{1/\theta} &= \begin{bmatrix}0.704 & 0.877 & 0.548 & 0.207 & 0.074\end{bmatrix}\\ \text{So, } C_\theta(u,v) = \exp\left(-\left[(-\ln u)^\theta + (-\ln v)^\theta\right]^{1/\theta}\right) &= \begin{bmatrix}0.416 & 0.813 & 0.578 & 0.495 & 0.928\end{bmatrix}\end{aligned}$$

Example 10.8.4

For the data given in Example 10.5.2, monthly anomaly of OLR is distributed normally with mean 0 W/m^2 and standard deviation 8 W/m^2. The monthly precipitation is found to follow exponential distribution with mean 4 cm. Fit Clayton, Frank, Ali–Mikhail–Haq, and Gaussian copula over the data set.

Solution The observed monthly mean OLR (X) and precipitation (Y) can be converted to their reduced variates (u and v, respectively).

u	0.9878	0.0304	0.9599	0.4013	0.0020	0.5497
v	0.2212	0.9179	0.2212	0.8262	0.9502	0.3935

From Example 10.5.2, $\tau = -0.93$

(a) Clayton Copula Fitting

The parameter for Clayton copula can be calculated as (Table 10.1)

$$\begin{aligned}\tau &= \frac{\theta}{\theta+2}\\ \text{or, } &-0.93(\theta+2) = \theta\\ \text{or, } &\theta = -0.964\end{aligned}$$

As $\theta \in [-1, \infty)\backslash\{0\}$, the value of θ is valid for Clayton copula. Hence,

$$u^{-\theta} = u^{0.964} = \begin{bmatrix} 0.9882 & 0.0345 & 0.9614 & 0.4147 & 0.0025 & 0.5617 \end{bmatrix}$$

$$v^{-\theta} = v^{0.964} = \begin{bmatrix} 0.2335 & 0.9207 & 0.2335 & 0.8319 & 0.9520 & 0.4069 \end{bmatrix}$$

$$\max(u^{-\theta} + v^{-\theta} - 1, 0) = \begin{bmatrix} 0.2218 & 0 & 0.1949 & 0.2466 & 0 & 0 \end{bmatrix}$$

$$\text{So, } C_\theta(u, v) = \left[\max(u^{-\theta} + v^{-\theta} - 1, 0)\right]^{-1/\theta} = \begin{bmatrix} 0.2096 & 0 & 0.1834 & 0.2341 & 0 & 0 \end{bmatrix}$$

(b) Frank Copula Fitting

Frank copula parameter (θ) is calculated as,

$$\begin{aligned}\tau &= 1 - \frac{4}{\theta}\left[D_1(\theta) - 1\right]\\ \text{or, } 1.93 &= \frac{4}{\theta}\left[D_1(\theta) - 1\right]\\ \text{where, } D_1(\theta) &= \frac{1}{\theta}\int_0^\theta \frac{t}{exp(t) - 1}dt\end{aligned}$$

Solving the above equation numerically, $\theta = -55.45$, which is a valid parameter value for Frank copula. Hence,

$$\begin{aligned}C_\theta(u, v) &= -\frac{1}{\theta}\ln\left(1 + \frac{\left(e^{-\theta u} - 1\right)\left(e^{-\theta v} - 1\right)}{e^{-\theta} - 1}\right)\\ &= \frac{1}{55.45}\ln\left(1 + \frac{\left(e^{55.45u} - 1\right)\left(e^{55.45v} - 1\right)}{e^{55.45} - 1}\right)\end{aligned}$$

The value of u and v can be substituted, and the Frank copula values can be calculated using above equation.

$$C_\theta(u, v) = \begin{bmatrix} 0.2090 & 0.0008 & 0.1811 & 0.2275 & 0.0001 & 0.0008 \end{bmatrix}$$

(c) Ali–Mikhail–Haq Copula Fitting

From Table 10.1, the parameter for Ali–Mikhail–Haq copula in terms of τ is given as

$$\tau = \left(\frac{3\theta - 2}{\theta}\right) - \frac{2\ln(1-\theta)}{3}\left(1 - \frac{1}{\theta}\right)^2$$

Hence, $\theta = -353.3$, and this value of parameter θ is invalid for Ali–Mikhail–Haq copula. So Ali–Mikhail–Haq copula cannot be used for modeling the relationship between mean monthly OLR and precipitation.

(d) Gaussian Copula Fitting

The standard normal reduced variates (Table B.1, p. 434) corresponding to u and v are,

$$\Phi^{-1}(u) = \begin{bmatrix} 2.250 & -1.875 & 1.750 & -0.250 & -2.875 & 0.125 \end{bmatrix}$$
$$\Phi^{-1}(v) = \begin{bmatrix} -0.768 & 1.391 & -0.768 & 0.939 & 1.647 & -0.270 \end{bmatrix}$$

For fitting Gaussian copula, the correlation matrix needs to be calculated between $\Phi^{-1}(u)$ and $\Phi^{-1}(v)$. As described in Sect. 3.5.6, the correlation coefficient is given by

$$\rho = \frac{\sigma_{\Phi^{-1}(u),\Phi^{-1}(v)}}{\sigma_{\Phi^{-1}(u)}\sigma_{\Phi^{-1}(v)}} = \frac{-2.074}{\sqrt{3.972 \times 1.200}} = -0.95$$

Hence, the correlation matrix between $\Phi^{-1}(u)$ and $\Phi^{-1}(v)$ is given by

$$R = \begin{bmatrix} 1 & -0.95 \\ -0.95 & 1 \end{bmatrix}$$

Using R, the Gaussian copula function is expressed as (Eq. 10.24)

$$C_R(u, v) = \Phi_R\left(\Phi^{-1}(u), \Phi^{-1}(v)\right)$$

where Φ_R is bivariate Gaussian distribution *CDF* with zero mean vector and covariance R. As bivariate normal distribution *CDF* cannot be solved analytically, hence, the numerical solution for the Gaussian copula for values of u and v is given as

$$C_R(u, v) = \begin{bmatrix} 0.2090 & 0.0009 & 0.1812 & 0.2281 & 0.0000 & 0.0262 \end{bmatrix}$$

10.9 Selection of Best-Fit Copula

Given the sample data, if there are more than one potential copulas, best-fit copula has to be selected. There are several goodness-of-fit (GOF) tests for statistically checking the suitability of a copula. Most of these approaches use (a) empirical copula, (b) Kendall's transform, and (c) Rosenblatt's transform.

10.9.1 Test Using Empirical Copula

These tests compare the distance between the empirical copula ($C_n(u, v)$) and parametric estimate $\left(C_n^\theta(u, v)\right)$ of C, where u and v are the reduced variates of the sample data X $(x_1, x_2, \ldots, x_n)$ and Y $(y_1, y_2, \ldots, y_n)$, respectively, and n is the number of observations. The empirical copula (C_n) is defined as:

$$C_n(u, v) = \frac{1}{n} \sum_{\forall u, v} \Im(U \le u, V \le v), \qquad u, v \in \mathrm{I} \qquad (10.34)$$

where $\Im(\bullet)$ is the indicator function that takes a value of 1 if the argument $(\bullet)$ is true and 0 if it is false. The Cramér-von Mises and Kolmogorov–Smirnov (KS) statistics are based on the distance between fitted copula and empirical copula. The Cramér-von Mises statistic (S_n) is a popular goodness-of-fit test statistic for copula models (Genest et al. 2007). The statistic S_n is expressed as:

$$S_n = \sum_{\forall u, v} \left(C_n(u, v) - C_n^\theta(u, v)\right)^2 \qquad (10.35)$$

The KS statistic (T_n) is based on the absolute maximum distance between C_n and C_n^θ. It is expressed as:

$$T_n = \max_{u, v \in \mathrm{I}} \left| \sqrt{n} \left(C_n(u, v) - C_n^\theta(u, v)\right) \right| \qquad (10.36)$$

10.9.2 Test Using Kendall's Transform

For the best-fit copula selection procedure using Kendall's transform, κ is obtained from the joint distribution, derived parametrically using a particular copula, C_n^θ. It is expressed as follows:

$$\kappa(t) = P(C_n^\theta(u, v) \le t) \qquad (10.37)$$

The κ is determined either parametrically $\left(\kappa_n^\theta\right)$ or nonparametrically (κ_n). κ_n is derived using the empirical distribution function C_n (Genest et al. 1993, 2009) as given below,

$$\kappa_n(t) = \frac{1}{n}\sum_{\forall u,v} \Im(C_n(u,v) \le t) \tag{10.38}$$

The test statistics $\left(S_n^{(\kappa)} \text{ and } T_n^{(\kappa)}\right)$ are basically the rank-based analogues of the Cramér-von Mises and KS statistics (Genest et al. 2009). The test statistics $S_n^{(\kappa)}$ and $T_n^{(\kappa)}$ are expressed as,

$$S_n^{(\kappa)} = \sum_{\forall t} (\kappa_n(t) - \kappa(t))^2 \tag{10.39}$$

$$T_n^{(\kappa)} = \sup_{\forall t} |\kappa_n(t) - \kappa(t)| \tag{10.40}$$

10.9.3 Test Using Rosenblatt's Probability Integral Transformation

The Rosenblatt's probability integral transformation of the copula is defined as $R(u, v) = (e_1, e_2)$, where $e_1 = \partial C_n^\theta \big/ \partial u$, $e_2 = \partial C_n^\theta \big/ \partial v$. Based on the properties of *Rosenblatt*'s transform, (u, v) is approximately distributed as C_n^θ, if and only if the $R(u, v)$ is a bivariate independent copula, i.e., $C_\perp(e_1, e_2) = e_1 \times e_2$, where $e_1, e_2 \in \mathrm{I}$. The R is estimated either parametrically $\left(R_n^\theta\right)$ or nonparametrically (R_n). The R_n is derived following Genest et al. (2009) as follows,

$$R_n(e_1, e_2) = \frac{1}{n}\sum_{i=1}^{n} \Im(E_1 \le e_1, E_2 \le e_2) \qquad \text{for } e_1, e_2 \in \mathrm{I} \tag{10.41}$$

R_n^θ as stated above is given by $C_\perp$. Further the two Cramér-von Mises statistics, $S_n^{(B)}$ and $S_n^{(C)}$, are estimated to check the distance between R_n and R_n^θ. $S_n^{(B)}$ can be calculated as:

$$S_n^{(B)} = n\sum_{i=1}^{n} \left(R_n(e_1, e_2) - C_\perp(e_1, e_2)\right)^2 \tag{10.42}$$

and $S_n^{(C)}$ can be estimated as

$$S_n^{(C)} = n\sum_{i=1}^{n} \left(R_n(e_1, e_2) - C_\perp(e_1, e_2)\right)^2 R_n(e_1, e_2) \tag{10.43}$$

For all the measures $\left(S_n, T_n, S_n^{(\kappa)}, T_n^{(\kappa)}, S_n^{(B)} \text{ and } S_n^{(C)}\right)$, the lower the value, the better is the fit. Thus, the copula function with the lowest value of these statistics indicates the best-fit copula. Further, when the best-fit copula is found to be different using different statistics, the more preferable statistic is honored while selecting the best-fit copula. The preference order is $S_n^{(B)} \succ S_n \succ S_n^{(\kappa)} \succ S_n^{(C)} \succ T_n \succ T_n^{\kappa}$ based on their power (Genest et al. 2009). The copula showing best-fit based on these criteria is selected for further analysis and denoted as $C(u, v)$.

Example 10.9.1
For the two copula models fitted in Example 10.8.3, calculate the Cramér-von Mises statistic and Kolmogorov–Smirnov (KS) statistic. Select the best copula based on these statistics.

Solution The empirical copula function is given by (Eq. 10.34)

$$C_n(u, v) = \frac{1}{n} \sum_{\forall u,v} \Im(U \le u, V \le v) = \begin{bmatrix} 0.2 & 0.6 & 0.4 & 0.2 & 0.8 \end{bmatrix}$$

(a) Goodness-of-fit statistics for fitted Clayton copula

The Cramér-von Mises statistic is given by (Eq. 10.35),

$$\begin{aligned} S_n &= \sum_{\forall u,v} \left(C_n(u, v) - C_n^{\theta}(u, v)\right)^2 \\ &= (0.404 - 0.2)^2 + (0.801 - 0.6)^2 + (0.557 - 0.4)^2 + (0.490 - 0.2)^2 + (0.926 - 0.8)^2 \\ &= 0.207 \end{aligned}$$

The KS statistic is given by (Eq. 10.36),

$$T_n = \max_{u,v \in \mathrm{I}} \left|\sqrt{n}\left(C_n(u, v) - C_n^{\theta}(u, v)\right)\right| = \sqrt{5} \times 0.29 = 0.648$$

(b) Goodness-of-fit statistics for fitted Gumbel–Hougaard copula

The Cramér-von Mises statistic for fitted Gumbel–Hougaard copula is given by (Eq. 10.35),

$$\begin{aligned} S_n &= \sum_{\forall u,v} \left(C_n(u, v) - C_n^{\theta}(u, v)\right)^2 \\ &= (0.416 - 0.2)^2 + (0.813 - 0.6)^2 + (0.578 - 0.4)^2 + (0.495 - 0.2)^2 + (0.928 - 0.8)^2 \\ &= 0.227 \end{aligned}$$

The KS statistic is given by (Eq. 10.36),

$$T_n = \max_{u,v \in \mathrm{I}} \left|\sqrt{n}\left(C_n(u, v) - C_n^{\theta}(u, v)\right)\right| = \sqrt{5} \times 0.295 = 0.659$$

In this example, lower values of both S_n and T_n suggest that Clayton copula is better compared to Gumbel–Hougaard copula. However, in some applications, both S_n and

T_n may not agree on best copula, and then, copula should be selected on the basis of S_n, as S_n has more power compared to T_n $\left(S_n^{(B)} \succ S_n \succ S_n^{(\kappa)} \succ S_n^{(C)} \succ T_n \succ T_n^{\kappa}\right)$.

10.10 Use of Copulas

Copula can be used in hydroclimatic studies for data generation, multivariate frequency analysis, probabilistic prediction of hydroclimatic variables, and many other applications. These uses of copula are discussed in the following subsections.

10.10.1 Data Generation

Data generation using copula preserves the dependence structure between the associated variables. Two methods exist for data generation using copula: One is specific to Archimedean copula and other can be applied to any copula. These methods are discussed as follows:

(i) Simulation of random variates preserving the dependence structure using Archimedean copula can be done using the following algorithm (Genest et al. 1986):

   (a) For an Archimedean copula, functional forms of $\phi^{[-1]}(\bullet)$, $\phi'(\bullet)$ and $\phi'^{[-1]}(\bullet)$ are obtained using $\phi_\theta(\bullet)$, which is the generator function with parameter θ. Equation 10.31 is used to obtain $\phi^{[-1]}(\bullet)$. Same can be used for $\phi'^{[-1]}(\bullet)$ after obtaining $\phi'(\bullet)$, which is derivative of $\phi(\bullet)$ with respect to $\bullet$.
   (b) Two independent uniformly distributed ($U(0, 1)$) random variates, u and r, are generated.
   (c) Two new variables, S and W, are obtained as $s = \phi'(u)\big/r$ and $w = \phi'^{[-1]}(s)$.
   (d) Another variable, v, is obtained as $v = \phi^{[-1]}(\phi(w) - \phi(u))$ (Genest et al. 1986). The pairs u and v are the simulated pair, preserving the dependence structure.
   (e) Both these u and v are in the range [0, 1]. These simulated pairs of u and v are then back-transformed through their corresponding cumulative marginal distributions.

(ii) The more generalized approach for data generation uses the conditional probability function developed from copula. Hence, before applying this method, marginal distribution of the associated variables and copula function or joint distribution should be known. If the reduced variate of the variables is denoted by u and v, respectively, then the steps are given as,

   (a) Conditional distribution for v given u is obtained from the copula or joint distribution (Eq. 10.54).

(b) Uniformly distributed ($U(0, 1)$) random variates u and p are generated.
(c) Substituting u in the expression for conditional distribution of v given u, the expression is equated to p to solve for v. This gives the values of v based on the dependence structure between u and v.
(d) Both u and v are in the range [0, 1]. These simulated pairs of u and v are then back-transformed using their corresponding inverse cumulative marginal distribution.

Example 10.10.1

From the historical records, monthly maximum rainfall duration (in hour) is found to follow exponential distribution with mean maximum rainfall duration of $1\big/2$ h. The monthly maximum discharge (in cumec) is found to follow a normal distribution with mean 500 cumec and standard deviation of 36.5 cumec, if the joint distribution between these variables can be obtained by using Ali–Mikhail–Haq copula with $\theta = 0.5$. Generate monthly maximum rainfall duration and monthly maximum discharge for a year.

Solution Let us assume that X and Y are two random variables (with corresponding reduced variates u and v) representing the monthly maximum rainfall duration and monthly maximum discharge, respectively. The *CDF* for X and Y is given by,

$$\begin{aligned} F_X(x) &= 1 - e^{-\lambda x} = 1 - e^{-2x} && x \geq 0 \\ F_Y(y) &= \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2} dy && -\infty < y < \infty \\ &= \int_{-\infty}^{y} \frac{1}{91.49} e^{-(y-500)^2/2664.5} dy && -\infty < y < \infty \end{aligned}$$

The joint distribution of variables X and Y, using Ali–Mikhail–Haq copula with $\theta = 0.5$, can be evaluated using (Table 10.1),

$$F_{X,Y}(x, y) = \frac{uv}{1 - 0.5(1-u)(1-v)} \qquad \text{for } u, v \in [0, 1]$$

The conditional distribution for Y conditioned on X is given by,

$$F_{Y/X}(y\big/X = x) = \frac{\partial F_{X,Y}(x, y)}{\partial u} = \frac{uv(v/2 - 1/2)}{((u/2 - 1/2)(v-1) - 1)^2} - \frac{v}{(u/2 - 1/2)(v-1) - 1}$$

The 12 randomly generated values for u are 0.74, 0.08, 0.13, 0.10, 0.97, 0.33, 0.56, 0.03, 0.20, 0.70, 0.04, and 0.94. Similarly, 12 random values for p are 0.44, 0.38, 0.77, 0.80, 0.19, 0.49, 0.45, 0.65, 0.71, 0.75, 0.63, and 0.25. For estimating the values of v, the conditional distribution of Y given X is solved by substituting u and equating to p. The obtained values of v are,

$$v = \begin{bmatrix} 0.51 & 0.26 & 0.68 & 0.71 & 0.29 & 0.45 & 0.48 & 0.50 & 0.63 & 0.79 & 0.48 & 0.35 \end{bmatrix}$$

The reduced variates u and v can be transformed back to the variables X and Y using the relationship given above. Hence, the back-transformed variables are given by,

$$X = \begin{bmatrix} 0.67 & 0.04 & 0.07 & 0.05 & 1.75 & 0.20 & 0.41 & 0.02 & 0.11 & 0.60 & 0.02 & 1.41 \end{bmatrix}$$
$$Y = \begin{bmatrix} 501.0 & 476.8 & 517.0 & 519.8 & 479.7 & 495.4 & 497.9 & 499.6 & 512.4 & 529.0 & 498.1 & 486.4 \end{bmatrix}$$

Example 10.10.2
For the best copula selected in Example 10.9.1, generate 10 random data for temperature of cities A and B by using the methodology proposed by Genest et al. 1986.

Solution The best selected copula is Clayton copula with $\theta = 0.5$. Hence, the generator function is given by:

$$\phi_\theta(t) = \frac{1}{\theta}(t^{-\theta} - 1) = 2(t^{-0.5} - 1)$$
$$\phi_\theta^{[-1]}(t) = (\theta t + 1)^{-1/\theta} = (0.5t + 1)^{-2}$$
$$\phi'_\theta(t) = \frac{d\phi_\theta(t)}{dt} = 2(-0.5t^{-1.5}) = -t^{-1.5}$$

Hence, $\phi'_\theta(0) = \infty$, so $\phi_\theta^{'[-1]}(t)$ is given by,

$$\phi_\theta^{'[-1]}(t) = (-t)^{-1/1.5}$$

Ten random numbers between 0 and 1 generated for u are 0.93, 0.69, 0.05, 0.18, 0.19, 0.75, 0.85, 0.36, 0.83, and 0.59. Similarly, the random numbers between 0 and 1 generated for r are 0.65, 0.01, 0.56, 0.51, 0.46, 0.75, 0.02, 0.07, 0.23, and 0.73. S and W are obtained as $s = \phi'(u)\big/r$ and $w = \phi'^{[-1]}(s)$. Hence,

$$S = \begin{bmatrix} -1.7 & -174.5 & -159.7 & -25.7 & -26.2 & -2.0 & -63.8 & -66.1 & -5.7 & -3.0 \end{bmatrix}$$

And

$$W = \begin{bmatrix} 0.69 & 0.03 & 0.03 & 0.11 & 0.11 & 0.62 & 0.06 & 0.06 & 0.31 & 0.48 \end{bmatrix}$$

Simulated v can be obtained using the relationship $v = \phi^{[-1]}(\phi(w) - \phi(u)) = \left(w^{-0.5} - u^{-0.5} + 1\right)^{-2}$

$$v = \begin{bmatrix} 0.75 & 0.03 & 0.19 & 0.36 & 0.34 & 0.8 & 0.06 & 0.09 & 0.35 & 0.77 \end{bmatrix}$$

The u and v can then be back-transformed to generate the temperature of cities A and B using the information about their marginals as given in Example 10.8.3,

$$T_A = \begin{bmatrix} 21.48 & 18.84 & 13.06 & 15.03 & 15.13 & 19.32 & 20.30 & 16.53 & 20.08 & 18.11 \end{bmatrix}$$
$$T_B = \begin{bmatrix} 24.83 & 14.10 & 18.31 & 20.49 & 20.27 & 25.53 & 15.47 & 16.37 & 20.38 & 25.10 \end{bmatrix}$$

Example 10.10.3

In the last example, considering the same values of u, generate 10 values of temperature of cities A and B by using the conditional relationship between the temperature of two cities.

Solution The best copula fitted between temperature of cities A and B is Clayton; hence, their joint distribution is given by,

$$F_{T_A,T_B}(t_A, t_B) = C(F_{T_A}(t_A), F_{T_B}(t_B)) = \left[\max\left(u^{-0.5} + v^{-0.5} - 1, 0\right)\right]^{-1/0.5}$$

where $u = F_{T_A}(t_A)$ and $v = F_{T_B}(t_B)$. The conditional distribution for temperature variate for city B when temperature variate for city A is available is given by:

$$\begin{aligned} F_{T_B/T_A}(t_B/T_A = t_A) &= \frac{\partial\left[\max\left(u^{-0.5} + v^{-0.5} - 1, 0\right)\right]^{-1/0.5}}{\partial u} \\ &= \begin{cases} \frac{\partial\left(u^{-0.5}+v^{-0.5}-1\right)^{-2}}{\partial u} & \text{for } \left(u^{-0.5} + v^{-0.5} - 1\right) > 0 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} u^{-1.5}\left(u^{-0.5} + v^{-0.5} - 1\right)^{-3} & \text{for } \left(u^{-0.5} + v^{-0.5} - 1\right) > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Ten random numbers between 0 and 1 generated for u are 0.93, 0.69, 0.05, 0.18, 0.19, 0.75, 0.85, 0.36, 0.83, and 0.59. Similarly, 10 random numbers for p between 0 and 1 are 0.35, 0.83, 0.59, 0.55, 0.92, 0.29, 0.76, 0.75, 0.38, and 0.57. For generation of random variates of temperature values at City B, the conditional distribution for temperature of city B given temperature of city A should be equal to p, i.e., $F_{T_B/T_A}(t_B/T_A = t_A) = p$. Hence, corresponding to the value of temperature variate of city B, (v) is evaluated as,

$$v = \begin{bmatrix} 0.49 & 0.86 & 0.29 & 0.43 & 0.88 & 0.40 & 0.82 & 0.73 & 0.50 & 0.62 \end{bmatrix}$$

The u and v can be converted into temperature values for the cities by using their marginal distributions. According to Example 10.8.3, the temperature for town A follows normal distribution with mean of 17.5 °C and standard deviation of 2.7 °C. Similarly, the temperature for town B is distributed normally with mean 22 °C and standard deviation of 4.2 °C.

$$T_A = \begin{bmatrix} 21.48 & 18.84 & 13.06 & 15.03 & 15.13 & 19.32 & 20.30 & 16.53 & 20.08 & 18.11 \end{bmatrix}$$
$$T_B = \begin{bmatrix} 21.85 & 26.57 & 19.66 & 21.29 & 26.98 & 20.89 & 25.85 & 24.62 & 21.97 & 23.30 \end{bmatrix}$$

10.10.2 Probabilistic Prediction Using Copulas

Another potential application of copulas is probabilistic prediction of hydrologic and hydroclimatic variables. Major steps to be followed are presented in a flow chart (Fig. 10.6). As shown in Fig. 10.6, there are three major steps: (A) data preprocessing and analysis; (B) fitting suitable copula model; and (C) prediction of the dependent variable. These steps are explained below in detail.

(A) Data preprocessing and analysis: Data preprocessing includes many general statistical operations such as missing value treatment, outlier removal. However, these are general steps for any statistical modeling and should be carried out with caution. For instance, outliers need not always be erroneous data. These may be the data for extreme events.
Specific to application of copulas, estimation of scale-free measure of association and fitting a suitable marginal distribution are essential.

(a) Estimation of scale-free measure of association:

As mentioned before (Sect. 10.5), there are two popular scale-free measures of association (nonparametric measures of association), namely Kendall's tau (τ) and Spearman's rho (ρ_s). Sample estimates of these statistics are computed from the data.

(b) Estimation of marginal distributions: Fitting univariate parametric marginal distribution to any random variable is discussed in Chap. 6, using the theory from Chaps. 4 and 5. Readers may refer to chi-square test, Kolmogorov–Smirnov test, Anderson–Darling test, etc., for this purpose (Sect. 6.4.4). However, a parametric distribution may not always fit to hydrologic or hydroclimatic data with reasonable accuracy. In such cases, a nonparametric distribution may be adopted. Methodology to fit a nonparametric distribution to the data is explained as follows:
Kernel density estimator is the most popular method for estimation of nonparametric density (Bosq 2012). The kernel estimate of probability density, for a real-valued time series, x_i, $i = 1, 2, \ldots, n$, can be expressed as,

$$\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \tag{10.44}$$

where $K_h(z) = 1\big/hKr\left(z\big/h\right)$, in which h is the smoothing parameter and Kr is the kernel function. Different types of kernel functions are naïve, normal, and Epanechnikov. Mathematical formulations of these kernel functions are shown below (Bosq 2012).

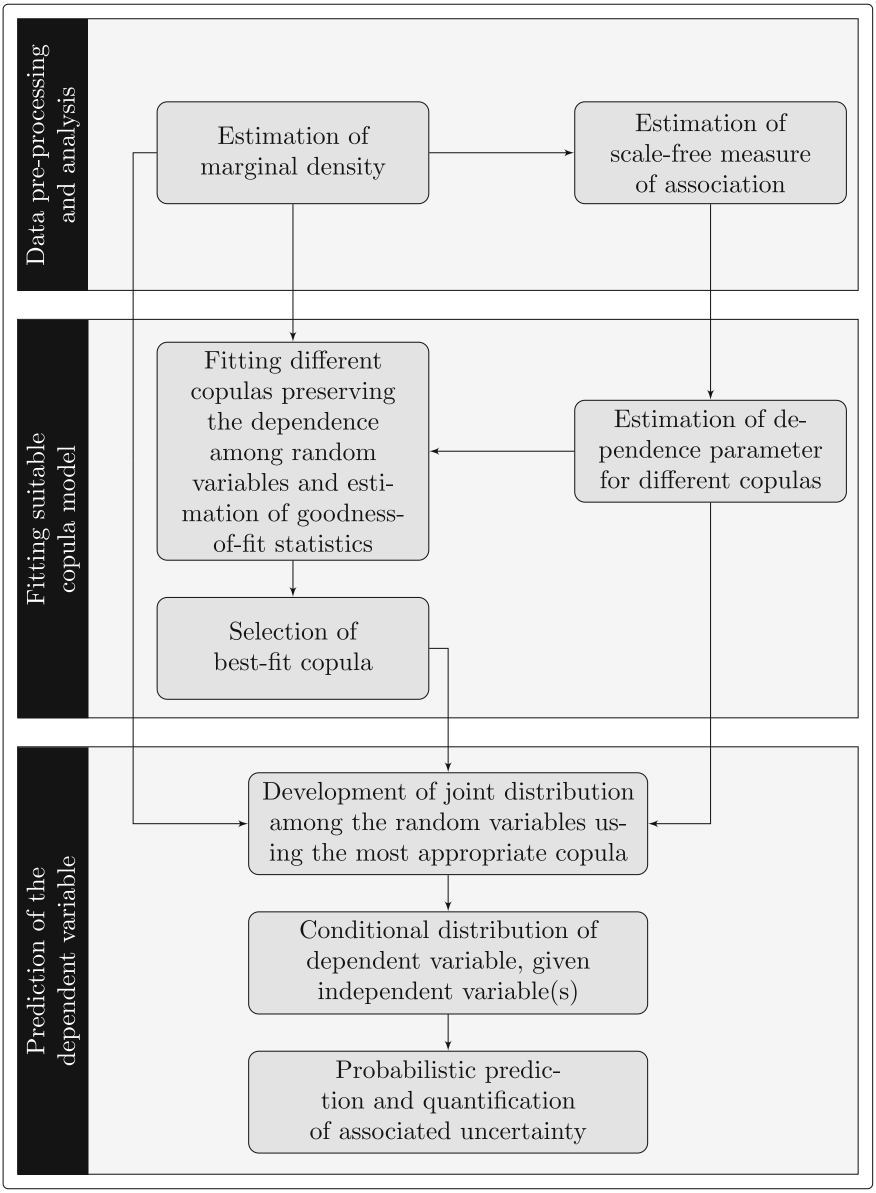

Data pre-processing and analysis
Estimation of marginal density
Estimation of scale-free measure of association
Fitting suitable copula model
Fitting different copulas preserving the dependence among random variables and estimation of goodness-of-fit statistics
Estimation of dependence parameter for different copulas
Selection of best-fit copula
Prediction of the dependent variable
Development of joint distribution among the random variables using the most appropriate copula
Conditional distribution of dependent variable, given independent variable(s)
Probabilistic prediction and quantification of associated uncertainty

Fig. 10.6 Flow chart for prediction using copula

Naive	$Kr(u) = 1$	$-\frac{1}{2} \le u \le \frac{1}{2}$
Normal	$Kr(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$	$-\infty < u < \infty$
Epanechnikov	$Kr(u) = \frac{3}{4\sqrt{5}}\left(1 - \frac{u^2}{5}\right)$	$-\sqrt{5} \le u \le \sqrt{5}$

The cumulative probability density is obtained from the corresponding nonparametrically estimated probability density. Either parametric or nonparametric marginal probabilistic distribution is used to obtain the reduced variate of the random variable.

(B) Fitting suitable copula model: Once the reduced variates are obtained using the fitted marginal distribution, a number of candidate copula functions are selected. These copula functions are fitted by estimating their parameters form the reduced variates. The best-fitted copula is then selected from all fitted candidate copula.

(a) Copula fitting: Let us consider ordered pairs of random variables X and Y represented as $(x_1, y_1), \ldots, (x_n, y_n)$. In general, parameter(s) of a copula function (C_θ) can be estimated using different methods, such as:

(i) Inversion of scale-free measure of association,
(ii) Maximum likelihood estimate (MLE),
(iii) Inference from margin (IFM),
(iv) Canonical maximum likelihood (CML).

Inversion of Scale-Free Measure of Association

In case of one-parameter bivariate copulas, the popular approach is the inversion of Spearman's or Kendall's rank correlation (Genest et al. 2007). The relationship between Kendall's tau ($\hat{\tau}$) and the dependence parameter θ is provided in Table 10.1 for some of the Archimedean copulas. After obtaining the sample estimate of Kendall's tau (τ), the copula parameter θ can be estimated.

Maximum Likelihood Estimate

In general, the method of maximum likelihood is discussed in Chap. 3 (Sect. 3.7.2). Using the method of maximum log-likelihood estimate (MLE), the copula structure with the copula parameters and individual marginal parameters can be estimated. For MLE of copula parameters, log-likelihood function of joint *pdf* is maximized with respect to all the parameters. A bivariate joint *pdf* can be expressed in terms of copula *pdf* as,

$$f_{X,Y}(x, y) = c(F_X(x), F_Y(y) : \theta) f_X(x : \alpha_1) f_Y(y : \alpha_2) \tag{10.45}$$

where θ is copula parameter and α_1 and α_2 are the parameters for marginal *pdf*'s. The log likelihood of above expression is given by,

$$\sum \log\left(f_{X,Y}(x,y)\right) = \sum_i \log\left(c(F_X(x_i : \alpha_1), F_Y(y_i : \alpha_2) : \theta)\right)$$

$$+ \sum_{i=1}^{n} \log f_X(x_i : \alpha_1) + \sum_{j=1}^{n} \log f_Y(y_j : \alpha_2) \tag{10.46}$$

$$L = L_C + L_{M1} + L_{M2}$$

$$L = L_C + L_M$$

where L, L_C, L_M show the total log likelihood of the joint *pdf*, copula *pdf*, and marginals, respectively. For MLE of parameters, following equations are required to be solved simultaneously:

$$\left(\partial L / \partial \alpha_1, \partial L / \partial \alpha_2, \partial L / \partial \theta\right) = 0 \tag{10.47}$$

Inference From Margins

Another approach to estimate the parameters is inference from marginal (IFM). In IFM, Eq. 10.46 decomposes the maximum log likelihood into two parts: one from copula dependence (L_C) and other from the marginals (L_M). In the first step of IFM, the marginal parameters are estimated for each of marginal functions individually, i.e., for ith marginal: $\hat{\alpha}_i = \text{argmax}_{\alpha_i} L_{Mi}(\alpha_i)$. In the second stage, using the estimated $(\hat{\alpha}_1, \hat{\alpha}_2)$, L_C is maximized to get an estimate for θ: $\hat{\theta} = \text{argmax}_\theta L_C(\theta)$. Hence, in IFM, following set of equations are solved for getting the estimates of parameters.

$$\left(\partial L_{M1} / \partial \alpha_1, \partial L_{M2} / \partial \alpha_2, \partial L_C / \partial \theta\right) = 0 \tag{10.48}$$

As per Joe (1997), the MLE and IFM estimation procedures are equivalent when all the variables $(X, Y, \ldots)$ follow univariate normal marginal and have multivariate joint normal *pdf* associated with them along with having a Gaussian copula.

The MLE and IFM can be extended to multivariate copula functions; however, it is computationally intensive to solve the equations simultaneously. Canonical maximum likelihood may be another alternative.

Canonical Maximum Likelihood

In the multivariate–multiparameter case, canonical maximum likelihood (CML) also known as maximum pseudo-likelihood estimator (MPE) method is a general estimation technique (Genest et al. 1995; Kojadinovic and Jun Yan 2011). For example, the parameters of the nested 3-copula families (Table 10.2) (θ_1 and θ_2) may be estimated using the CML method. This method performs a nonparametric estimation of the margins by using the respective scaled ranks. The dependence parameters θ_1 and θ_2 are obtained by simply maximizing the log-likelihood function given by:

$$l(\theta) = \sum_{i=1}^{n} \log\left[c_\theta\left(u_i^1, u_i^2, \ldots, u_i^d\right)\right] \tag{10.49}$$

where c_θ denotes the density of the copula C_θ and $u_i^k = \hat{F}_k\left(X_{ik}\right)$ for $k = 1, 2, \ldots, d$ is the rank-based nonparametric marginal probability of k^{th} variable given by:

$$\hat{F}_k(X_{ik}) = \frac{1}{n+1} \sum_{i=1}^{n} \Im(X_{ik} \le x) \tag{10.50}$$

where $\Im(\bullet)$, as defined before, is the indicator function that takes a value 1 if the argument $\bullet$ is true and 0 if it is false.

(b) Selection of best-fit copula: Aforementioned procedure of parameter estimation is carried out for all tentatively selected copulas. Among different alternatives, the best-fit copula can be selected using the steps explained in Sect. 10.9.

(C) Probabilistic prediction of dependent variable: The joint distribution is obtained using the best-fit copula. If the best-fit copula is C, then the joint distribution is obtained through Sklar's theorem (Eq. 10.1) as follows:

$$F_{X,Y}(x, y) = C(F_X(x), F_Y(y)) \tag{10.51}$$

The probabilistic estimation is carried out by employing the conditional distribution. In general, the conditional distribution is obtained from joint distribution as follows:

$$\left.\begin{aligned} f_{X/Y}(x\big/Y = y) &= \frac{f_{X,Y}(x, y)}{f_Y(y)} \\ F_{X/Y}(x\big/Y = y) &= \frac{\int\limits_{-\infty}^{x} f_{X,Y}(u, y)du}{f_Y(y)} \\ F_{X/Y}(x\big/Y \le y) &= \frac{F_{X,Y}(x, y)}{F_Y(y)} \end{aligned}\right\} \tag{10.52}$$

where $f_{X/Y}(x\big/Y = y)$ is the conditional *pdf*. $F_{X/Y}(x\big/Y = y)$ and $F_{X/Y}(x\big/Y \le y)$ conditional *CDFs*, respectively. These expressions become relatively easier to execute through copula function. The conditional distribution function of U (i.e., $F_X(x)$) given $V = v$ (i.e., $F_Y(y)$) can be expressed in terms of copulas as:

$$C_{U/V=v} = \frac{\partial}{\partial v} C(u, v)\bigg|_{V=v} \tag{10.53}$$

Similarly, the conditional distribution function of V (i.e., $F_Y(y)$) given $U = u$ (i.e., $F_X(x)$) can be expressed in terms of copulas as:

$$C_{V/U=u} = \frac{\partial}{\partial u} C(u, v)\bigg|_{U=u} \tag{10.54}$$

The conditional distribution function of U given $V \leq v$ can be expressed in terms of copula as:

$$C_{U/V\leq v} = \frac{C(u, v)}{v} \tag{10.55}$$

Similarly, the conditional distribution function of V given $U \leq u$ can be expressed in terms of copula as:

$$C_{V/U\leq u} = \frac{C(u, v)}{u} \tag{10.56}$$

Depending on the condition, the respective equation for conditional distribution function can be used. Different probabilistic assessment can be done using this conditional distribution. For example, the expected value (EV) of the target variable can be obtained from the 50th quantile value of the distribution. Assessment of range of uncertainty can be obtained from different quantile values. For instance, 95% confidence interval can be obtained from 2.5th quantile (used as lower limit (LL)) and 97.5th quantile (used as upper limit (UL)).

Example 10.10.4
Using the joint distribution obtained in Example 10.9.1, find the most expected and 95% confidence interval of the temperature in city B if the temperature (in °C) in city A for 4 different days are 25, 22, 15.5, and 19, respectively.

Solution Using the marginal for temperature of city A (Example 10.8.3), the temperature of 4 different days can be converted to their reduced variate (say u). Hence, u is given by,

$$u = \begin{bmatrix} 0.997 & 0.952 & 0.229 & 0.711 \end{bmatrix}$$

According to the Example 10.10.3, the conditional distribution for temperature of city B given the temperature of city A is given by (Eq. 10.54),

$$F_{T_B/T_A}(t_B/T_A = t_A) = \begin{cases} u^{-1.5}\left(u^{-0.5} + v^{-0.5} - 1\right)^{-3} & \text{for } \left(u^{-0.5} + v^{-0.5} - 1\right) > 0 \\ 0 & \text{otherwise} \end{cases}$$

For the most expected temperature of city B, the conditional distribution is solved for being equal to 0.5 using the values of u. Hence, the most expected values of reduced variates of temperature of city B (say v) are given by,

$$\begin{aligned}&0.997^{-1.5}\left(0.997^{-0.5}+v^{-0.5}-1\right)^{-3}=0.5\\&v=0.630\end{aligned}$$

Similarly, $v_{0.5}$ can be obtained for all other values of u.

$$v_{0.5}=\begin{bmatrix}0.630 & 0.623 & 0.420 & 0.584\end{bmatrix}$$

Similarly, the 95% confidence interval for v can be obtained, and the conditional distribution is solved for 97.5% and 2.5% probabilities.

$$v_{0.975}=\begin{bmatrix}0.983 & 0.983 & 0.966 & 0.980\end{bmatrix}$$

$$v_{0.025}=\begin{bmatrix}0.085 & 0.083 & 0.027 & 0.067\end{bmatrix}$$

These v values can be back-transformed into temperature for city B, as temperature for city B follows normal distribution with mean 22 °C and standard deviation 4.2 °C (Example 10.8.3).

$$\begin{aligned}T_{B\ (0.025)}&=\begin{bmatrix}16.24 & 16.18 & 13.91 & 15.71\end{bmatrix}\\T_{B\ (0.50)}&=\begin{bmatrix}23.39 & 23.33 & 21.15 & 22.89\end{bmatrix}\\T_{B\ (0.975)}&=\begin{bmatrix}30.90 & 30.90 & 29.67 & 30.63\end{bmatrix}\end{aligned}$$

10.11 MATLAB Example

The examples discussed in this chapter can be solved using MATLAB. Some of the important built-in functions in this regard are following:

- `tau=corr(X,Y,'type','kendall')`
  This built-in function returns Kendall's tau (τ) between X and Y.
- `theta = copulaparam(family,tau)`
  This built-in function gives the value of `theta` for selected copula. The parameter `family` can either be `Clayton`, `Gumbel`, or `Frank`. The parameter `tau` is Kendall's tau (τ).
- `paramhat = copulafit(family,u)`
  This built-in function is used to fit a copula family (either of `Gaussian`, `t`, `Clayton`, `Gumbel` or `Frank`) over the data `u`. `u` must be a $n \times 2$ matrix, where n is number of observations. `paramhat` is estimate of parameter for selected copula model.
- `y = copulacdf(family,u,theta)`
  The `copulacdf` function is used to calculate the *CDF* for data set `u` using the specified copula and `theta`.

For instance, the following script (Box 10.1) can be used to solve Example 10.5.1 and associated examples (Examples 10.8.3, 10.9.1, 10.10.3, and 10.10.4).

Box 10.1 Sample MATLAB script for solving Example 10.5.1 and associated examples

```
close all; clear; clc;
T_A=[18.1,22.3,18.7,17.5,24.5];
T_B=[23.3,26.0,25.5,30.0,28.2];

%% Measure of scale-free association and reduced variates
tau=corr(T_A',T_B','type','kendall');
spearman_rho=corr(T_A',T_B','type','spearman');
u_val=normcdf(T_A,17.5,2.7);
v_val=normcdf(T_B,22,4.2);

%% Fitting Clayton and Gumbel Copula
clayton_theta=copulaparam('clayton',tau);
C_clayton=copulacdf('clayton',[u_val' v_val'],clayton_theta)';
gumbel_theta=copulaparam('gumbel',tau);
C_gumbel=copulacdf('gumbel',[u_val' v_val'],gumbel_theta)';

%% Empirical Copula
C_emp=zeros(size(C_gumbel));
for i=1:length(u_val)
    C_emp(i)=sum((u_val<=u_val(i)).*(v_val<=v_val(i)))/length(
        u_val);
end

%% Goodness-of-fit for two copula
S_n(1)=sum((C_emp-C_clayton).^2);
S_n(2)=sum((C_emp-C_gumbel).^2);
T_n(1)=sqrt(5)*max(abs((C_emp-C_clayton)));
T_n(2)=sqrt(5)*max(abs((C_emp-C_gumbel)));
[~, S_n_min_index]=min(S_n);
[~, T_n_min_index]=min(T_n);
switch S_n_min_index
    case 1
        best_copula="Clayton";
    case 2
        best_copula="Gumbel-Hougaard";
end

%% Data generation using Clayton copula
% u_val_random=rand(10,1)  % Generate random data, however we are
% using pre-generated random data
u_val_random=[0.93, 0.69, 0.05, 0.18, 0.19, 0.75, 0.85, 0.36,
    0.83, 0.59];
syms u v;
[~,copula_cdf]=clayton_copula(u,v,clayton_theta);
cond_prob_v_given_u=diff(copula_cdf,u);
v_val_generated=zeros(size(u_val_random));
for i=1:length(u_val_random)
    v_val_generated(i)=eval(solve(cond_prob_v_given_u(u_val_random
        (i),v)-0.5));
end
T_A_random=norminv(u_val_random,17.5,2.7);
T_B_generated=norminv(v_val_generated,22,4.2);

%% Prediction using Clayton Copula - Problem 1
T_A_observed=[25, 22, 15.5, 19];
u_val_observed=normcdf(T_A_observed,17.5,2.7);
```

```
v_val_expected=zeros(size(u_val_observed));
v_val_ll=v_val_expected;v_val_ul=v_val_expected;
for i=1:length(v_val_expected)
    v_val_ll(i)=eval(solve(cond_prob_v_given_u(u_val_observed(i),v)-0.025));
    v_val_expected(i)=eval(solve(cond_prob_v_given_u(u_val_observed(i),v)-0.5));
    v_val_ul(i)=eval(solve(cond_prob_v_given_u(u_val_observed(i),v)-0.975));
end
T_B_ll=norminv(v_val_ll,22,4.2);
T_B_expected=norminv(v_val_expected,22,4.2);
T_B_ul=norminv(v_val_ul,22,4.2);

%% Display Results
output_file=['output' filesep() 'code_1_result.txt'];
delete(output_file);diary(output_file);diary on;
fprintf("tau=%2.2f\t Spearman's rho=%2.2f\n", tau, spearman_rho);
fprintf("v")
disp(v_val)
fprintf("u")
disp(u_val)

fprintf('\n\nGOF statistic\n')
disp('First row correspond to Clayton and other correspond to Gumbel-Hougaard')
fprintf("\tS_n\tT_n\n")
disp([S_n' T_n'])
fprintf("The selected copula is %s.\n", best_copula);

fprintf("\n\nRandom Data Generation\n")
fprintf("Random u\t");disp(u_val_random);
fprintf("Generated v\t");disp(v_val_generated);
fprintf("Random T_A\t");disp(T_A_random);
fprintf("Generated T_B\t");disp(T_B_generated);

fprintf("\n\nPrediction for temperature of city B\n")
fprintf("u\t");disp(u_val_observed);
fprintf("v_0.025");disp(v_val_ll);
fprintf("Expected v\t");disp(v_val_expected);
fprintf("v_0.975");disp(v_val_ul);
fprintf("T_B_0.025\t");disp(T_B_ll);
fprintf("Expected T_B\t");disp(T_B_expected);
fprintf("T_B_0.975\t");disp(T_B_ul);

diary off;
```

Here, it should be noted that in the script, random values of u are generated and used, as done in Example 10.10.3. The output of Box 10.1 is shown in Box 10.2. The results match with the solution obtained in the corresponding examples. Similar scripts can be used to solve other examples in this chapter.

Box 10.2 Results for Box 10.1

```
tau=0.20    Spearman's rho=0.00
v     0.6215     0.8295     0.7977     0.9716     0.9301

u     0.5879     0.9623     0.6716     0.5000     0.9952
```

```


GOF statistic
First row correspond to Clayton and other correspond to Gumbel-
     Hougaard
   S_n T_n
     0.2067     0.6482
     0.2270     0.6593

The selected copula is Clayton.


Random Data Generation
Random u       0.9300     0.6900     0.0500     0.1800     0.1900
     0.7500     0.8500     0.3600     0.8300     0.5900

Generated v      0.6205     0.5801     0.2139     0.3845     0.3924
     0.5916     0.6085     0.4868     0.6053     0.5583

Random T_A      21.4846    18.8388    13.0589    15.0285    15.1297
     19.3211    20.2984    16.5322    20.0762    18.1144

Generated T_B     23.2881    22.8494    18.6690    20.7668    20.8535
       22.9729    23.1570    21.8614    23.1221    22.6155



Prediction for temperature of city B
u     0.9973     0.9522     0.2294     0.7107

v_0.025     0.0853     0.0826     0.0273     0.0668

Expected v       0.6296     0.6236     0.4202     0.5842

v_0.975     0.9832     0.9829     0.9655     0.9802

T_B_0.025     16.2457    16.1706    13.9274    15.6983

Expected T_B      23.3894    23.3224    21.1543    22.8934

T_B_0.975     30.9284    30.8898    29.6390    30.6426
```

Exercise

10.1 Check whether the following functions are valid copula functions or not?

(a) $C(u, v) = \frac{uv}{u-v+uv}$

(b) $C(u, v) = \sqrt{\max(u^2 + v^2 - 1, 0)}$

(c) $C(u, v) = \max(e^{\ln(u)} + e^{\ln(v)} - 1, 0)$

(d) $C(u, v) = \frac{\sqrt{u^2+v^2}}{2}$

(e) $C(u, v) = |u + v - 1|$

(f) $C(u, v) = \frac{uv}{1-0.3(1-u)(1-v)}$

(Ans. Only (d) and (e) are not copula functions.)

10.2 Using the first 6 values of total monthly precipitation depth and mean monthly specific humidity in Table A.1 (p. 429), calculate the Kendall's tau and Spearman's rho.
(Ans. Kendall's $\tau = 0.20$, Spearman's rho $(\rho_s) = 0.43$)

10.3 A location is frequently hit by cyclone. For the location, in a cyclonic event, the total rainfall depth and maximum pressure difference between eye and periphery of cyclone are assumed to be associated. For last six cyclones, the maximum pressure difference (in millibar) between the eye and periphery and total rainfall received (in cm) are (30, 70), (35, 77), (27, 75), (32, 81), (37, 87), and (25, 70). Calculate the Kendall's tau and Spearman's rho for the data set.
(Ans. Kendall's $\tau = 0.69$, Spearman's rho $(\rho_s) = 0.84$)

10.4 For the data used in Exercise 10.2, total monthly precipitation depth is distributed exponentially with mean 82 mm and the specific humidity is distributed normally with mean 10.7 and standard deviation of 2. Fit following copulas to the data:

(a) Independent Copula
(b) Gaussian Copula
(c) Clayton Copula
(d) Frank Copula

(Hint: Numerical solution may be needed for fitting some of the copula functions.)

Ans. (a) For independent copula, $C(u, v) = \begin{bmatrix} 0.001 & 0 & 0 & 0 & 0.001 & 0.970 \end{bmatrix}$
(b) For Gaussian copula, $\rho = 0.994$, $C(u, v) = \begin{bmatrix} 0.01 & 0 & 0 & 0 & 0.02 & 0.97 \end{bmatrix}$
(c) For Clayton copula, $\theta = 0.5$, $C(u, v) = \begin{bmatrix} 0.006 & 0 & 0 & 0 & 0.009 & 0.970 \end{bmatrix}$
(d) For Frank copula, $\theta = 1.86$, $C(u, v) = \begin{bmatrix} 0.001 & 0 & 0 & 0 & 0.002 & 0.970 \end{bmatrix}$

10.5 For the data given in Exercise 10.3, assume that the pressure difference (in millibar) between the eye and periphery of cyclone follows normal distribution with mean 30 and standard deviation 3.2. Similarly, the rainfall is gamma distributed with $\alpha = 35$ and $\beta = 2.5$. Fit following copula functions over the data,

(a) Frank Copula
(b) Gumbel–Hougaard Copula
(c) Clayton Copula
(d) Ali–Mikhail–Haq Copula

Ans. (a) For Frank copula, $\theta = 10.968$,
$C(u, v) = \begin{bmatrix} 0.1112 & 0.2471 & 0.1301 & 0.3455 & 0.5089 & 0.0375 \end{bmatrix}$
(b) For Gumbel–Hougaard copula, $\theta = 3.226$,
$C(u, v) = \begin{bmatrix} 0.1103 & 0.2471 & 0.1253 & 0.3446 & 0.5090 & 0.0422 \end{bmatrix}$
(c) For Clayton copula, $\theta = 4.45$,
$C(u, v) = \begin{bmatrix} 0.1121 & 0.2471 & 0.1588 & 0.3447 & 0.5086 & 0.0583 \end{bmatrix}$
(d) Ali–Mikhail–Haq copula cannot be fitted over the data.

10.6 From historical records, the daily rainfall depths (in mm) in two nearby cities A and B are found to have Kendall's tau as 0.45. Fit Gumbel–Hougaard, Ali–Mikhail–Haq, and Clayton copulas between the daily rainfall depth of two cities.

(Ans. For Gumbel–Hougaard $\theta = 1.818$. For Clayton copula $\theta = 1.636$. Ali–Mikhail–Haq copula can not be fitted on the data.)

10.7 Select the best copula function for Example 10.4 using Kolmogorov–Smirnov statistic and Cramér-von Mises statistic. (Ans. Gaussian copula)

10.8 For a location A, monthly mean potential evaporation follows an exponential distribution with mean 4 mm/day, and mean monthly air temperature follows normal distribution with mean 25 °C and standard deviation 3.7 °C. Clayton copula with $\theta = 0.7$ is found to be the best-fit copula. Generate the mean potential evaporation and mean monthly temperature values for a year.

(Answers may vary depending on random number generated. Refer to Sect. 10.10.1.)

10.9 Using the Frank copula fitted in Exercise 10.6, predict the daily rainfall depth (in mm) for city A, if the daily rainfall depths (in mm) for city B recorded in last week are 0, 2, 5, 20, 8, 0, and 3. Also, calculate 90% confidence interval for predictions. Assume that daily rainfall in cities A and B follows exponential distribution with mean 4.5 and 3 mm/day, respectively.
(Ans. Expected rainfall (in mm/day) for city $A = \begin{bmatrix} 1.7 & 3.1 & 4.3 & 5.2 & 4.9 & 1.7\ 3.6 \end{bmatrix}$ and corresponding 90% confidence interval are (0.1, 9.7), (0.2, 13.0), (0.4, 15.2), (0.5, 16.6), (0.5, 16.1), (0.1, 9.7), and (0.3, 14.0).)

10.10 For the last 5 months due to some technical problem at site A (Exercise 10.8), evaporation was not recorded. However, if the observed mean monthly air temperature (in °C) for last 5 months is 24, 28, 30, 32, and 27, then using the copula function given in Exercise 10.8, calculate the expected value of mean monthly evaporation for these month along with their interquartile range (25–75% range).
(Ans. Expected evaporation (in mm/day) is 2.7, 3.9, 4.2, 4.32, and 3.7. The interquartile range of evaporation is (1.3, 5.3), (2.0, 6.9), (2.2, 7.2), (2.3, 7.4), and (1.9, 6.6).)

10.11 In the Table A.1 (p. 429), model the association of the total precipitation depth and mean monthly pressure using Frank copula. Assume that pressure is distributed normally and total monthly rainfall follows exponential distribution with mean 95 mm. Furthermore, predict the total monthly rainfall depth and 95% confidence interval if the mean monthly pressure is 960 mb.
(Hint: Numerical solution may be needed for fitting some of the copula function.)
(Ans. For Frank copula $\theta = -5$. The expected total monthly rainfall depth for 960 mb pressure is 50 mm and its 95% confidence interval is 6.2 and 182.3 mm.)

10.12 In the Table A.2 (p. 431), the locations A1 and A2 are 50 km apart, and their monthly mean sea surface temperature is assumed to associated. Assuming that SST at both the places are normally distributed, fit a Clayton copula to model the relationship of SST between these two places. Using the fitted copula, generate the monthly mean sea surface temperature for one year.
(Hint: Numerical solution may be needed for fitting some of the copula function.)
(Ans. For Clayton copula $\theta = 28.21$.)

References

Bosq, Denis. 2012. *Nonparametric statistics for stochastic processes: Estimation and prediction*, vol. 110. New York: Springer Science & Business Media.

Genest, Christian, and Jock MacKay. 1986. The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician* 40 (4): 280–283.

Genest, Christian, and Louis-Paul Rivest. 1993. Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association* 88 (423): 1034–1043.

Genest, Christian, and Anne-Catherine Favre. 2007. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering* 12 (4): 347–368.

Genest, Christian, Kilani Ghoudi, and L.-P. Rivest. 1995. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* 82 (3): 543–552.

Genest, Christian, Bruno Rémillard, and David Beaudoin. 2009. Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics* 44 (2): 199–213.

Joe, Harry. 1997. *Multivariate models and multivariate dependence concepts*. Boca Raton: CRC Press.

Kojadinovic, Ivan, and Jun Yan. 2011. A goodness-of-fit test for multivariate multiparameter copulas based on multiplier central limit theorems. *Statistics and Computing* 21 (1): 17–30.

Nelsen, Roger B. 1999. *An introduction to copulas*, 269. Berlin: Springer.

Salvadori, G., and C. De Michele. 2007. On the use of copulas in hydrology: Theory and practice. *Journal of Hydrologic Engineering* 12 (4): 369–380.

Sklar, A. 1959. Fonctions de réepartition à n dimensions et leurs marges. Publications de l'Institut de Statistique de l'Universitée de Paris.

Appendix A
Data Set

See Tables A.1, A.2, A.3, A.4 and A.5.

Table A.1 Different hydrocliamtic variables for Upper Mahanadi Basin for 24 consecutive months (January, 1980–December, 1981)

Precipitation	Surface air temperature	Precipitable water	Pressure	Air temperature*	Specific humidity*	Geo-potential height*	Zonal wind^{+}	Meridional wind^{+}
6.33	21.49	17.76	964.42	20.96	5.64	802.36	0.21	0.09
0.86	23.27	15.04	962.91	23.27	4.48	791.30	1.02	–0.84
8.67	28.47	18.56	961.09	27.80	4.93	779.47	1.51	–0.32
7.31	33.78	20.81	957.42	33.06	5.40	750.21	2.38	–0.91
4.62	35.88	24.74	954.30	35.40	6.55	722.29	3.33	–1.20
287.65	27.90	52.12	951.67	27.02	16.91	692.85	2.76	0.47
387.35	24.66	56.61	952.05	23.44	17.51	694.13	4.88	–1.28
229.80	24.58	54.82	953.22	23.31	17.32	704.62	4.27	–0.84
404.99	24.48	44.21	956.97	23.66	15.38	739.06	2.51	–3.03
25.40	24.23	27.93	961.14	23.58	11.80	776.56	–1.01	–1.82
0.00	22.37	20.41	964.18	21.94	7.43	802.34	–2.13	–0.89
6.56	21.16	17.51	963.93	20.55	5.84	797.93	–1.14	–0.63
12.72	19.43	16.16	965.22	18.69	5.41	805.18	0.26	–0.08
0.45	24.45	13.13	963.36	23.93	3.83	794.97	0.60	–0.53
33.26	27.85	20.27	961.46	26.87	5.51	781.94	1.02	0.26
0.76	32.67	16.74	957.17	32.31	4.75	747.79	1.91	–0.71
5.99	35.27	29.69	953.51	34.11	8.23	715.36	1.48	0.63
79.17	33.34	39.07	950.13	31.96	11.05	681.90	3.57	–0.91
236.62	24.77	53.89	952.40	23.55	17.55	696.73	4.56	0.30
420.82	24.51	53.97	952.28	23.34	17.38	696.11	4.14	–0.83
231.00	24.64	47.04	956.70	23.38	16.24	737.47	–0.19	0.20
13.24	24.32	30.74	961.48	23.63	11.07	779.48	–0.75	–1.46
1.06	21.74	19.71	963.00	21.37	7.60	790.59	–1.43	–1.71
2.70	20.17	17.24	966.26	19.47	5.82	817.45	–1.82	–1.27

*measured at 925 mb
$^{+}$measured at 200 mb

R. Maity, *Statistical Methods in Hydrology and Hydroclimatology*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-10-8779-0

Table A.2 Monthly average sea surface temperature for Arabian Sea (near Mumbai) for 24 consecutive months (January, 2011–December, 2012)

Locations*																								
A1	B1	C1	D1	E1	A2	B2	C2	D2	E2	A3	B3	C3	D3	E3	A4	B4	C4	D4	E4	A5	B5	C5	D5	E5
25.0	25.2	25.4	25.6	25.6	25.3	25.5	25.9	26.4	26.7	25.6	25.9	26.4	26.9	27.3	26.0	26.3	26.7	27.2	27.7	26.5	26.8	27.2	27.6	27.9
24.5	24.6	24.8	24.9	24.9	25.0	25.1	25.5	25.9	26.1	25.5	25.7	26.1	26.6	27.0	26.0	26.2	26.6	27.1	27.5	26.5	26.8	27.2	27.5	27.9
25.1	25.2	25.3	25.4	25.3	25.5	25.7	25.9	26.2	26.4	26.0	26.1	26.4	26.8	27.2	26.4	26.6	27.0	27.3	27.7	26.8	27.1	27.5	27.8	28.1
27.3	27.4	27.4	27.4	27.4	27.8	27.8	27.9	28.0	28.1	28.1	28.2	28.3	28.5	28.7	28.4	28.5	28.7	28.9	29.1	28.7	28.9	29.0	29.2	29.5
29.0	29.1	29.1	29.0	28.9	29.3	29.4	29.4	29.4	29.3	29.5	29.6	29.6	29.6	29.7	29.6	29.7	29.7	29.8	29.8	29.7	29.8	29.8	29.9	29.9
28.8	29.1	29.4	29.4	29.4	28.8	29.1	29.3	29.3	29.3	28.8	29.1	29.2	29.2	29.2	28.7	28.9	29.1	29.1	29.1	28.7	28.9	29.0	29.1	29.1
27.3	27.9	28.4	28.7	28.8	27.1	27.7	28.1	28.4	28.5	27.0	27.5	27.9	28.1	28.3	27.0	27.5	27.9	28.1	28.2	27.3	27.7	28.0	28.2	28.3
26.3	26.9	27.4	27.7	27.8	26.2	26.7	27.2	27.5	27.7	26.2	26.7	27.1	27.4	27.7	26.5	26.9	27.3	27.6	27.8	26.9	27.4	27.8	28.0	28.2
27.2	27.8	28.3	28.6	28.7	27.0	27.5	28.0	28.3	28.6	26.8	27.3	27.8	28.2	28.4	26.9	27.4	27.8	28.2	28.4	27.2	27.7	28.1	28.4	28.6
28.5	28.6	28.8	28.9	29.1	28.4	28.6	28.7	28.9	29.1	28.4	28.5	28.6	28.8	29.0	28.3	28.5	28.6	28.8	29.0	28.4	28.6	28.7	28.9	29.0
27.7	27.8	27.9	28.1	28.2	27.8	27.9	28.1	28.4	28.6	27.8	28.0	28.2	28.5	28.8	27.9	28.1	28.3	28.7	29.0	28.1	28.3	28.5	28.8	29.1
26.5	26.6	26.7	26.9	26.9	26.7	26.8	27.1	27.5	27.8	27.0	27.1	27.5	27.9	28.4	27.3	27.5	27.8	28.3	28.7	27.7	27.9	28.2	28.6	29
25.0	25.1	25.3	25.4	25.5	25.4	25.5	25.9	26.3	26.6	25.9	26.1	26.4	26.9	27.4	26.4	26.6	26.9	27.4	27.9	26.9	27.1	27.5	27.8	28.2
24.7	24.9	25.1	25.3	25.4	25.1	25.3	25.7	26.2	26.5	25.6	25.8	26.3	26.9	27.3	26.1	26.4	26.9	27.4	27.9	26.7	27.0	27.4	27.9	28.2
26.1	26.3	26.5	26.6	26.6	26.6	26.8	27.1	27.4	27.6	27.0	27.3	27.6	28.0	28.3	27.5	27.7	28.1	28.4	28.7	27.9	28.2	28.5	28.8	29.1
27.8	27.9	28.0	28.0	27.9	28.2	28.4	28.6	28.7	28.7	28.7	28.8	29.0	29.1	29.2	29.0	29.2	29.3	29.5	29.6	29.3	29.4	29.6	29.7	29.8
29.3	29.6	29.6	29.6	29.5	29.7	29.9	30.0	30.0	29.9	30.0	30.1	30.2	30.2	30.1	30.1	30.2	30.2	30.2	30.2	30.0	30.1	30.2	30.2	30.2
28.6	29.1	29.4	29.5	29.5	28.5	29.0	29.3	29.5	29.5	28.4	28.8	29.1	29.3	29.4	28.3	28.7	29.0	29.1	29.2	28.3	28.6	28.9	29.0	29.1
26.6	27.4	28.0	28.3	28.4	26.2	27.0	27.6	27.9	28.1	26.1	26.8	27.3	27.6	27.9	26.1	26.7	27.3	27.6	27.8	26.4	27.0	27.5	27.8	27.9
25.6	26.4	27.1	27.6	27.8	25.2	26.0	26.7	27.1	27.4	25.1	25.8	26.5	26.9	27.2	25.3	26.0	26.6	27.0	27.2	25.8	26.4	27.0	27.3	27.5
27.0	27.5	28.0	28.2	28.4	26.6	27.2	27.7	28.0	28.2	26.5	27.0	27.5	27.8	28.1	26.5	27.1	27.5	27.8	28.0	26.9	27.4	27.8	28.0	28.1
28.4	28.6	28.6	28.6	28.6	28.3	28.4	28.5	28.7	28.7	28.2	28.4	28.5	28.6	28.8	28.2	28.3	28.4	28.6	28.7	28.2	28.3	28.5	28.6	28.7
27.8	27.9	28.0	28.2	28.2	27.9	28.0	28.2	28.4	28.6	28.0	28.1	28.3	28.6	28.8	28.1	28.2	28.4	28.7	28.9	28.3	28.4	28.6	28.8	29
26.0	26.2	26.4	26.6	26.6	26.2	26.4	26.8	27.2	27.4	26.4	26.7	27.1	27.5	27.9	26.7	27.0	27.3	27.8	28.2	27.1	27.4	27.7	28.1	28.4

*Locations are gridded—A to E shows different latitude, and 1–5 shows different longitude

Table A.3 Monthly minimum or maximum temperature (January, 2008–December, 2010) for Bhadra Reservoir and Holehonnur* in Bhadra Basin

Months	Bhadra Reservoir		Holehonnur station	
	Max temperature	Min temperature	Max temperature	Min temperature
1	30.79	15.75	32.67	21.43
2	30.68	17.75	32.68	21.99
3	31.41	19.10	28.65	20.81
4	32.42	21.11	28.08	20.94
5	31.96	21.36	26.92	20.22
6	27.82	20.83	27.97	21.03
7	27.06	20.61	30.62	21.30
8	26.14	20.26	30.16	18.05
9	27.83	20.22	29.87	18.41
10	29.72	20.33	30.27	15.40
11	30.12	17.97	32.40	17.52
12	29.79	17.85	33.64	20.78
13	30.24	15.55	34.28	22.44
14	31.99	16.99	33.04	22.17
15	33.21	19.84	29.38	21.21
16	33.75	21.71	26.14	21.02
17	32.54	21.91	27.75	21.41
18	28.93	21.03	28.32	20.72
19	25.88	20.46	29.57	20.63
20	27.49	21.00	30.25	20.78
21	26.99	20.37	30.08	18.70
22	28.47	20.11	30.08	18.65
23	29.58	19.88	31.45	16.88
24	29.98	18.42	35.23	19.67
25	29.89	18.36	35.36	22.14
26	31.25	16.92	33.89	22.73
27	33.62	19.84	30.35	22.58
28	34.07	21.78	27.63	21.97
29	33.13	22.44	27.21	21.84
30	29.89	21.79	27.61	21.02
31	27.68	20.70	29.26	20.75
32	26.52	20.55	28.56	21.01
33	26.24	20.44	29.10	19.56
34	28.99	20.57	28.93	18.74
35	28.23	20.04	30.85	20.63
36	28.67	18.75	30.23	18.41

*Holehonnur is a town 50 km downstream from Bhadra Reservoir

Table A.4 Monthly average volumetric soil moisture content* in Upper Mahandai Basin for 26 years

Year	January	February	March	April	May	June	July	August	September	October	November	December
1981	268.4	225.8	178.6	142.4	115.3	327.0	567.4	627.4	605.5	518.1	422.7	373.4
1982	337.3	282.4	240.8	202.7	171.3	204.4	352.9	511.9	583.5	523.1	418.8	367.4
1983	348.1	350.9	321.9	260.8	211.6	233.4	354.4	539.7	572.7	495.1	413.2	348.5
1984	299.8	263.1	207	163.3	145.4	179.0	285.6	431.9	524	510.0	432.8	382.5
1985	355.0	328.2	261.1	204.8	159.3	208	414.6	587.2	569.7	458.9	369.0	315.3
1986	284.0	256.2	200.7	153.1	117.4	136.8	291.3	488.2	574.6	564.3	485.5	414.1
1987	376.8	353.8	302.9	248.7	202.0	318.8	488.4	580.4	578.4	491.9	427.9	389.6
1988	351.7	307.6	262.7	195.6	156.7	165.6	308.8	436.1	452.7	443.6	414.7	374.3
1989	325.1	288.4	240.5	186.1	138.8	190.1	286.9	371.7	415.5	361.4	296.6	252.7
1989	218.7	183.7	154.5	115.3	86.33	229.0	429.6	554.0	579.8	475.5	380.9	342.7
1990	297.0	269.1	229.4	185.2	193.3	309.9	453.7	551.2	607.1	608.2	528.1	448.1
1991	387.2	318.1	243.2	178.4	127.9	188.6	362	549.4	561.5	456.5	382.5	333.5
1992	290.0	246.7	194.6	157.0	137.0	154.4	281.1	503.3	582.4	487.4	392.9	335.6
1993	287.5	251.7	217.3	168.3	132.4	189.3	413.4	596.4	622.9	544.1	433.7	373.1
1994	325.8	289.4	227.3	179.4	155.2	284.8	545.8	651.2	641.7	570.5	465.7	399.7
1995	379.2	348.5	317.1	252.9	215.1	219.2	395.9	577.1	586.6	514.1	428.1	367.6
1996	327.2	290.4	241.3	199.2	151.3	161.5	311.2	505.0	527.3	457.6	385.7	329.9
1997	301.5	258.5	206.3	180.7	156.9	178.3	354.6	578.1	581.2	495.8	438.4	413.4
1998	396.7	358.6	311.9	252.3	195.9	194.0	261.7	369.3	456.7	456.4	421.9	376.3
1999	329.1	283.7	217.3	157.7	129.4	172.9	284.5	479.7	585.5	533.3	433.3	370.6
2000	320.3	288.8	230.1	168.7	148.9	259.4	441.8	495.9	458.2	365.4	291.0	246.6
2001	215.8	178.1	152.0	128.7	117.1	253.3	467.8	581.6	562.2	491.5	407.6	344.0
2002	310.1	268.6	224.2	166.3	135.9	211.7	298.8	412.8	474.9	408.3	327.8	273.6
2003	232.7	219.8	207.0	173.2	125.2	146.8	298.3	533.5	635.3	602.7	502.6	442.4
2004	414.3	363.4	297.5	236.4	194.7	270.8	427.0	515.3	531.1	476.9	388.6	332.1
2005	335.8	310.8	247.9	192.2	159.7	192.6	425.5	608.5	629.8	593.3	501.7	430.8

*The soil moisture content is expressed as (% volume by volume (% v/v) soil moisture content × 1000)

Table A.5 Daily soil moisture* data at a location since January 1, 2017

Days	0	1	2	3	4	5	6	7	8	9
0	0.0179	0.1157	0.3552	0.3363	0.1374	0.0798	0.1701	0.2167	0.1297	0.0959
1	0.1641	0.1222	0.0444	0.1326	0.0938	0.0443	0.0917	0.2501	0.1010	0.1474
2	0.1818	0.1298	0.1422	0.1699	0.2162	0.3209	0.2856	0.2008	0.1225	0.3459
3	0.3709	0.0882	0.1805	0.1251	0.2169	0.2009	0.1621	0.1234	0.2105	0.1754
4	0.3314	0.3665	0.0439	0.2603	0.1682	0.4307	0.1673	0.1552	0.1218	0.2349
5	0.0490	0.2135	0.3055	0.1900	0.3236	0.1797	0.0774	0.1162	0.4122	0.1960
6	0.3838	0.2368	0.1025	0.2296	0.1169	0.1199	0.2844	0.2500	0.2140	0.1634
7	0.0171	0.4016	0.2316	0.1554	0.1303	0.3428	0.0305	0.1393	0.3813	0.1853
8	0.1822	0.2350	0.3696	0.1001	0.2454	0.2072	0.1740	0.3359	0.1216	0.1097
9	0.2586	0.2536	0.3718	0.2753	0.0757	0.1131	0.2885	0.1540	0.0468	0.1662

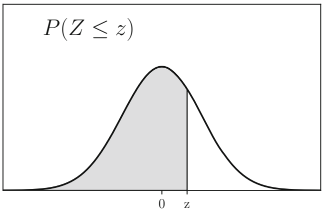
*The soil moisture content is expressed as volume by volume fraction

Appendix B
Statistical Tables

See Tables B.1, B.2, B.3, B.4, B.5, B.6, B.7 and B.8.

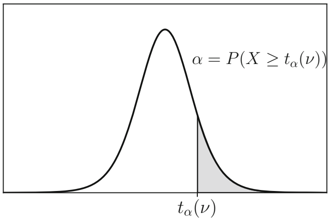
R. Maity, *Statistical Methods in Hydrology and Hydroclimatology*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-10-8779-0

Table B.1 Standard normal table



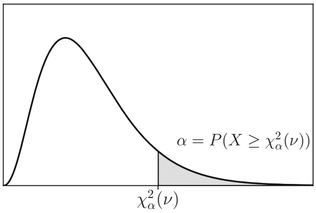
z	**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Table B.2 Student's t-distribution percentage points



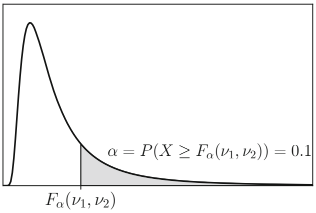
ν	α								
	0.25	0.10	0.05	0.025	0.01	0.00833	0.00625	0.005	0.0025
1	1.250	3.178	6.364	12.731	31.831	38.212	50.929	63.662	127.324
2	1.066	1.986	2.970	4.328	6.975	7.659	8.866	9.930	14.092
3	1.015	1.738	2.403	3.207	4.551	4.866	5.398	5.846	7.456
4	0.991	1.633	2.182	2.801	3.757	3.970	4.321	4.609	5.600
5	0.977	1.576	2.065	2.596	3.375	3.543	3.816	4.037	4.776
6	0.968	1.540	1.993	2.472	3.153	3.296	3.527	3.712	4.319
7	0.961	1.515	1.945	2.390	3.008	3.136	3.342	3.504	4.032
8	0.956	1.497	1.910	2.331	2.906	3.024	3.212	3.360	3.835
9	0.953	1.483	1.883	2.287	2.831	2.942	3.117	3.255	3.692
10	0.950	1.472	1.862	2.253	2.774	2.879	3.044	3.174	3.584
11	0.947	1.463	1.846	2.226	2.728	2.829	2.987	3.111	3.499
12	0.945	1.456	1.832	2.204	2.691	2.788	2.941	3.060	3.431
13	0.944	1.450	1.821	2.185	2.660	2.754	2.902	3.017	3.375
14	0.942	1.445	1.811	2.170	2.634	2.726	2.870	2.982	3.328
15	0.941	1.441	1.803	2.156	2.612	2.702	2.843	2.952	3.289
16	0.940	1.437	1.796	2.145	2.593	2.682	2.819	2.926	3.254
17	0.939	1.433	1.790	2.135	2.577	2.664	2.799	2.903	3.225
18	0.938	1.430	1.784	2.126	2.562	2.648	2.781	2.883	3.199
19	0.938	1.428	1.779	2.118	2.549	2.634	2.765	2.866	3.176
20	0.937	1.425	1.775	2.111	2.538	2.621	2.751	2.850	3.156
21	0.936	1.423	1.771	2.105	2.528	2.610	2.738	2.836	3.138
22	0.936	1.421	1.767	2.099	2.518	2.600	2.726	2.824	3.121
23	0.935	1.419	1.764	2.094	2.510	2.591	2.716	2.812	3.106
24	0.935	1.418	1.761	2.089	2.502	2.582	2.706	2.802	3.093
25	0.934	1.416	1.758	2.085	2.495	2.574	2.698	2.792	3.081
26	0.934	1.415	1.756	2.081	2.489	2.567	2.690	2.784	3.069
27	0.934	1.414	1.753	2.077	2.483	2.561	2.683	2.776	3.059
28	0.933	1.413	1.751	2.073	2.477	2.555	2.676	2.768	3.049
29	0.933	1.411	1.749	2.070	2.472	2.549	2.669	2.761	3.041
30	0.933	1.410	1.747	2.067	2.467	2.544	2.664	2.755	3.032
40	0.931	1.403	1.734	2.046	2.433	2.507	2.622	2.709	2.974
50	0.929	1.399	1.726	2.034	2.413	2.486	2.598	2.683	2.939
60	0.929	1.396	1.721	2.025	2.400	2.471	2.581	2.665	2.917
120	0.927	1.389	1.708	2.005	2.368	2.436	2.542	2.622	2.862
1000	0.925	1.382	1.696	1.987	2.340	2.406	2.508	2.586	2.816

Table B.3 χ^2 distribution percentage points



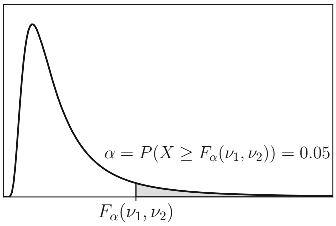
	α									
ν	**0.999**	**0.990**	**0.950**	**0.900**	**0.500**	**0.100**	**0.050**	**0.010**	**0.005**	**0.001**
1	0.000	0.000	0.004	0.016	0.455	2.706	3.841	6.635	7.879	10.828
2	0.002	0.020	0.103	0.211	1.386	4.605	5.991	9.210	10.597	13.816
3	0.024	0.115	0.352	0.584	2.366	6.251	7.815	11.345	12.838	16.266
4	0.091	0.297	0.711	1.064	3.357	7.779	9.488	13.277	14.860	18.467
5	0.210	0.554	1.145	1.610	4.351	9.236	11.070	15.086	16.750	20.515
6	0.381	0.872	1.635	2.204	5.348	10.645	12.592	16.812	18.548	22.458
7	0.598	1.239	2.167	2.833	6.346	12.017	14.067	18.475	20.278	24.322
8	0.857	1.646	2.733	3.490	7.344	13.362	15.507	20.090	21.955	26.125
9	1.152	2.088	3.325	4.168	8.343	14.684	16.919	21.666	23.589	27.877
10	1.479	2.558	3.940	4.865	9.342	15.987	18.307	23.209	25.188	29.588
11	1.834	3.053	4.575	5.578	10.341	17.275	19.675	24.725	26.757	31.264
12	2.214	3.571	5.226	6.304	11.340	18.549	21.026	26.217	28.300	32.910
13	2.617	4.107	5.892	7.042	12.340	19.812	22.362	27.688	29.819	34.528
14	3.041	4.660	6.571	7.790	13.339	21.064	23.685	29.141	31.319	36.123
15	3.483	5.229	7.261	8.547	14.339	22.307	24.996	30.578	32.801	37.697
16	3.942	5.812	7.962	9.312	15.338	23.542	26.296	32.000	34.267	39.252
17	4.416	6.408	8.672	10.085	16.338	24.769	27.587	33.409	35.718	40.790
18	4.905	7.015	9.390	10.865	17.338	25.989	28.869	34.805	37.156	42.312
19	5.407	7.633	10.117	11.651	18.338	27.204	30.144	36.191	38.582	43.820
20	5.921	8.260	10.851	12.443	19.337	28.412	31.410	37.566	39.997	45.315
21	6.447	8.897	11.591	13.240	20.337	29.615	32.671	38.932	41.401	46.797
22	6.983	9.542	12.338	14.041	21.337	30.813	33.924	40.289	42.796	48.268
23	7.529	10.196	13.091	14.848	22.337	32.007	35.172	41.638	44.181	49.728
24	8.085	10.856	13.848	15.659	23.337	33.196	36.415	42.980	45.559	51.179
25	8.649	11.524	14.611	16.473	24.337	34.382	37.652	44.314	46.928	52.620
26	9.222	12.198	15.379	17.292	25.336	35.563	38.885	45.642	48.290	54.052
27	9.803	12.879	16.151	18.114	26.336	36.741	40.113	46.963	49.645	55.476
28	10.391	13.565	16.928	18.939	27.336	37.916	41.337	48.278	50.993	56.892
29	10.986	14.256	17.708	19.768	28.336	39.087	42.557	49.588	52.336	58.301
30	11.588	14.953	18.493	20.599	29.336	40.256	43.773	50.892	53.672	59.703
35	14.688	18.509	22.465	24.797	34.336	46.059	49.802	57.342	60.275	66.619
40	17.916	22.164	26.509	29.051	39.335	51.805	55.758	63.691	66.766	73.402
45	21.251	25.901	30.612	33.350	44.335	57.505	61.656	69.957	73.166	80.077
50	24.674	29.707	34.764	37.689	49.335	63.167	67.505	76.154	79.490	86.661
55	28.173	33.570	38.958	42.060	54.335	68.796	73.311	82.292	85.749	93.168
60	31.738	37.485	43.188	46.459	59.335	74.397	79.082	88.379	91.952	99.607

Table B.4 F distribution percentage points $\alpha = 0.1$



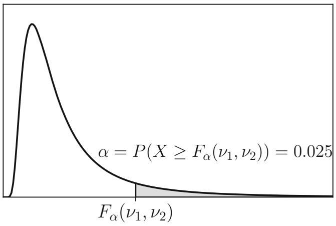
ν_1/ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	30	50	60
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.26	62.69	62.79
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.46	9.47	9.47
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.17	5.15	5.15
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.82	3.80	3.79
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.17	3.15	3.14
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.80	2.77	2.76
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.56	2.52	2.51
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.38	2.35	2.34
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.25	2.22	2.21
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.16	2.12	2.11
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.08	2.04	2.03
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.01	1.97	1.96
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.96	1.92	1.90
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.91	1.87	1.86
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.87	1.83	1.82
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.84	1.79	1.78
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.81	1.76	1.75
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.78	1.74	1.72
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.76	1.71	1.70
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.74	1.69	1.68
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.72	1.67	1.66
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.70	1.65	1.64
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.69	1.64	1.62
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.67	1.62	1.61
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.66	1.61	1.59
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.65	1.59	1.58
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.64	1.58	1.57
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.63	1.57	1.56
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.62	1.56	1.55
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.61	1.55	1.54
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.54	1.48	1.47
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.68	1.63	1.57	1.50	1.44	1.42
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.48	1.41	1.40
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.41	1.34	1.32
1000	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64	1.61	1.55	1.49	1.43	1.35	1.27	1.25

Table B.5 F distribution percentage points $\alpha = 0.05$



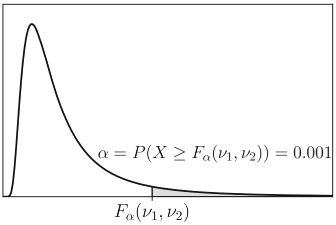
ν_1/ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	30	50	60
1	161	199	216	224	230	234	237	239	241	242	244	246	248	250	252	252
2	18.5	19.0	19.2	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.58	8.57
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.70	5.69
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.44	4.43
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.75	3.74
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.32	3.30
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.02	3.01
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.80	2.79
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.70	2.64	2.62
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.57	2.51	2.49
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.40	2.38
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.38	2.31	2.30
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.31	2.24	2.22
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.18	2.16
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.19	2.12	2.11
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.15	2.08	2.06
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.11	2.04	2.02
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.07	2.00	1.98
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.97	1.95
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.01	1.94	1.92
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	1.98	1.91	1.89
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	1.96	1.88	1.86
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.94	1.86	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.92	1.84	1.82
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.90	1.82	1.80
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.88	1.81	1.79
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.87	1.79	1.77
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.85	1.77	1.75
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.76	1.74
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.74	1.66	1.64
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.87	1.78	1.69	1.60	1.58
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65	1.56	1.53
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.55	1.46	1.43
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.68	1.58	1.47	1.36	1.33

Table B.6 F distribution percentage points $\alpha = 0.025$



ν_1/ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	30	50	60
1	648	799	864	900	922	937	948	957	963	969	977	985	993	1001	1008	1010
2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5
3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.2	14.2	14.1	14.0	14.0
4	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.46	8.38	8.36
5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.23	6.14	6.12
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.07	4.98	4.96
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.36	4.28	4.25
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.89	3.81	3.78
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.56	3.47	3.45
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.31	3.22	3.20
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.12	3.03	3.00
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	2.96	2.87	2.85
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.84	2.74	2.72
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.73	2.64	2.61
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.64	2.55	2.52
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.57	2.47	2.45
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.50	2.41	2.38
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.44	2.35	2.32
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.39	2.30	2.27
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.35	2.25	2.22
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.31	2.21	2.18
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.27	2.17	2.14
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.24	2.14	2.11
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.21	2.11	2.08
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.18	2.08	2.05
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.16	2.05	2.03
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.13	2.03	2.00
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.11	2.01	1.98
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.09	1.99	1.96
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.07	1.97	1.94
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	1.94	1.83	1.80
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.22	2.11	1.99	1.87	1.75	1.72
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.82	1.70	1.67
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.69	1.56	1.53
1000	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13	2.06	1.96	1.85	1.72	1.58	1.45	1.41

Table B.7 F distribution percentage points $\alpha = 0.001$



ν_1/ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	30	50	60
1	4052	4999	5403	5624	5763	5859	5928	5981	6022	6056	6106	6157	6209	6261	6302	6313
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.4	27.2	27.0	26.9	26.7	26.5	26.3	26.3
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.8	13.7	13.6
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.0	9.89	9.72	9.55	9.38	9.24	9.20
6	13.8	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23	7.09	7.06
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.99	5.86	5.82
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.20	5.07	5.03
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.65	4.52	4.48
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25	4.12	4.08
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	3.94	3.81	3.78
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.70	3.57	3.54
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.51	3.38	3.34
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.35	3.22	3.18
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.21	3.08	3.05
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.10	2.97	2.93
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.00	2.87	2.83
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.92	2.78	2.75
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.84	2.71	2.67
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.78	2.64	2.61
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.72	2.58	2.55
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.67	2.53	2.50
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.62	2.48	2.45
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.58	2.44	2.40
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.54	2.40	2.36
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.50	2.36	2.33
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.47	2.33	2.29
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.44	2.30	2.26
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.41	2.27	2.23
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.39	2.25	2.21
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.20	2.06	2.02
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.42	2.27	2.10	1.95	1.91
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.03	1.88	1.84
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.86	1.70	1.66
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.20	2.06	1.90	1.72	1.54	1.50

Table B.8 Kolmogorov–Smirnov two-sided test

n	*α*		
	0.10	0.05	0.01
1	0.9500	0.9750	0.9950
2	0.7764	0.8419	0.9293
3	0.6360	0.7076	0.8290
4	0.5652	0.6239	0.7342
5	0.5094	0.5633	0.6685
6	0.4680	0.5193	0.6166
7	0.4361	0.4834	0.5758
8	0.4096	0.4543	0.5418
9	0.3875	0.4300	0.5133
10	0.3687	0.4092	0.4889
11	0.3524	0.3912	0.4677
12	0.3382	0.3754	0.4490
13	0.3255	0.3614	0.4325
14	0.3142	0.3489	0.4176
15	0.3040	0.3376	0.4042
16	0.2947	0.3273	0.3920
17	0.2863	0.3180	0.3809
18	0.2785	0.3094	0.3706
19	0.2714	0.3014	0.3612
20	0.2647	0.2941	0.3524
21	0.2586	0.2872	0.3443
22	0.2528	0.2809	0.3367
23	0.2475	0.2749	0.3295
24	0.2424	0.2693	0.3229
25	0.2377	0.2640	0.3166
26	0.2332	0.2591	0.3106
27	0.2290	0.2544	0.3050
28	0.2250	0.2499	0.2997
29	0.2212	0.2457	0.2947
30	0.2176	0.2417	0.2899
31	0.2141	0.2379	0.2853
32	0.2108	0.2342	0.2809
33	0.2077	0.2308	0.2768
34	0.2047	0.2274	0.2728
35	0.2018	0.2242	0.2690
36	0.1991	0.2212	0.2653
37	0.1965	0.2183	0.2618
38	0.1939	0.2154	0.2584
39	0.1915	0.2127	0.2552
40	0.1891	0.2101	0.2521
>40	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.63/\sqrt{n}$

Index

© Springer Nature Singapore Pte Ltd. 2018
R. Maity, *Statistical Methods in Hydrology and Hydroclimatology*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-10-8779-0