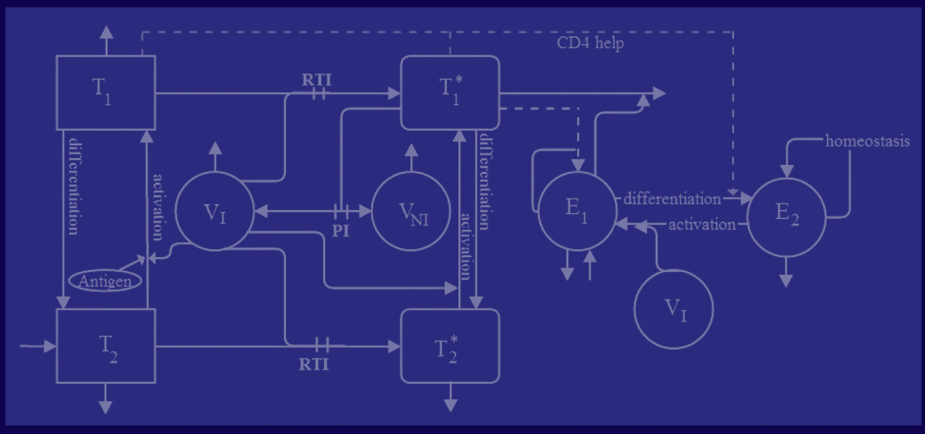
MONOGRAPHS AND RESEARCH NOTES IN MATHEMATICS

Modeling and Inverse Problems in the Presence of Uncertainty



H. T. Banks, Shuhua Hu, and W. Clayton Thompson

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Modeling and Inverse Problems in the Presence of Uncertainty

MONOGRAPHS AND RESEARCH NOTES IN MATHEMATICS

Series Editors

Published Titles

Iterative Optimization in Inverse Problems, Charles L. Byrne

Modeling and Inverse Problems in the Presence of Uncertainty, H. T. Banks, Shuhua Hu, and W. Clayton Thompson

Forthcoming Titles

Sinusoids: Theory and Technological Applications, Prem K. Kythe

Stochastic Cauchy Problems in Infinite Dimensions: Generalized and Regularized Solutions, Irina V. Melnikova and Alexei Filinkov

Signal Processing: A Mathematical Approach, Charles L. Byrne

Monomial Algebra, Second Edition, Rafael Villarreal

Groups, Designs, and Linear Algebra, Donald L. Kreher

Geometric Modeling and Mesh Generation from Scanned Images, Yongjie Zhang

Difference Equations: Theory, Applications and Advanced Topics, Third Edition, Ronald E. Mickens

Set Theoretical Aspects of Real Analysis, Alexander Kharazishvili

Method of Moments in Electromagnetics, Second Edition, Walton C. Gibson

The Separable Galois Theory of Commutative Rings, Second Edition, Andy R. Magid

Dictionary of Inequalities, Second Edition, Peter Bullen

Actions and Invariants of Algebraic Groups, Second Edition, Walter Ferrer Santos and Alvaro Rittatore

Practical Guide to Geometric Regulation for Distributed Parameter Systems, Eugenio Aulisa and David S. Gilliam

Analytical Methods for Kolmogorov Equations, Second Edition, Luca Lorenzi

Handbook of the Tutte Polynomial, Joanna Anthony Ellis-Monaghan and Iain Moffat

Blow-up Patterns for Higher-Order: Nonlinear Parabolic, Hyperbolic Dispersion and Schrödinger Equations, Victor A. Galaktionov, Enzo L. Mitidieri and Stanislav Pohozaev

Application of Fuzzy Logic to Social Choice Theory, John N. Mordeson, Davendar Malik and Terry D. Clark

Microlocal Analysis on Rˆn and on NonCompact Manifolds, Sandro Coriasco

Cremona Groups and Icosahedron, Ivan Cheltsov and Constantin Shramov

MONOGRAPHS AND RESEARCH NOTES IN MATHEMATICS

Modeling and Inverse Problems in the Presence of Uncertainty

H. T. Banks, Shuhua Hu,
and W. Clayton Thompson
North Carolina State University
Raleigh, USA

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20140224

International Standard Book Number-13: 978-1-4822-0643-2 (eBook - PDF)

Contents

This page intentionally left blank

Preface

Writing a research monograph on a "hot topic" such as "uncertainty propagation" is a somewhat daunting undertaking. Nonetheless, we decided to collect our own views, supported by our own research efforts over the past 12–15 years on a number of aspects of this topic, and summarize these for the possible enlightenment they might provide (for us, our students and others). The research results discussed below are thus necessarily filled with a preponderance of references to our own research reports and papers. In numerous references below (given at the conclusion of each chapter), we refer to CRSC-TRXX-YY. This refers to early Technical Report versions of manuscripts which can be found on the Center for Research in Scientific Computation website at North Carolina State University where XX refers to the year, e.g., XX = 03 is 2003, XX = 99 is 1999, while the YY refers to the number of the report in that year. These can be found at and downloaded from http://www.ncsu.edu/crsc/reports.html where they are listed by year.

Our presentation here has an intended audience from the community of investigators in applied mathematics interested in deterministic and/or stochastic models and their interactions as well as scientists in biology, medicine, engineering and physics interested in basic modeling and inverse problems, uncertainty in modeling, propagation of uncertainty and statistical modeling.

We owe great thanks to our former and current students, postdocs and colleagues for their patience in enduring lectures, questions, feedback and some proofreading. Special thanks are due (in no particular order) to Zack Kenz, Keri Rehm, Dustin Kapraun, Jared Catenacci, Katie Link, Kris Rinnovatore, Kevin Flores, John Nardini, Karissa Cross and Laura Poag for careful reading of notes and suggested corrections/revisions on subsets of the material for this monograph. However, in a sincere attempt to give credit where it is due, each of the authors firmly insists that any errors in judgment, mathematical content, grammar or typos in the material presented in this monograph are entirely the responsibility of his/her two co-authors!!

We (especially young members of our research group) have been generously supported by research grants and fellowships from US federal funding agencies including AFSOR, DARPA, NIH, NSF, DED, and DOE. For this support and encouragement we are all most grateful.

H.T. Banks
Shuhua Hu
W. Clayton Thompson

This page intentionally left blank

Chapter 1

Introduction

The terms ***uncertainty quantification*** and ***uncertainty propagation*** have become so widely used as to almost have little meaning unless they are further explained. Here we focus primarily on two basic types of problems:

1. Modeling and inverse problems where one assumes that a precise mathematical model without modeling error is available. This is a standard assumption underlying a large segment of what is taught in many modern statistics courses with a frequentist philosophy. More precisely, a mathematical model is given by a dynamical system

$$\frac{d\boldsymbol{x}}{dt}(t) = \boldsymbol{g}(t, \boldsymbol{x}(t), \boldsymbol{q}) \tag{1.1}$$

$$\boldsymbol{x}(t_0) = \boldsymbol{x}_0 \tag{1.2}$$

   with **observation process**

$$\boldsymbol{f}(t; \boldsymbol{\theta}) = \mathcal{C}\boldsymbol{x}(t; \boldsymbol{\theta}), \tag{1.3}$$

   where $\boldsymbol{\theta} = (\boldsymbol{q}, \boldsymbol{x}_0)$. The mathematical model is an n-dimensional deterministic system and there is a corresponding "truth" parameter $\boldsymbol{\theta}_0 = (\boldsymbol{q}_0, \boldsymbol{x}_{00})$ so that in the presence of no measurement error the data can be described exactly by the deterministic system at $\boldsymbol{\theta}_0$. Thus, uncertainty is present entirely due to some **statistical model** of the form

$$\boldsymbol{Y}_j = \boldsymbol{f}(t_j; \boldsymbol{\theta}_0) + \boldsymbol{\mathcal{E}}_j, \quad j = 1, \ldots, N, \tag{1.4}$$

   where $\boldsymbol{f}(t_j; \boldsymbol{\theta}) = \mathcal{C}\boldsymbol{x}(t_j; \boldsymbol{\theta})$, $j = 1, \ldots, N$, corresponds to the observed part of the solution of the mathematical model (1.1)–(1.2) at the jth covariate or observation time and $\boldsymbol{\mathcal{E}}_j$ is some type of (possibly state dependent) measurement error. For example, we consider errors that include those of the form $\boldsymbol{\mathcal{E}}_j = \boldsymbol{f}(t_j; \boldsymbol{\theta}_0)^{\gamma} \circ \widetilde{\boldsymbol{\mathcal{E}}}_j$ where the operation ${}^{\gamma}\circ$ denotes component-wise exponentiation by γ followed by component-wise multiplication and $\gamma \geq 0$.

2. An alternate problem wherein the mathematical modeling ***itself*** is a major source of uncertainty and this uncertainty usually propagates in time. That is, the mathematical model has major uncertainties in its form and/or its parametrization and/or its initial/boundary data, and this uncertainty is propagated dynamically via some framework as yet to be determined.

Before we begin the inverse problem discussions, we give a brief but useful review of certain basic probability and statistical concepts. After the probability and statistics review we present a chapter summarizing both mathematical and statistical aspects of inverse problem methodology which includes ordinary, weighted and generalized least-squares formulations. We discuss asymptotic theories, bootstrapping and issues related to evaluation of the correctness of the assumed form of statistical models. We follow this with a discussion of methods for evaluating and comparing the validity of appropriateness of a collection of models for describing a given data set, including statistically based model selection and model comparison techniques.

In Chapter 5 we present a summary of recent results on the estimation of probability distributions when they are embedded in complex mathematical models and only aggregate (not individual) data are available. This is followed by a brief chapter on optimal design (what to measure? when and where to measure?) of experiments to be carried out in support of inverse problems for given models.

The last two chapters focus on the uncertainty in model formulation itself (the second item listed above as the focus of this monograph). In Chapter 7 we consider the general problem of evolution of probability density functions in time. This is done in the context of associated processes resulting from stochastic differential equations (SDE), which are driven by white noise, and those resulting from random differential equations (RDE), which are driven by colored noise. We also discuss their respective wide applications in a number of different fields including physics and biology. We also consider the general relationship between SDE and RDE and establish that there are classes of problems for which there is an equivalence between the solutions of the two formulations. This equivalence, which we term pointwise equivalence, is in the sense that the respective probability density functions are the same at each time t. We show, however, that the stochastic processes resulting from the SDE and its corresponding pointwise equivalent RDE are generally not the same in that they may have different covariance functions.

In a final chapter we consider questions related to the appropriateness of discrete versus continuum models in transitions from small numbers of individuals (particles, populations, molecules, etc.) to large numbers. These investigations are carried out in the context of continuous time Markov chain (CTMC) models and the Kurtz limit theorems for approximations for large number stochastic populations by ordinary differential equations for corresponding mean populations. Algorithms for simulating CTMC models and CTMC models with delays (discrete and random) are explained and simulations are presented for problems arising in specific applications.

The monograph contains illustrative examples throughout, many of them directly related to research projects carried out by our group at North Carolina State University over the past decade.

Chapter 2

Probability and Statistics Overview

The theory of probability and statistics is an essential mathematical tool in the formulation of inverse problems, in the development of subsequent analysis and approaches to statistical hypothesis testing and model selection criteria, and in the study of uncertainty propagation in dynamic systems. Our coverage of these fundamental and important topics is brief and limited in scope. Indeed, we provide in this section a few definitions and basic concepts in the theory of probability and statistics that are essential for the understanding of estimators, confidence intervals, model selection criteria and stochastic processes. For more information on the topics in this chapter, selected references are provided at the end of these as well as subsequent chapters.

2.1 Probability and Probability Space

The set of all possible outcomes in a statistical experiment is called the *sample space* and is denoted by Ω. Each element $\omega \in \Omega$ is called a *sample point*. A collection of outcomes in which we are interested is called an *event*; that is, an event is a subset of Ω. For example, consider the experiment of rolling a six-sided die. In this case, there are six possible outcomes, and the sample space can be represented as

$$\Omega = \{1, 2, 3, 4, 5, 6\}. \tag{2.1}$$

An event $\mathbb{A}$ might be defined as

$$\mathbb{A} = \{1, 5\}, \tag{2.2}$$

which consists of the outcomes 1 and 5. Note that we say the event $\mathbb{A}$ occurs if the outcome of the experiment is in the set $\mathbb{A}$. Consider another experiment of tossing a coin three times. A sample point in this experiment indicates the result of each toss; for example, HHT indicates that two heads and then a tail were observed. The sample space for this experiment has eight sample points; that is,

$$\Omega = \{\text{HHH}, \text{HHT}, \text{HTH}, \text{THH}, \text{TTH}, \text{THT}, \text{HTT}, \text{TTT}\}. \tag{2.3}$$

An event $\mathbb{A}$ might be defined as the set of outcomes for which the first toss is a head; that is,

$$\mathbb{A} = \{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}\}.$$

Thus, we see that the sample point could be either a numerical value or a character value. Based on the number of sample points contained in the sample space, the sample space can be either finite (as we illustrate above) or infinitely countable (e.g., the number of customers to arrive in a bank) or uncountable (e.g., the sample space for the lifetime of a bulb or the sample space for the reaction time to a certain stimulus).

Definition 2.1.1 *Let Ω be the given sample space. Then the σ-algebra $\mathcal{F}$ is a collection of subsets of Ω with the following properties:*

(i) The empty set $\emptyset$ is an element of $\mathcal{F}$; that is, $\emptyset \in \mathcal{F}$.

(ii) (closed under complementation): If $\mathbb{A} \in \mathcal{F}$, then $\mathbb{A}^c \in \mathcal{F}$, where $\mathbb{A}^c$ denotes the complement of the event $\mathbb{A}$, which consists of all sample points in Ω that are not in $\mathbb{A}$.

(iii) (closed under countable unions): If $\mathbb{A}_1, \mathbb{A}_2, \mathbb{A}_3, \ldots, \in \mathcal{F}$, then $\bigcup_j \mathbb{A}_j \in \mathcal{F}$.

The pair $(\Omega, \mathcal{F})$ is called a measurable space. A subset $\mathbb{A}$ of Ω is said to be a measurable set (or event) if $\mathbb{A} \in \mathcal{F}$.

It is worth noting that a σ-algebra is also called a *σ-field* in the literature. In addition, one can define many different σ-algebras associated with the sample space Ω. The σ-algebra we will mainly consider is the smallest one that contains all of the open sets in the given sample space. In other words, it is the algebra generated by a topological space, whose definition is given as follows.

Definition 2.1.2 *A topology $\mathcal{T}$ on a set Ω is a collection of subsets of Ω having the following properties:*

- *$\emptyset \in \mathcal{T}$ and $\Omega \in \mathcal{T}$.*
- *(closed under finite intersection): If $\mathbb{U}_i \in \mathcal{T}$, $i = 1, 2, \ldots, l$, with l being a positive integer, then $\bigcap_{i=1}^{l} \mathbb{U}_i \in \mathcal{T}$.*
- *(closed under arbitrary union): If $\{\mathbb{U}_\alpha\}$ is an arbitrary collection of members of $\mathcal{T}$ (finite, countable, uncountable), then $\bigcup_\alpha \mathbb{U}_\alpha \in \mathcal{T}$.*

The pair $(\Omega, \mathcal{T})$ is called a topological space. A subset $\mathbb{U}$ of Ω is said to be an open set if $\mathbb{U} \in \mathcal{T}$. A subset $\mathbb{A}$ of Ω is said to be closed if $\mathbb{A}^c$ is an open set.

A σ-algebra generated by all the open sets in a given sample space Ω is often called a *Borel algebra* or *Borel σ-algebra* (denoted by $\mathcal{B}(\Omega)$), and the sets in a Borel algebra are called *Borel sets*.

Associated with an event $\mathbb{A} \in \mathcal{F}$ is its probability $\text{Prob}\{\mathbb{A}\}$, which indicates the likelihood that event $\mathbb{A}$ occurs. For example, in a fair experiment of rolling a die, where one assumes that each possible sample point has probability $\frac{1}{6}$, then the event $\mathbb{A}$ as defined by (2.2) has probability $\text{Prob}\{\mathbb{A}\} = \frac{2}{6} = \frac{1}{3}$. The strict definition for the probability is given as follows.

Definition 2.1.3 *A probability Prob on the measurable space $(\Omega, \mathcal{F})$ is a set function $Prob : \mathcal{F} \to [0, 1]$ such that*

(i) $Prob\{\emptyset\} = 0$, $Prob\{\Omega\} = 1$.

(ii) (completely additive): If $\mathbb{A}_1, \mathbb{A}_2, \mathbb{A}_3, \ldots,$ is a finite or an infinite sequence of disjoint subsets in $\mathcal{F}$, then $Prob\{\cup_j \mathbb{A}_j\} = \sum_j Prob\{\mathbb{A}_j\}$.

The triplet $(\Omega, \mathcal{F}, Prob)$ is called a probability space.

It is worth noting that a probability is also called a *probability measure* or a *probability function* (these names will be used interchangeably in this monograph), and a probability measure defined on a Borel algebra is called a *Borel probability measure*.

Using Definition 2.1.3 for probability, a number of immediate consequences can also be derived which have important applications. For example, the probability that an event will occur and that it will not occur always sum to 1. That is,

$$\text{Prob}\{\mathbb{A}\} + \text{Prob}\{\mathbb{A}^c\} = 1.$$

In addition, if $\mathbb{A}, \mathbb{B} \in \mathcal{F}$, then we have

$$\text{Prob}\{\mathbb{A}\} \leq \text{Prob}\{\mathbb{B}\} \text{ if } \mathbb{A} \subset \mathbb{B}.$$

It can also be found that if $\mathbb{A}_j \in \mathcal{F}$, $j = 1, 2, \ldots$, then

$$\text{Prob}\left\{\cup_{j=1}^{\infty} \mathbb{A}_j\right\} \leq \sum_{j=1}^{\infty} \text{Prob}\{\mathbb{A}_j\}.$$

If $\text{Prob}\{\mathbb{A}\} = 1$, then we say that the event $\mathbb{A}$ occurs "with probability 1" or "*almost surely* (a.s.)." A set $\mathbb{A} \in \mathcal{F}$ is called a *null set* if $\text{Prob}\{\mathbb{A}\} = 0$. In addition, a probability space $(\Omega, \mathcal{F}, \text{Prob})$ is said to be *complete* if for any two sets $\mathbb{A}$ and $\mathbb{B}$ the following condition holds: If $\mathbb{A} \subset \mathbb{B}$, $\mathbb{B} \in \mathcal{F}$ and $\text{Prob}\{\mathbb{B}\} = 0$, then $\mathbb{A} \in \mathcal{F}$. It is worth noting that any probability space can be extended into a complete probability space (e.g., see [15, p. 10] and the references therein). Hence, we will assume that all the probability spaces are complete in the remainder of this monograph.

Remark 2.1.1 *We remark that we can generalize a probability space to more general measure spaces. Specifically, a measure ν on the measurable space $(\Omega, \mathcal{F})$ is a set function $\nu : \mathcal{F} \to [0, \infty]$ such that $\nu(\emptyset) = 0$ and ν is completely additive (that is, the second property of probability in Definition 2.1.3 holds). The triplet $(\Omega, \mathcal{F}, \nu)$ is called a measure space. If $\nu(\Omega)$ is finite, then ν is said to be a finite measure. In particular, Prob is a normalized finite measure with $Prob(\Omega) = 1$. Hence, a probability possesses all the general properties of a finite measure.*

Remark 2.1.2 *Let $(\Omega, \mathcal{F}, \nu)$ be a measure space. If $\Omega = \bigcup_{j=1}^{\infty} \mathbb{A}_j$ and $\nu(\mathbb{A}_j)$ is finite for all j, then we say that ν is a σ-finite measure and $(\Omega, \mathcal{F}, \nu)$ is a σ-finite measure space. Another measure that we will consider in this monograph is the Lebesgue measure, which is defined on the measurable space $(\mathbb{R}^l, \mathcal{B}(\mathbb{R}^l))$ and is given by*

$$\nu((a_1, b_1) \times \cdots \times (a_l, b_l)) = \prod_{j=1}^{l} (b_j - a_j),$$

that is, the volume of the interval $(a_1, b_1) \times \cdots \times (a_l, b_l)$. We thus see that the Lebesgue measure of a countable set of points is zero (that is, a countable set of points is a null set with respect to the Lebesgue measure), and that the Lebesgue measure of a k-dimensional plane in $\mathbb{R}^l$ $(l > k)$ is also zero. We refer the interested reader to some real analysis textbooks such as [11, 23] for more information on a measure as well as its properties.

2.1.1 Joint Probability

Instead of considering a single experiment, let us perform two experiments and consider their outcomes. For example, the two experiments may be two separate tosses of a single die or a single toss of two dice. The sample space in this case consists of 36 pairs (k, j), where $k, j = 1, 2, \ldots, 6$. Note that in a fair dice game, each sample point in the sample space has probability $\frac{1}{36}$. We now consider the probability of joint events, such as $\{k = 2, j = \text{odd}\}$. We begin by denoting the event of one experiment by $\mathbb{A}_k$, $k = 1, 2, \ldots, l$, and the event of the second experiment by $\mathbb{B}_j$, $j = 1, 2, \ldots, m$. The combined experiment has the joint events $(\mathbb{A}_k, \mathbb{B}_j)$, where $k = 1, 2, \ldots, l$ and $j = 1, 2, \ldots, m$.

The joint probability $\text{Prob}\{\mathbb{A}_k, \mathbb{B}_j\}$, also denoted by $\text{Prob}\{\mathbb{A}_k \cap \mathbb{B}_j\}$ (which will be occasionally used in this monograph for notational convenience) or $\text{Prob}\{\mathbb{A}_k \mathbb{B}_j\}$ in the literature, indicates the likelihood that the events $\mathbb{A}_k$ and $\mathbb{B}_j$ occur simultaneously. By Definition 2.1.3, a number of immediate consequences can also be derived for the joint probability. For example, $\text{Prob}\{\mathbb{A}_k, \mathbb{B}_j\}$ satisfies the condition $0 \leq \text{Prob}\{\mathbb{A}_k, \mathbb{B}_j\} \leq 1$. In addition,

if $\mathbb{B}_j$ for $j = 1, 2, \ldots, m$ are mutually exclusive (i.e., $\mathbb{B}_i \cap \mathbb{B}_j = \emptyset, i \neq j$) such that $\bigcup_{j=1}^{m} \mathbb{B}_j = \Omega$, then

$$\sum_{j=1}^{m} \text{Prob}\{\mathbb{A}_k, \mathbb{B}_j\} = \text{Prob}\{\mathbb{A}_k\}. \tag{2.4}$$

Furthermore, if all the outcomes of the two experiments are mutually exclusive such that $\bigcup_{k=1}^{l} \mathbb{A}_k = \Omega$ and $\bigcup_{j=1}^{m} \mathbb{B}_j = \Omega$, then $\sum_{k=1}^{l} \sum_{j=1}^{m} \text{Prob}\{\mathbb{A}_k, \mathbb{B}_j\} = 1$. The generalization of the above concept to more than two experiments follows in a straightforward manner.

2.1.2 Conditional Probability

Next, we consider a joint event with probability $\text{Prob}\{\mathbb{A}, \mathbb{B}\}$. Assuming that event $\mathbb{A}$ has occurred and $\text{Prob}\{\mathbb{A}\} > 0$, we wish to determine the probability of the event $\mathbb{B}$. This is called the *conditional probability* of event $\mathbb{B}$ given the occurrence of event $\mathbb{A}$ and is given by

$$\text{Prob}\{\mathbb{B}|\mathbb{A}\} = \frac{\text{Prob}\{\mathbb{A}, \mathbb{B}\}}{\text{Prob}\{\mathbb{A}\}}. \tag{2.5}$$

Definition 2.1.4 *Two events, $\mathbb{A}$ and $\mathbb{B}$, are said to be statistically independent if and only if*

$$Prob\{\mathbb{A}, \mathbb{B}\} = Prob\{\mathbb{A}\} Prob\{\mathbb{B}\}. \tag{2.6}$$

Statistical independence is often simply called *independence*. By (2.5) and (2.6), we see that if $\mathbb{A}$ and $\mathbb{B}$ are independent, then

$$\text{Prob}\{\mathbb{B}|\mathbb{A}\} = \text{Prob}\{\mathbb{B}\}. \tag{2.7}$$

In addition, we observe that if (2.7) holds, then by (2.5) and Definition 2.1.4 we know that $\mathbb{A}$ and $\mathbb{B}$ are independent. Thus, (2.7) can also be used as a definition for the independence of two events.

Two very useful relationships for conditional probabilities can be given. If $\mathbb{A}_k$, $k = 1, 2, \ldots, l$, are mutually exclusive events such that $\bigcup_{k=1}^{l} \mathbb{A}_k = \Omega$ and $\mathbb{B}$ is an arbitrary event with $\text{Prob}\{\mathbb{B}\} > 0$, then by (2.4) and (2.5) we have

$$\text{Prob}\{\mathbb{B}\} = \sum_{j=1}^{l} \text{Prob}\{\mathbb{A}_j, \mathbb{B}\} = \sum_{j=1}^{l} \text{Prob}\{\mathbb{B}|\mathbb{A}_j\} \text{Prob}\{\mathbb{A}_j\}, \tag{2.8}$$

and

$$\text{Prob}\{\mathbb{A}_k|\mathbb{B}\} = \frac{\text{Prob}\{\mathbb{A}_k, \mathbb{B}\}}{\text{Prob}\{\mathbb{B}\}} = \frac{\text{Prob}\{\mathbb{B}|\mathbb{A}_k\}\text{Prob}\{\mathbb{A}_k\}}{\sum_{j=1}^{l} \text{Prob}\{\mathbb{B}|\mathbb{A}_j\}\text{Prob}\{\mathbb{A}_j\}}. \tag{2.9}$$

Equation (2.8) is often called the *law of total probability*. Equation (2.9) is known as Bayes' formula or Bayes' Theorem. Here $\text{Prob}\{\mathbb{A}_k\}$ is called a *prior* probability of event $\mathbb{A}_k$, $\text{Prob}\{\mathbb{B}|\mathbb{A}_k\}$ is called the likelihood of $\mathbb{B}$ given $\mathbb{A}_k$, and $\text{Prob}\{\mathbb{A}_k|\mathbb{B}\}$ is called a *posterior* probability of event $\mathbb{A}_k$ obtained by using the information gained from $\mathbb{B}$.

2.2 Random Variables and Their Associated Distribution Functions

In most applications of probability theory, we are not interested in the details associated with each sample point but rather in some numerical description of the outcome of an experiment. For example, in the experiment of tossing a coin three times, we might only be interested in the number of heads obtained in these three tosses. In the language of probability and statistics, the number of heads obtained in these three tosses is called a *random variable*. The values of this particular random variable corresponding to each sample point in (2.3) are given by

HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
3	2	2	2	1	1	1	0.

This implies that the range (i.e., the collection of all possible values) of this random variable is $\{0, 1, 2, 3\}$. The strict definition of a random variable is as follows.

Definition 2.2.1 *A function $X : \Omega \to \mathbb{R}$ is said to be a random variable defined on a measurable space $(\Omega, \mathcal{F})$ if for any $x \in \mathbb{R}$ we have $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$. Such a function is said to be measurable with respect to $\mathcal{F}$. In addition, for any fixed $\omega \in \Omega$, $X(\omega)$ is called a realization of this random variable.*

As is usually done and also for notational convenience, we suppress the dependence of the random variable on ω if no confusion occurs; that is, we denote $X(\cdot)$ by X. Under this convention, the set $\{\omega \in \Omega : X(\omega) \leq x\}$ is simply written as $\{X \leq x\}$. In addition, a realization of a random variable is simply denoted by its corresponding lower case letter; for example, a realization of random variable X is denoted by x. We point out that when we consider

a sequence of random variables in this monograph, we assume that they are defined on the same probability space.

The σ-algebra generated by the random variable X, often denoted by $\sigma(X)$, is given by

$$\sigma(X) = \{\mathbb{A} \subset \Omega \mid \mathbb{A} = X^{-1}(\mathbb{B}),\ \mathbb{B} \in \mathcal{B}(\mathbb{R})\},$$

where $\mathcal{B}(\mathbb{R})$ is the Borel algebra on $\mathbb{R}$. It is worth noting that this is the smallest σ-algebra with respect to which X is measurable. This means that if X is $\mathcal{F}$-measurable, then $\sigma(X) \subset \mathcal{F}$.

Remark 2.2.1 *The concept of a random variable can be generalized so that its range can be some complicated space rather than* $\mathbb{R}$. *Let* $(\Omega, \mathcal{F}, Prob)$ *be a probability space, and* $(\mathbb{S}, \mathscr{S})$ *be a measurable space. A function* $X : \Omega \to \mathbb{S}$ *is said to be a random element if for any* $\mathbb{B} \in \mathscr{S}$ *we have* $\{\omega \in \Omega : X(\omega) \in \mathbb{B}\} \in \mathcal{F}$. *Specifically, if* $\mathbb{S} = \mathbb{R}^m$, *then we say that random element* X *is an* m*-dimensional random vector.*

2.2.1 Cumulative Distribution Function

For any random variable, there is an associated function called a cumulative distribution function, which is defined as follows.

Definition 2.2.2 *The* cumulative distribution function (CDF) *of random variable* X *is the function* $P : \mathbb{R} \to [0, 1]$ *defined by*

$$P(x) = Prob\{X \leq x\}, \quad x \in \mathbb{R}. \tag{2.10}$$

The cumulative distribution function is sometimes simply called the *distribution function*. It has the following properties (inherited from the probability measure):

(i) P is a right continuous function of x; that is, $\lim_{\Delta x \to 0+} P(x + \Delta x) = P(x)$.

(ii) P is a non-decreasing function of x; that is, if $x_1 \leq x_2$, then $P(x_1) \leq P(x_2)$.

(iii) $P(-\infty) = 0$, $P(\infty) = 1$.

The last two properties imply that the cumulative distribution function P has bounded variation, where the variation of a function is defined as follows.

Definition 2.2.3 *The* m*-variation of a real-valued function* h *on the interval* $[\underline{x}, \bar{x}] \subset \mathbb{R}$ *is defined as*

$$[h]^{(m)}([\underline{x}, \bar{x}]) = \sup \sum_{j=0}^{l-1} |h(x_{j+1}^l) - h(x_j^l)|^m, \tag{2.11}$$

where the supremum is taken over all partitions $\{x_j^l\}_{j=0}^l$ *of* $[\underline{x}, \bar{x}]$. *For the case* $m = 1$ *it is simply called variation (it is also called total variation in the literature), and for the case* $m = 2$ *it is called quadratic variation. If* $[h]^{(m)}([\underline{x}, \bar{x}])$ *is finite, then we say that* h *has bounded (or finite)* m*-variation on the given interval* $[\underline{x}, \bar{x}]$.

If h *is a function of* $x \in \mathbb{R}$, *then* h *is said to have finite* m*-variation if* $[h]^{(m)}([\underline{x}, \bar{x}])$ *is finite for any given* $\underline{x}$ *and* $\bar{x}$. *In addition,* h *is said to have bounded* m*-variation if there exists a constant* c_h *such that* $[h]^{(m)}([\underline{x}, \bar{x}]) < c_h$ *for any* $\underline{x}$ *and* $\bar{x}$, *where* c_h *is independent of* $\underline{x}$ *and* $\bar{x}$.

The properties of the cumulative distribution function also imply that P has derivatives almost everywhere with respect to the Lebesgue measure (that is, the set of points at which P is not differentiable is a null set with respect to the Lebesgue measure), but it should be noted that P does not have to be equal to the integral of its derivative. These properties also imply that the cumulative distribution function can only have jump discontinuities and it has at most countably many jumps. Thus, we see that the cumulative distribution function could be a step function (illustrated in the upper left panel of Figure 2.1), a continuous function (illustrated in the upper right panel of Figure 2.1) or a function with a mixture of continuous pieces and jumps (illustrated in the bottom plots of Figure 2.1). The last type of cumulative distribution function could result from a convex combination of the first two types of cumulative distribution functions. In fact, this is how we obtained the cumulative distribution functions demonstrated in the bottom plots of Figure 2.1. Specifically, let P_j be the cumulative distribution function of some random variable, $j = 1, 2, \ldots, m$, and $\varpi_j, j = 1, 2, \ldots, m$, be some non-negative numbers such that $\sum_{j=1}^{m} \varpi_j = 1$. Then we easily see that $\sum_{j=1}^{m} \varpi_j P_j$ is also a cumulative distribution function. This type of distribution is often called a *mixture distribution*.

Remark 2.2.2 *It is worth noting that there are many different ways found in the literature to define the continuity of a random variable. One is based on the range of the random variable. In this case, a discrete random variable is one with its range consisting of a countable subset in* $\mathbb{R}$ *with either a finite or infinite number of elements. For example, the random variable defined in the experiment of tossing a coin three times is a discrete random variable. A continuous random variable is one that takes values in a continuous interval. For example, the random variable defined in the experiment for recording the reaction time to a certain stimulus is a continuous random variable.*

Another way to define the continuity of random variables is based on the continuity of the cumulative distribution function for a random variable. Specifically, if the cumulative distribution function is continuous (as illustrated in the upper right panel of Figure 2.1*), then the associated random variable is said to be continuous. If the cumulative distribution function is a*

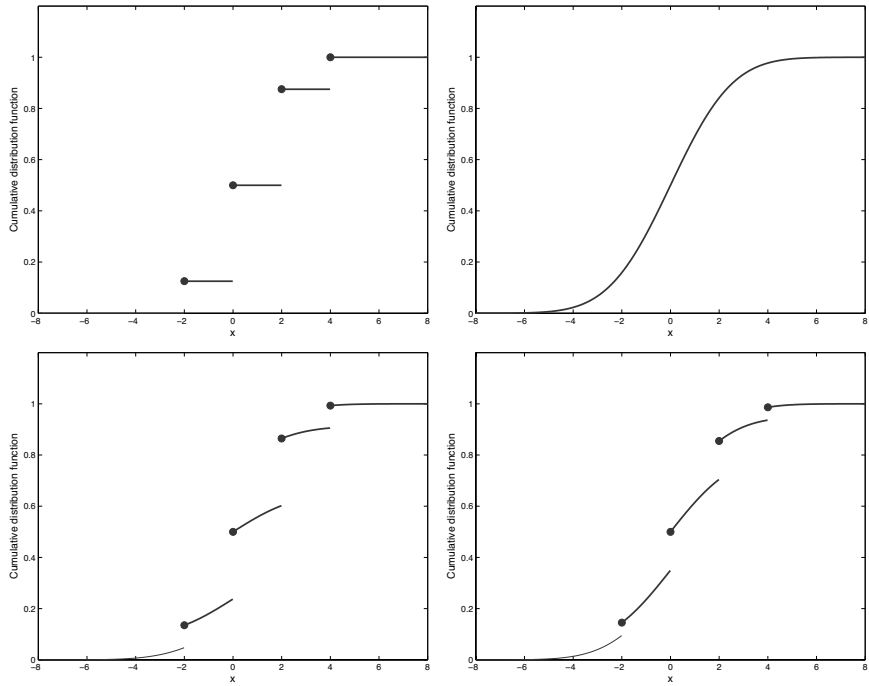


FIGURE 2.1: Cumulative distribution function: (upper left) a step function P_1, (upper right) a continuous function P_2, (lower left) $0.7P_1 + 0.3P_2$, (lower right) $0.4P_1 + 0.6P_2$.

step function (as illustrated in the upper left panel of Figure 2.1*), then the associated random variable is said to be discrete. We see that the discrete random variable defined here is equivalent to that in the first case (as the cumulative distribution function has at most countably many jumps). However, the continuous random variable defined here is more restrictive than that in the first case where the cumulative distribution function of a continuous random variable may be either a continuous function or a function with a mixture of continuous pieces and jumps (as illustrated in the bottom plots of Figure* 2.1*).*

The last way to define the continuity of random variables is based on whether or not a random variable has an associated probability density function (discussed blow); specifically, a random variable is said to be continuous if there is a probability density function associated with it. As we shall see below, this definition is even stronger than the second one. In this monograph, we define a random variable to be discrete or continuous based on its range. However, the continuous random variables we will mainly consider in this monograph are those with associated probability density functions.

2.2.2 Probability Mass Function

Any discrete random variable has an associated *probability mass function*. Without loss of generality, we assume the range of discrete random variable X is $\{x_j\}$. Then the probability mass function of X is defined by

$$\Phi(x_j) = \text{Prob}\{X = x_j\}, \quad j = 1, 2, 3, \ldots. \tag{2.12}$$

Hence, we see that the value of the probability mass function at x_j is the probability associated with x_j, $j = 1, 2, 3, \ldots$. In addition, by the definition of probability we know that

$$\sum_j \Phi(x_j) = 1.$$

For example, the probability mass function of the discrete random variable defined in the experiment of tossing a coin three times is

$$\Phi(0) = \frac{1}{8}, \quad \Phi(1) = \frac{3}{8}, \quad \Phi(2) = \frac{3}{8}, \quad \Phi(3) = \frac{1}{8}. \tag{2.13}$$

The relationship between the probability mass function and the cumulative distribution function is given by

$$P(x) = \sum_{\{j:\, x_j \le x\}} \Phi(x_j), \quad \Phi(x_j) = P(x_j) - \lim_{x \to x_j -} P(x).$$

In such a case, the cumulative distribution function is a step function, and it is said to be discrete. Figure 2.2 illustrates the probability mass function (2.13) of the discrete random variable defined in the experiment of tossing a coin three times as well as the corresponding cumulative distribution function.

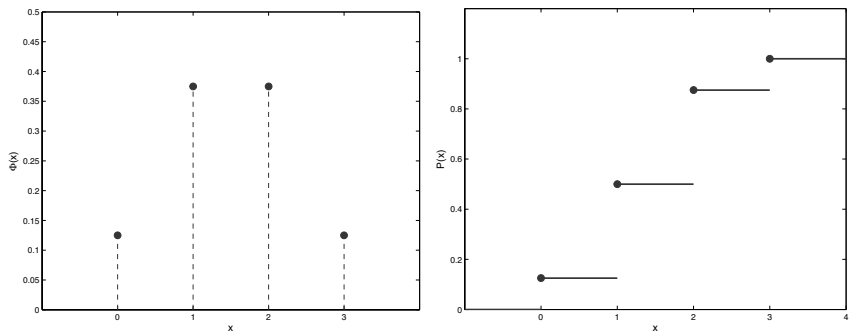


FIGURE 2.2: (left) The probability mass function of the discrete random variable defined in the experiment of tossing a coin three times; (right) the corresponding cumulative distribution function.

2.2.3 Probability Density Function

If the derivative p of the cumulative distribution function P exists for almost all x, and for all x

$$P(x) = \int_{-\infty}^{x} p(\xi) d\xi,$$

then p is called the *probability density function*. It should be noted that a necessary and sufficient condition for a cumulative distribution function to have an associated probability density function is that P is *absolutely continuous* in the sense that for any positive number ϵ, there exists a positive number c_l such that for any finite collection of disjoint intervals $(x_j, y_j) \subset \mathbb{R}$ satisfying $\sum_j |y_j - x_j| < c_l$ then $\sum_j |P(y_j) - P(x_j)| < \epsilon$. Thus, we see that the requirement for absolute continuity is much stronger than that for continuity. This implies that there are cases in which the cumulative distribution function is continuous but not absolutely continuous. An example is the *Cantor distribution* (also called the Cantor–Lebesgue distribution; e.g., see [10, p. 169] or [17, p. 38]), where the cumulative distribution function P is a constant between the points of the Cantor set (which is an uncountable set in $\mathbb{R}$ that has Lebesgue measure zero). Such a distribution is often called a *singular continuous distribution* (the derivative of its corresponding cumulative distribution function is zero almost everywhere), and the associated random variable is often termed a *singular continuous random variable*. This type of distribution is rarely encountered in practice. Based on Lebesgue's decomposition theorem, any cumulative distribution function can be written as a convex combination of a discrete, an absolutely continuous and a singular continuous cumulative distribution function. We refer the interested reader to [2, Section 31] for more information on this topic.

The name "density function" comes from the fact that the probability of the event $x_1 \leq X \leq x_2$ is given by

$$\begin{aligned} \text{Prob}\{x_1 \leq X \leq x_2\} &= \text{Prob}\{X \leq x_2\} - \text{Prob}\{X \leq x_1\} \\ &= P(x_2) - P(x_1) \\ &= \int_{x_1}^{x_2} p(x)\, dx. \end{aligned} \tag{2.14}$$

In addition, the probability density function p satisfies the following properties:

$$p(x) \geq 0, \qquad \int_{-\infty}^{\infty} p(x)\, dx = P(\infty) - P(-\infty) = 1.$$

Figure 2.3 illustrates the probability density function p (shown in the left panel) of a continuous random variable X and the corresponding cumulative distribution function P (shown in the right panel). By (2.14) we know that the shaded area under the probability density function between $x = 0$ and

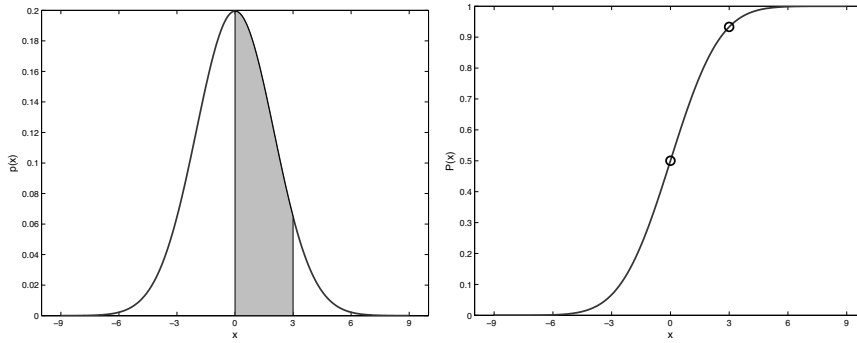


FIGURE 2.3: (left) Probability density function of a continuous random variable; (right) the corresponding cumulative distribution function.

$x = 3$ is equal to the probability that the value of X is between 0 and 3, and is also equal to $P(3) - P(0)$ (where the points $(0, P(0))$ and $(3, P(3))$ are indicated by the circles in the right panel of Figure 2.3).

For a discrete random variable, the corresponding probability density function does not exist in the ordinary sense. But it can be constructed in the general sense (a generalized function sense, sometimes called a "distributional sense") with the help of Dirac delta "functions" [26]. Specifically, for a discrete random variable X with range $\{x_j\}$, its probability density function can be written as

$$p(x) = \sum_j \Phi(x_j)\delta(x - x_j), \tag{2.15}$$

where δ is the Dirac delta "function," that is, $\delta(x - x_j) = \begin{cases} 0 & \text{if } x \neq x_j \\ \infty & \text{if } x = x_j \end{cases}$ with the property that $\displaystyle\int_{-\infty}^{\infty} h(x)\delta(x - x_j)dx = h(x_j)$. This construction has many advantages; for example, as we shall see later, it can put the definition of moments in the same framework for a discrete random variable and for a continuous random variable with a probability density function.

2.2.4 Equivalence of Two Random Variables

One can define the equivalence between two random variables in several senses.

Definition 2.2.4 *Two random variables are said to be equal in distribution (or identically distributed) if their associated cumulative distribution functions are equal.*

Definition 2.2.5 *Two random variables X and Y are said to be equal almost surely if Prob$\{X = Y\} = 1$.*

It is worth noting that two random variables that are equal almost surely are equal in distribution. However, the fact that two random variables are equal in distribution does not necessarily imply that they are equal almost surely. For example, in the experiment of tossing a fair coin three times, if we define X as the number of heads obtained in three tosses and Y as the number of tails observed in three tosses, then it can easily be seen that X and Y have the same probability mass function and hence are equal in distribution. However, $X(\omega) \neq Y(\omega)$ for any $\omega \in \Omega$, where Ω is defined in (2.3).

2.2.5 Joint Distribution Function and Marginal Distribution Function

The definition of cumulative distribution and probability density functions can be extended from one random variable to two or more random variables. In this case, the cumulative distribution function is often called the *joint cumulative distribution function* or simply the *joint distribution function*, and the probability density function is often called the *joint probability density function*. For example, the joint distribution function of two random variables X and Y is defined as

$$P(x, y) = \text{Prob}\{X \leq x, Y \leq y\}. \tag{2.16}$$

Similar to the one-dimensional case, the joint distribution function P is non-negative, non-decreasing and right continuous with respect to each variable. In addition,

$$\begin{array}{lll} P(-\infty, -\infty) = 0, & P(x, -\infty) = 0, & P(-\infty, y) = 0, \\ P(\infty, \infty) = 1, & P(x, \infty) = P_X(x), & P(\infty, y) = P_Y(y). \end{array} \tag{2.17}$$

Here P_X and P_Y are, respectively, the cumulative distribution functions of X and Y, and the subscript in P_X (P_Y) is used to emphasize that it is the cumulative distribution function of X (Y). (It should be noted that the subscript in the cumulative distribution function is always suppressed if no confusion occurs. Otherwise, we index a distribution function by the random variable it refers to.) In the context of two or more random variables, each random variable is often called a *marginal variable*, and the cumulative distribution function of each random variable is often called a *marginal distribution function*.

The corresponding joint probability density function of X and Y, if it exists, is defined as

$$p(x, y) = \frac{\partial^2}{\partial x \partial y} P(x, y). \tag{2.18}$$

Similar to the one-dimensional case, the joint probability density function of X and Y is non-negative and

$$\begin{aligned} &\text{Prob}\{x_1 \le X \le x_2, y_1 \le Y \le y_2\} = \int_{y_1}^{y_2} \int_{x_1}^{x_2} p(x,y)dxdy, \\ &\int_{\mathbb{R}^2} p(x,y)dxdy = 1. \end{aligned} \tag{2.19}$$

By (2.17) and (2.19), we see that the probability density function of X can be derived from the joint probability density function of X and Y, and is given as

$$p_X(x) = \int_{-\infty}^{\infty} p(x,y)dy, \tag{2.20}$$

where the subscript X in p_X is used to emphasize that this is the probability density function of X. (It should be noted that the subscript in the probability density function is always suppressed if no confusion occurs. Otherwise, we index a probability density function by the random variable it refers to.) Again, in the context of two or more random variables, the probability density function of each random variable is called the *marginal probability density function* (or simply the *marginal density function*). Similarly, the probability density function of Y can be derived from the joint probability density function of X and Y,

$$p_Y(y) = \int_{-\infty}^{\infty} p(x,y)dx. \tag{2.21}$$

Again the subscript Y in p_Y is used to emphasize that it is the probability density function of Y. Thus, by (2.20) and (2.21) we see that the probability density function of each random variable can be derived from the joint probability density function of X and Y by integrating across the other variable. However, the converse is usually not true; that is, the joint probability density function usually cannot be derived based on the associated marginal probability density functions.

In general, the joint cumulative distribution function of m random variables $X_1, X_2, \ldots, X_m$ is defined as

$$P(x_1, x_2, \ldots, x_m) = \text{Prob}\{X_1 \le x_1, X_2 \le x_2, \ldots, X_m \le x_m\}, \tag{2.22}$$

and the corresponding joint probability density function of $X_1, X_2, \ldots, X_m$, if it exists, is defined by

$$p(x_1, x_2, \ldots, x_m) = \frac{\partial^m}{\partial x_1 \partial x_2 \ldots \partial x_m} P(x_1, x_2, \ldots, x_m). \tag{2.23}$$

We can view these m random variables $X_1, X_2, \ldots, X_m$ as the components of an m-dimensional random vector (i.e., $\mathbf{X} = (X_1, X_2, \ldots, X_m)^T$). Hence, (2.22) and (2.23) can be respectively viewed as the cumulative distribution

function and probability density function of the random vector $\mathbf{X}$. In this case, we write these functions more compactly as $P(\mathbf{x})$ and $p(\mathbf{x})$ with $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$.

Similar to the two-dimensional case, for any $j \in \{1, 2, \ldots, m\}$, the probability density function of X_j can be derived from the joint probability density function of $X_1, X_2, \ldots, X_m$ by integrating with respect to all the other $m-1$ variables. For example, the probability density function of X_2 is given by

$$p_{X_2}(x_2) = \int_{\mathbb{R}^{m-1}} p(x_1, x_2, \ldots, x_m) dx_1 dx_3 dx_4 \cdots dx_m. \tag{2.24}$$

Similarly, the joint probability density function of any two or more random variables in the set of $\{X_j\}_{j=1}^m$ can be derived from the joint probability density function of $X_1, X_2, \ldots, X_m$ by integrating with respect to all the rest of the variables. For example, the joint probability density function of X_1 and X_2 is

$$p_{X_1X_2}(x_1, x_2) = \int_{\mathbb{R}^{m-2}} p(x_1, x_2, \ldots, x_m) dx_3 dx_4 \cdots dx_m.$$

Here the subscript X_1X_2 in $p_{X_1X_2}$ is used to emphasize that it is the joint probability density function of X_1 and X_2. Again, the (joint) probability density function of any subset of random variables in the set of $\{X_j\}_{j=1}^m$ is often called the *marginal probability density function*.

2.2.6 Conditional Distribution Function

The conditional distribution function of X given $Y = y$ is defined by

$$\mathcal{P}_{X|Y}(x|y) = \text{Prob}\{X \leq x | Y = y\}. \tag{2.25}$$

Then in analogy to (2.5), we have that

$$\mathcal{P}_{X|Y}(x|y) = \frac{\int_{-\infty}^{x} p_{XY}(\xi, y) d\xi}{p_Y(y)}. \tag{2.26}$$

Here p_{XY} is the joint probability density function of X and Y, and p_Y is the probability density function of Y. The corresponding *conditional probability density function* (or simply *conditional density function*) of X given $Y = y$, if it exists, is

$$\rho_{X|Y}(x|y) = \frac{d}{dx}\mathcal{P}_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)}. \tag{2.27}$$

By (2.27) we see that

$$p_{XY}(x, y) = \rho_{X|Y}(x|y) p_Y(y). \tag{2.28}$$

Definition 2.2.6 *Two random variables X and Y are said to be independent if the conditional distribution of X given Y is equal to the unconditional distribution of X.*

Definition 2.2.6 implies that if X and Y are independent, then we have $\rho_{X|Y}(x|y) = p_X(x)$. Hence, by (2.28), we obtain

$$p_{XY}(x, y) = p_X(x)p_Y(y), \tag{2.29}$$

which is also used as a definition for the independence of two random variables. Thus, we see that if two random variables are independent, then the joint probability density function can be determined by the marginal probability density functions. It should be noted that the definition for the independence of two general (either discrete or continuous) random variables is based on the independence of two events, and is given as follows.

Definition 2.2.7 *Two random variables X and Y are said to be independent if the σ-algebras they generate, $\sigma(X)$ and $\sigma(Y)$, are independent (that is, for any $\mathbb{A} \in \sigma(X)$ and for any $\mathbb{B} \in \sigma(Y)$, events $\mathbb{A}$ and $\mathbb{B}$ are independent).*

The above definition implies that if X and Y are independent, then their joint distribution function is given by the product of their marginal distribution functions; that is,

$$P_{XY}(x, y) = P_X(x)P_Y(y), \tag{2.30}$$

which is also used as a general definition for the independence of two random variables.

In general, for any positive integer $k \geq 2$, the *conditional distribution function* of X_k given $X_1 = x_1, \ldots, X_{k-1} = x_{k-1}$ is

$$\begin{aligned} &\mathcal{P}_{X_k|X_{k-1},\cdots,X_1}(x_k|x_{k-1}, \cdots, x_1) \\ &\quad = \text{Prob}\{X_k \leq x_k | X_{k-1} = x_{k-1}, \cdots, X_1 = x_1\} \\ &\quad = \frac{\int_{-\infty}^{x_k} p_{X_1,\ldots,X_k}(x_1, \ldots, x_{k-1}, \xi)d\xi}{p_{X_1,\ldots,X_{k-1}}(x_1, \ldots, x_{k-1})}. \end{aligned} \tag{2.31}$$

Here $p_{X_1,\ldots,X_{k-1}}$ is the joint probability density function of $X_1, \ldots, X_{k-1}$, and $p_{X_1,\ldots,X_k}$ is the joint probability density function of $X_1, \ldots, X_k$. The corresponding *conditional density function* of X_k given $X_1 = x_1, \ldots, X_{k-1} = x_{k-1}$ is

$$\begin{aligned} &\rho_{X_k|X_{k-1},\cdots,X_1}(x_k|x_{k-1}, \cdots, x_1) \\ &\quad = \frac{d}{dx_k}\mathcal{P}_{X_k|X_{k-1},\cdots,X_1}(x_k|x_{k-1}, \cdots, x_1) \\ &\quad = \frac{p_{X_1,\ldots,X_k}(x_1, \ldots, x_k)}{p_{X_1,\ldots,X_{k-1}}(x_1, \ldots, x_{k-1})}, \end{aligned} \tag{2.32}$$

which implies that

$$\begin{aligned} &p_{X_1,\ldots,X_k}(x_1, \ldots, x_k) \\ &\quad = \rho_{X_k|X_{k-1},\cdots,X_1}(x_k|x_{k-1}, \cdots, x_1)p_{X_1,\ldots,X_{k-1}}(x_1, \ldots, x_{k-1}). \end{aligned} \tag{2.33}$$

Hence, by (2.27) and the above equation we see that the joint probability density function $p_{X_1,\ldots,X_m}$ of any m random variables $X_1, X_2, \ldots, X_m$ can be written as

$$\begin{aligned} & p_{X_1,\ldots,X_m}(x_1, x_2, \ldots, x_m) \\ & \quad = p_{X_1}(x_1)\rho_{X_2|X_1}(x_2|x_1)\rho_{X_3|X_2,X_1}(x_3|x_2, x_1) \\ & \quad \cdots \rho_{X_m|X_{m-1},\cdots,X_1}(x_m|x_{m-1}, \cdots, x_1). \end{aligned} \tag{2.34}$$

The concept of the independence of two random variables can be extended to the case of a sequence of random variables, and it is given as follows.

Definition 2.2.8 *Random variables $X_1, X_2, \ldots, X_m$ are said to be mutually independent if*

$$p_{X_1,\ldots,X_m}(x_1, x_2, \ldots, x_m) = \prod_{j=1}^{m} p_{X_j}(x_j), \tag{2.35}$$

where $p_{X_1,\ldots,X_m}$ is the joint probability density function of m random variables $X_1, X_2, \ldots, X_m$, and p_{X_j} is the probability density function of X_j, $j = 1, 2, \ldots, m$.

In general, the conditional probability density function of a subset of the coordinates of $(X_1, X_2, \ldots, X_m)$ given the values of the remaining coordinates is obtained by dividing the joint probability density function of $(X_1, X_2, \ldots, X_m)$ by the marginal probability density function of the remaining coordinates. For example, the conditional probability density function of $(X_{k+1}, \ldots, X_m)$ given $X_1 = x_1, \ldots, X_k = x_k$ (where k is a positive integer such that $1 < k < m$) is defined as

$$\rho_{X_{k+1},\ldots,X_m|X_1,\ldots,X_k}(x_{k+1}, \ldots, x_m \mid x_1, \ldots, x_k) = \frac{p_{X_1,\ldots,X_m}(x_1, \ldots, x_m)}{p_{X_1,\ldots,X_k}(x_1, \ldots, x_k)}, \tag{2.36}$$

where $p_{X_1,\ldots,X_k}$ is the joint probability density function of $X_1, \ldots, X_k$, and $p_{X_1,\ldots,X_m}$ is the joint probability density function of $X_1, \ldots, X_m$. The concept of mutually independent random variables can also be extended to that of mutually independent random vectors. The definition is given as follows.

Definition 2.2.9 *Random vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ are said to be mutually independent if*

$$p_{\mathbf{X}_1,\ldots,\mathbf{X}_m}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) = \prod_{j=1}^{m} p_{\mathbf{X}_j}(\mathbf{x}_j), \tag{2.37}$$

where $p_{\mathbf{X}_1,\ldots,\mathbf{X}_m}$ is the joint probability density function of m random vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$, and $p_{\mathbf{X}_j}$ is the probability density function of $\mathbf{X}_j$, $j = 1, 2, \ldots, m$.

2.2.7 Function of a Random Variable

Let X be a random variable with cumulative distribution function P_X and probability density function p_X. Then by Definition 2.2.1 we know that for any measurable function $\eta : \mathbb{R} \to \mathbb{R}$, the composite function $\eta(X)$ is measurable and indeed is a random variable. Let $Y = \eta(X)$. Then the cumulative distribution function of Y is

$$P_Y(y) = \text{Prob}\{\eta(X) \le y\}. \tag{2.38}$$

If we assume that η is a monotone function, then by (2.38) we have

$$P_Y(y) = \begin{cases} \text{Prob}\{X \le \eta^{-1}(y)\} = P_X(\eta^{-1}(y)), & \eta \text{ is increasing} \\ \text{Prob}\{X \ge \eta^{-1}(y)\} = 1 - P_X(\eta^{-1}(y)), & \eta \text{ is decreasing.} \end{cases} \tag{2.39}$$

If we further assume that η^{-1} is differentiable, then differentiating both sides of (2.39) yields the probability density function of Y

$$p_Y(y) = p_X(\eta^{-1}(y)) \left| \frac{d\eta^{-1}(y)}{dy} \right|. \tag{2.40}$$

In general, we consider an m-dimensional random vector $\mathbf{X}$ with probability density function $p_{\mathbf{X}}$. Let $\mathbf{Y} = \boldsymbol{\eta}(\mathbf{X})$ with $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_m)^T$ and $\eta_j : \mathbb{R}^m \to \mathbb{R}$ be a measurable function for all j. Assume that $\boldsymbol{\eta}$ has a unique inverse $\boldsymbol{\eta}^{-1}$. Then the probability density function of the m-dimensional random vector $\mathbf{Y}$ is given by

$$p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{X}}(\boldsymbol{\eta}^{-1}(\mathbf{y}))|\mathcal{J}|, \tag{2.41}$$

where $\mathcal{J}$ is the determinant of the Jacobian matrix $\dfrac{\partial \mathbf{x}}{\partial \mathbf{y}}$ with its (j,k)th element being $\dfrac{\partial x_j}{\partial y_k}$.

In the following, whenever we talk about transformation of a random variable or random vector, we always assume that the transformation is measurable so that the resulting function is also a random variable. The following theorem is about the transformation of two independent random variables.

Theorem 2.2.3 *Let X and Z be independent random variables, η_X be a function only of x and η_Z be a function only of z. Then the random variables $U = \eta_X(X)$ and $V = \eta_Z(Z)$ are independent.*

The above theorem is very important in theory, and it can be generalized as follows.

Theorem 2.2.4 *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ be mutually independent random vectors, and η_j be a function only of $\mathbf{x}_j$, $j = 1, 2, \ldots, m$. Then random variables $U_j = \eta_j(\mathbf{X}_j)$, $j = 1, 2, \ldots, m$, are mutually independent.*

2.3 Statistical Averages of Random Variables

The concepts of moments of a single random variable and the joint moments between any pair of random variables in a multi-dimensional set of random variables are of particular importance in practice. We begin the discussion of these statistical averages by considering first a single random variable X and its cumulative distribution function P. The *expectation* (also called an *expected value* or *mean*) of the random variable X is defined by

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x dP(x). \tag{2.42}$$

Here $\mathbb{E}(\cdot)$ is called the *expectation operator* (or statistical averaging operator), and the integral on the right side is interpreted as a Riemann–Stieltjes integral. We remark that the Riemann–Stieltjes integral is a generalization of the Riemann integral and its definition is given as follows.

Definition 2.3.1 *Let φ and h be real-valued functions defined on $[\underline{x}, \bar{x}] \subset \mathbb{R}$, $\{x_j^l\}_{j=0}^l$ be a partition of $[\underline{x}, \bar{x}]$, $\Delta_l = \max\limits_{0\le j\le l-1}\{x_{j+1}^l - x_j^l\}$, and $s_j^l \in [x_j^l, x_{j+1}^l]$ denote intermediate points of the partition. Then φ is said to be Riemann–Stieltjes integrable with respect to h on $[\underline{x}, \bar{x}]$ if*

$$\lim_{\substack{l\to\infty \\ \Delta_l\to 0}} \sum_{j=0}^{l-1} \varphi(s_j^l)[h(x_{j+1}^l) - h(x_j^l)] \tag{2.43}$$

exists and the limit is independent of the choice of the partition and their intermediate points. The limit of (2.43) *is called the Riemann–Stieltjes integral of φ with respect to h on $[\underline{x}, \bar{x}]$, and is denoted by $\int_{\underline{x}}^{\bar{x}} \varphi(x)dh(x)$; that is,*

$$\int_{\underline{x}}^{\bar{x}} \varphi(x)dh(x) = \lim_{\substack{l\to\infty \\ \Delta_l\to 0}} \sum_{j=0}^{l-1} \varphi(s_j^l)[h(x_{j+1}^l) - h(x_j^l)],$$

where φ and h are called the integrand and the integrator, respectively.

It is worth noting that the Riemann–Stieltjes integral $\int_{\underline{x}}^{\bar{x}} \varphi(x)dh(x)$ does not exist for all continuous functions φ on $[\underline{x}, \bar{x}]$ unless h has bounded variation. This is why (2.42) can be interpreted as a Riemann–Stieltjes integral (as P has bounded variation). In general, the Riemann–Stieltjes integral $\int_{\underline{x}}^{\bar{x}} \varphi(x)dh(x)$ exists if the following conditions are satisfied:

- The functions φ and h have no discontinuities at the same point $x \in [\underline{x}, \bar{x}]$.
- The function φ has bounded κ_φ-variation on $[\underline{x}, \bar{x}]$ and the function h has bounded κ_h-variation on $[\underline{x}, \bar{x}]$, where κ_φ and κ_h are some positive constants such that $\dfrac{1}{\kappa_\varphi} + \dfrac{1}{\kappa_h} > 1$.

We refer the interested reader to [18, Section 2.1], [21], and the references therein for more information on Riemann–Stieltjes integrals.

If we assume that P is absolutely continuous (that is, the corresponding probability density function p exists), then we can rewrite (2.42) as

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xp(x)\,dx. \tag{2.44}$$

We note from (2.12) and (2.15) that if X is a discrete random variable with range $\{x_j\}$, then the expectation of X is given by

$$\mathbb{E}(X) = \sum_j x_j \mathrm{Prob}\{X = x_j\}.$$

Thus, we see that with the help of the Dirac delta function one can put the definition of expectation in the same framework for a discrete random variable and for a continuous random variable with a probability density function (as we stated earlier). Since we are mainly interested in discrete random variables and those continuous random variables associated with probability density functions, we will define the statistical average of a random variable in terms of its probability density function in the following presentation.

The expectation of a random variable is also called the *first moment*. In general, the kth moment of a random variable X is defined as

$$\mathbb{E}(X^k) = \int_{-\infty}^{\infty} x^k p(x)\,dx.$$

We can also define the *central moments*, which are the moments of the difference between X and $\mathbb{E}(X)$. For example, the kth central moment of X is defined by

$$\mathbb{E}\left((X - \mathbb{E}(X))^k\right).$$

Of particular importance is the second central moment, called the *variance* of X, which is defined as

$$\sigma^2 = \mathrm{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 p(x)\,dx. \tag{2.45}$$

The square root σ of the variance of X is called the *standard deviation* of X. Variance is a measure of the "randomness" of the random variable X. It is

related to the first and second moments through the relationship

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left((X-\mathbb{E}(X))^2\right)\\ &= \mathbb{E}\left(X^2 - 2X\mathbb{E}(X) + (\mathbb{E}(X))^2\right)\\ &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2.\end{aligned} \tag{2.46}$$

One of the useful concepts in understanding the variation of a random variable X is the *coefficient of variation* (CV), which is defined as the ratio of the standard deviation to the mean (that is, $\mathrm{CV} = \sqrt{\mathrm{Var}(X)}/\mathbb{E}(X)$). It is the inverse of the so-called *signal-to-noise ratio*. A random variable with $\mathrm{CV} < 1$ is considered to have low variation, while one with $\mathrm{CV} > 1$ is considered to have high variation.

Note that the moments of a function of a random variable, $Y = \eta(X)$, can be defined in the same way as above. For example, the kth moment of Y is

$$\mathbb{E}(Y^k) = \int_{\Omega_y} y^k p_Y(y)dy, \tag{2.47}$$

where Ω_y denotes the range of Y, and p_Y is the probability density function of Y. However, due to the relation (2.40) between p_Y and the probability density function p_X of X, we can also calculate the kth moment of Y by

$$\mathbb{E}(Y^k) = \mathbb{E}\{\eta^k(X)\} = \int_{-\infty}^{\infty} \eta^k(x)p_X(x)dx. \tag{2.48}$$

For any real numbers a and b, by (2.48) we observe that

$$\mathbb{E}(aX+b) = a\mathbb{E}(X) + b, \quad \mathrm{Var}(aX+b) = a^2\mathrm{Var}(X). \tag{2.49}$$

Remark 2.3.1 *We remark that the infinite sequence of moments is in general not enough to uniquely determine a distribution function. Interested readers can refer to [7, Section 2.3] for a counterexample of two random variables having the same moments but different probability density functions. However, if two random variables have bounded support, then an infinite sequence of moments does uniquely determine the distribution function.*

2.3.1 Joint Moments

The definition of moments and central moments can be extended from one random variable to two or more random variables. In this context, the moments are often called joint moments, and central moments are often called joint central moments. For example, the joint moment of two random variables X and Y is defined as

$$\mathbb{E}(X^{k_x}Y^{k_y}) = \int_{\mathbb{R}^2} x^{k_x}y^{k_y}p(x,y)dxdy, \tag{2.50}$$

where k_x and k_y are positive integers, and p is the joint probability density function of X and Y. The joint central moment of X and Y is given by

$$\mathbb{E}\left((X-\mathbb{E}(X))^{k_x}(Y-\mathbb{E}(Y))^{k_y}\right). \tag{2.51}$$

However, the joint moment that is most useful in practical applications is the *correlation* of two random variables X and Y, defined as

$$\mathrm{Cor}\{X,Y\}=\mathbb{E}(XY)=\int_{\mathbb{R}^2} xyp(x,y)dxdy, \tag{2.52}$$

which implies that the second moment of X is the correlation of X with itself. Also of particular importance is the *covariance* of two random variables X and Y, defined as

$$\mathrm{Cov}\{X,Y\}=\mathbb{E}\left((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right)=\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y), \tag{2.53}$$

which indicates that the variance of X is the covariance of X with itself. It is worth noting that the correlation of two random variables should not be confused with the *correlation coefficient*, r, which is defined as the covariance of the two random variables divided by the product of their standard deviations.

$$r=\frac{\mathrm{Cov}\{X,Y\}}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

By (2.52) and (2.53) we see that both correlation and covariance are symmetric. That is,

$$\mathrm{Cor}\{X,Y\}=\mathrm{Cor}\{Y,X\},\quad \mathrm{Cov}\{X,Y\}=\mathrm{Cov}\{Y,X\}.$$

Moreover, they are both linear in each variable. That is, for any real numbers a and b we have

$$\begin{aligned}
\mathrm{Cor}\{aX+bY,Z\}&=a\mathrm{Cor}\{X,Z\}+b\mathrm{Cor}\{Y,Z\},\\
\mathrm{Cor}\{Z,aX+bY\}&=a\mathrm{Cor}\{Z,X\}+b\mathrm{Cor}\{Z,Y\},\\
\mathrm{Cov}\{aX+bY,Z\}&=a\mathrm{Cov}\{X,Z\}+b\mathrm{Cov}\{Y,Z\},\\
\mathrm{Cov}\{Z,aX+bY\}&=a\mathrm{Cov}\{Z,X\}+b\mathrm{Cov}\{Z,Y\}.
\end{aligned}$$

By the above equations, we find that the variance of $aX+bY$ is given by

$$\mathrm{Var}(aX+bY)=a^2\mathrm{Var}(X)+2ab\mathrm{Cov}\{X,Y\}+b^2\mathrm{Var}(Y).$$

Definition 2.3.2 *Two random variables X and Y are called uncorrelated if $Cov\{X,Y\}=0$.*

By the above definition and (2.53) we see that if

$$\mathbb{E}(XY)=\mathbb{E}(X)\mathbb{E}(Y), \tag{2.54}$$

then X and Y are uncorrelated. We observe from (2.29) that if X and Y are independent, then (2.54) also holds. Therefore, the independence of two random variables implies that these two random variables are uncorrelated. However, in general the converse is not true.

2.3.2 Conditional Moments

The kth *conditional moment* of X given $Y = y$ is

$$\mathbb{E}(X^k|Y = y) = \int_{-\infty}^{\infty} x^k \rho_{X|Y}(x|y)dx. \tag{2.55}$$

This implies that if random variables X and Y are independent (that is, $\rho_{X|Y}(x|y) = p_X(x)$), then

$$\mathbb{E}(X^k|Y = y) = \mathbb{E}(X^k).$$

The first conditional moment of X given $Y = y$ is called the *conditional expectation* of X given $Y = y$.

Observe that $\mathbb{E}(X|Y = y)$ is a function of y. Hence, $\mathbb{E}(X|Y)$ is a random variable, which is called the *conditional expectation* of X given Y. By (2.20), (2.27) and (2.55) we find

$$\begin{aligned}
\mathbb{E}\left(\mathbb{E}(X|Y)\right) &= \int_{-\infty}^{\infty} p_Y(y) \left(\int_{-\infty}^{\infty} x \rho_{X|Y}(x|y)dx \right) dy \\
&= \int_{-\infty}^{\infty} p_Y(y) \left(\int_{-\infty}^{\infty} x \frac{p_{XY}(x,y)}{p_Y(y)} dx \right) dy \\
&= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} p_{XY}(x,y)dy \right) dx \\
&= \int_{-\infty}^{\infty} x p_X(x)dx \\
&= \mathbb{E}(X),
\end{aligned}$$

which indicates that the expected value of the conditional expectation of X given Y is the same as the expected value of X. This formula is often called the *law of total expectation*. Similarly, we can show that for any positive integer k we have

$$\mathbb{E}\left(\mathbb{E}(X^k|Y)\right) = \mathbb{E}(X^k). \tag{2.56}$$

The *conditional variance* of X given Y is defined as

$$\mathrm{Var}(X|Y) = \mathbb{E}\left((X - \mathbb{E}(X|Y))^2|Y\right) = \mathbb{E}(X^2|Y) - \left(\mathbb{E}(X|Y)\right)^2. \tag{2.57}$$

By (2.46), (2.56) and (2.57), we find that the unconditional variance is related to the conditional variance by

$$\mathrm{Var}(X) = \mathbb{E}\left(\mathrm{Var}(X|Y)\right) + \mathrm{Var}\left(\mathbb{E}(X|Y)\right), \tag{2.58}$$

which indicates that the unconditional variance is equal to the sum of the mean of the conditional variance and the variance of the conditional mean. Equation (2.58) is often called the *law of total variance*, *variance decomposition formula* or *conditional variance formula*.

2.3.3 Statistical Averages of Random Vectors

For an m-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_m)^T$, its *mean vector* and *covariance matrix* are of particular importance. Specifically, the mean vector of $\mathbf{X}$ is given by

$$\boldsymbol{\mu} = \mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_m))^T,$$

and its *covariance matrix* is defined by

$$\boldsymbol{\Sigma} = \mathrm{Var}(\mathbf{X}) = \mathbb{E}\{(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\}. \tag{2.59}$$

Hence, we see that $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times m}$ is a non-negative definite matrix with its (k, j)th element being the covariance of random variables X_k and X_j:

$$\begin{aligned}\mathrm{Cov}\{X_k, X_j\} &= \mathbb{E}\left((X_k - \mathbb{E}(X_k))(X_j - \mathbb{E}(X_j))\right)\\ &= \int_{\mathbb{R}^2} (x_k - \mathbb{E}(X_k))(x_j - \mathbb{E}(X_j)) p_{X_k X_j}(x_k, x_j)\, dx_k dx_j,\end{aligned}$$

where $p_{X_k X_j}$ is the joint probability density function of X_k and X_j. For any $\mathcal{A} \in \mathbb{R}^{l\times m}$ and $\mathbf{a} \in \mathbb{R}^l$, it can be easily shown that

$$\mathbb{E}(\mathcal{A}\mathbf{X} + \mathbf{a}) = \mathcal{A}\mathbb{E}(\mathbf{X}) + \mathbf{a}, \quad \mathrm{Var}(\mathcal{A}\mathbf{X} + \mathbf{a}) = \mathcal{A}\mathrm{Var}(\mathbf{X})\mathcal{A}^T. \tag{2.60}$$

Similarly, we can extend the covariance of two random variables to the *cross-covariance matrix* between two random vectors. Let $\mathbf{X} = (X_1, X_2, \ldots, X_{m_x})^T$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_{m_y})^T$, where m_x and m_y are positive integers. Then the *cross-covariance matrix* between $\mathbf{X}$ and $\mathbf{Y}$ is

$$\begin{aligned}\mathrm{Cov}\{\mathbf{X}, \mathbf{Y}\} &= \mathbb{E}\left((\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{Y} - \mathbb{E}(\mathbf{Y}))^T\right)\\ &= \mathbb{E}(\mathbf{X}\mathbf{Y}^T) - \mathbb{E}(\mathbf{X})(\mathbb{E}(\mathbf{Y}))^T,\end{aligned} \tag{2.61}$$

which implies that $\mathrm{Var}(\mathbf{X}) = \mathrm{Cov}\{\mathbf{X}, \mathbf{X}\}$. By (2.60) it can be easily shown that for any matrix $\mathcal{A} \in \mathbb{R}^{\kappa_x\times m_x}$ and $\mathcal{B} \in \mathbb{R}^{\kappa_y\times m_y}$ (κ_x and κ_y are positive integers) we have

$$\mathrm{Cov}\{\mathcal{A}\mathbf{X}, \mathcal{B}\mathbf{Y}\} = \mathcal{A}\mathrm{Cov}\{\mathbf{X}, \mathbf{Y}\}\mathcal{B}^T. \tag{2.62}$$

The concept of uncorrelatedness of two random variables can also be extended to two random vectors. Specifically, two random vectors $\mathbf{X}$ and $\mathbf{Y}$ are called uncorrelated if $\mathrm{Cov}\{\mathbf{X}, \mathbf{Y}\} = 0$.

2.3.4 Important Inequalities

In this section, we list a few important inequalities that will be used in this monograph.