Mathematical Statistics

Guo-Liang TIAN

Department of Statistics and Data Science
Southern University of Science and Technology
Shenzhen, Guangdong Province, P.R. China

Xuejun JIANG

Department of Statistics and Data Science,
Southern University of Science and Technology
Shenzhen, Guangdong Province, P.R. China

Science Press, Beijing

1 December 2020

To Yanli, Margaret and Adam

To Zhengzhuo and Zhuomin

Preface

There are so many textbooks of mathematical statistics in the world, why do we need another one? We found that most traditional English textbooks of mathematical statistics seem not quite appropriate for readers such as typical undergraduates and instructors in the mainland of China. This is partly because they are generally big books with explanations beyond their level of English. Based on the first author's experience of teaching *Mathematical Statistics* course for 12 semesters at the Department of Statistic and Actuarial Science (September 2008–July 2016) of the University of Hong Kong (HKU), the Department of Mathematics (August 2016–August 2019) and the Department of Statistics and Data Science (September 2019–Present) of Southern University of Science and Technology (SUSTech), we have updated the original lecture notes from those 12 years and finally coined this textbook with the following unique features:

(1) It includes some important contents such as inverse Bayes formulae (§1.10), definition of valid categorical distribution (§1.11), and a unified expectation technique (Appendix B) which do not generally appear in other similar textbooks.

(2) This textbook is written in a style combining elements of the traditional book-writing with easy adaptation to ppt (Microsoft Office Power Point) presentation.

(3) Traditional textbooks typically have one line threading through chapters, sections and subsections, like the warp; while this text has another line, a self-contained weft. The advantage of this design is to make the structure of the whole book clear and precise so that relevant topics can be quickly captured even though the readers are not quite familiar with the book. At the same time, such a design makes it easy for readers to find and comb through the main topics.

(4) We heavily emphasized the motivation when introducing some fundamental concepts such as score function, Fisher information, the Cramér–Rao inequality, sufficiency, completeness and so on, and how to understand them. We made the best attempt to organize the contents to give students a good reading experience.

(5) It is quite user-friendly for instructors who want to adapt the materials from this book to make their own teaching ppt when using this textbook. In addition, we will provide online sample syllabus and tentative teaching plan, ten tutorials, five assignments with solutions, and 100 extra questions with solutions for instructors who choose the book as an undergraduate textbook.

(6) To facilitate the preparation of midterm/class test papers and final examination papers, we are preparing a companion book with 1000 exercises and their solutions.

The book is intended to be an undergraduate text for one semester with 48 lectures (3 lectures per week; 16 weeks $\times$ $3 = 48$ hours), where 10 lectures will be tutorials (10 hours) plus the midterm test (2 hours). Knowledge of calculus, linear algebra and probability is a prerequisite for this book, although reviews of basic materials in Chapter 1, Appendix A on distributions and Appendix C on Newton–Raphson and Fisher scoring algorithms, make the book quite self-contained. On successful completion of the course, students should be able to

— understand sufficient statistic(s) and its/their importance in data reduction and statistical inferences such as point estimation, confidence interval estimation, and testing hypothesis;

— derive maximum likelihood estimators of parameters to calculate maximum likelihood estimates;

— locate pivotal quantity to construct confidence intervals of parameters;

— find testing statistic to test hypotheses associated with one-sample and/or two-sample normal distributions with small sample sizes and non-normal distributions with large sample sizes.

A book such as this cannot be completed without substantial assistance from outside the team. The first author (GLT) would like to thank Professor Kai-Wang NG (Department of Statistics and Actuarial Science at HKU, HK) for discussing some important statistical concepts with him and comments on the first draft of the text. We are grateful to several undergraduates at HKU and SUSTech for their correcting some typos and grammatical errors. Professor Jun-Wu YU (School of Mathematics and Computational Science, Hunan University of Science and Technology), Dr. Chunling LIU (Department of Applied Mathematics at The Hong Kong Polytechnic University), Dr. Jing YAO (Department of Mathematics at SUSTech) and Dr. Tao LI (Department of Statistics and Data Science at SUSTech), Dr. Deng PAN (School of Mathematics and Statistics at Huazhong University of Science and Technology) used an earlier version of the text for a graduate/undergraduate course, respectively. Their constructive comments are greatly appreciated. We would like to thank Samuel HAO and Professor Jun-Wu YU who helped with the exercise solutions. We would also like to thank Professor Ming Tony TAN (Department of Biostatistics, Bioinformatics & Biomathematics at Georgetown University Medical Center, USA), Professor Jianxin PAN (School of Mathematics, University of Manchester, UK), Professor Yu FEI (School of Statistics and Mathematics at Yunnan University of Finance and Economics), and several anonymous reviewers who read the manuscript at various stages. Their reviews led to a much improvement version of the book.

Shenzhen Guo-Liang TIAN
Shenzhen Xuejun JIANG
December 2020

Contents

Chapter 1

Probability and Distributions

1.1 Probability

1.1.1 Permutation, combination and binomial coefficients

$1^{\bullet}$ SEVERAL NOTIONS

$1.1^{\bullet}$ **Factorial**

— The product $n(n-1) \times \cdots \times 3 \times 2 \times 1$ is represented by the symbol $n!$, which is read "n factorial."

$1.2^{\bullet}$ **Permutation**

— The number of permutations of n distinct objects taken r at a time is

$$ {}_nP_r = n(n-1) \times \cdots \times (n-r+1) = \frac{n!}{(n-r)!}, \quad r = 0, 1, \ldots, n. $$

$1.3^{\bullet}$ **Combination**

— The number of combinations of n distinct objects taken r at a time is

$$ \binom{n}{r} = \frac{n(n-1) \times \cdots \times (n-r+1)}{r!} = \frac{n!}{r!(n-r)!}, \quad r = 0, 1, \ldots, n. $$

$1.4^{\bullet}$ **Binomial coefficient**

— The binomial coefficient of the term $x^r y^{n-r}$ in the expansion of

$$ (x+y)^n = \sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r} $$

is $\binom{n}{r}$, where n is a positive integer and r is a non-negative integer that is smaller than or equal to n.

1.5• Multinomial coefficient

— The number of ways in which a set of n distinct objects can be partitioned into k subsets with n_1 objects in the first subset, n_2 objects in the second subset, . . . ,and n_k objects in the k-th subset is

$$\binom{n}{n_1, \ldots, n_k} = \frac{n!}{n_1! \cdots n_k!},$$

which is the multinomial coefficient of the term $x_1^{n_1} \cdots x_k^{n_k}$ in the expansion of $(x_1 + \cdots + x_k)^n$, where $n_1 + \cdots + n_k = n$.

2• Some useful formulae

- $\binom{n}{r} = \binom{n}{n-r}$.
- $\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}$.
- When $n = x$ is not a positive integer or zero, the generalized binomial coefficient is defined by
$$\binom{x}{r} = \frac{x(x-1)\cdots(x-r+1)}{r!}.$$
- $\binom{-x}{r} = (-1)^r \binom{x+r-1}{r}$ for any $x > 0$.
- $\binom{n}{r_1, \ldots, r_k} = \binom{n}{r_1}\binom{n-r_1}{r_2}\cdots\binom{n-r_1-\cdots-r_{k-1}}{r_k}$.

Example 1.1 (Some important identities). By equating the coefficients of x^n in the expressions on both sides of the equation

$$(1+x)^{a+b} = (1+x)^a(1+x)^b, \tag{1.1}$$

we have

$$\binom{a}{0}\binom{b}{n} + \binom{a}{1}\binom{b}{n-1} + \cdots + \binom{a}{n}\binom{b}{0} = \binom{a+b}{n}. \tag{1.2}$$

Especially, in (1.2) let $a = b = n$, we obtain

$$\binom{n}{0}\binom{n}{n} + \binom{n}{1}\binom{n}{n-1} + \cdots + \binom{n}{n}\binom{n}{0} = \binom{2n}{n}. \tag{1.3}$$

Note that $\binom{n}{r} = \binom{n}{n-r}$, then (1.3) becomes

$$\binom{n}{0}^2 + \binom{n}{1}^2 + \cdots + \binom{n}{n}^2 = \binom{2n}{n}. \tag{1.4}$$

If we compare the coefficients of x^{a-r} on both sides of (1.1), we have

$$\begin{aligned}\binom{a+b}{a-r} &= \sum_{i+j=a-r} \binom{a}{i}\binom{b}{j} = \sum_{k=0}^{a-r} \binom{a}{a-r-k}\binom{b}{k} \\ &= \sum_{k=0}^{a-r} \binom{a}{r+k}\binom{b}{k}. \end{aligned} \tag{1.5}$$

Furthermore, by differentiating the identity $(1+x)^n = \sum_{i=0}^{n} \binom{n}{i} x^i$ on both sides with respect to x, and setting $x = 1$, we obtain

$$n2^{n-1} = \sum_{i=1}^{n} i \binom{n}{i}. \tag{1.6}$$

Similarly, by differentiating the identity $(1-x)^n = \sum_{i=0}^{n} (-1)^i \binom{n}{i} x^i$ on both sides with respect to x, and substituting $x = 1$, we have

$$0 = \sum_{i=1}^{n} (-1)^i i \binom{n}{i}. \qquad \|$$

1.1.2 Sample space

3• OUTCOMES OF AN EXPERIMENT

- An experiment is a process of observation or measurement.
- The results obtained from an experiment are called *outcomes* of the experiment.
- The set of all possible outcomes of an experiment is called the *sample space*, denoted by $\mathbb{S}$.

- Each outcome in a sample space is called an *element* or a *sample point.*
- An *event* is a subset of a sample space.

4• Sample space

- According to the number of elements contained, sample spaces can be classified into *discrete* sample space and *continuous* sample space.

4.1• Discrete sample space

— A sample space is discrete, if the number of elements is finite or countable.

4.2• Continuous sample space

— A sample space is continuous, if the sample space consists of a continuum.

— For example, a set of real numbers includes both the rational numbers and the irrational numbers.

1.1.3 Events

5• Complement, union and intersection

- Given two events $\mathbb{A} \subset \mathbb{S}$ and $\mathbb{B} \subset \mathbb{S}$, we define three events as follows:

$$\begin{aligned} \mathbb{A}' &\hat{=} \{e\colon\ e \in \mathbb{S} \quad \text{but} \quad e \notin \mathbb{A}\}, \\ \mathbb{A} \cup \mathbb{B} &\hat{=} \{e\colon\ e \in \mathbb{A} \quad \text{or} \quad e \in \mathbb{B}\}, \quad \text{and} \\ \mathbb{A} \cap \mathbb{B} &\hat{=} \{e\colon\ e \in \mathbb{A} \quad \text{and} \quad e \in \mathbb{B}\}. \end{aligned}$$

  They are called the *complement* of $\mathbb{A}$, the *union* of $\mathbb{A}$ and $\mathbb{B}$, and the *intersection* of $\mathbb{A}$ and $\mathbb{B}$, respectively.

- Let $\emptyset$ denote the empty set. If $\mathbb{A} \cap \mathbb{B} = \emptyset$, then $\mathbb{A}$ and $\mathbb{B}$ are *mutually exclusive.*
- If $\mathbb{A} \subset \mathbb{B}$, then $\mathbb{A}$ is contained in $\mathbb{B}$ or $\mathbb{A}$ is a subset of $\mathbb{B}$.
- Figures 1.1 and 1.2 illustrate these concepts.

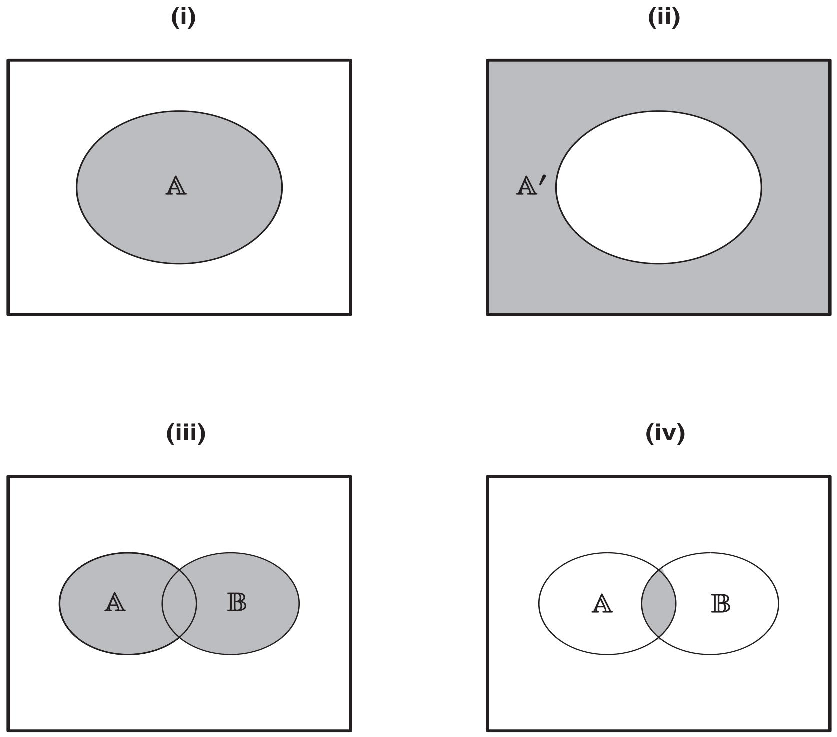

Figure 1.1 Venn diagrams. (i) $\mathbb{A}$; (ii) $\mathbb{A}'$; (iii) $\mathbb{A} \cup \mathbb{B}$; (iv) $\mathbb{A} \cap \mathbb{B}$.

1.1.4 Properties of probability

6• Definition of probability

Definition 1.1 (Probability of a set). Let $\mathbb{A}$ be a subset of the sample space $\mathbb{S}$, then $\Pr(\mathbb{A})$ is said to be the probability of $\mathbb{A}$ if

(1) $\Pr(\mathbb{A}) \geqslant 0$ and $\Pr(\mathbb{S}) = 1$;

(2) If $\{\mathbb{A}_1, \mathbb{A}_2, \ldots\}$ is a sequence of mutually exclusive events of $\mathbb{S}$, then

$$\Pr\left(\bigcup_{i=1}^{\infty} \mathbb{A}_i\right) = \sum_{i=1}^{\infty} \Pr(\mathbb{A}_i). \tag{1.7}$$

$\|$

7• Some properties of probability

Property 1.1 $\Pr(\emptyset) = 0$. $\|$

Proof. Since $\emptyset = \cup_{i=1}^{\infty} \emptyset$, from (1.7), we have $\Pr(\emptyset) = \sum_{i=1}^{\infty} \Pr(\emptyset)$ and thus $\Pr(\emptyset) = 0$. □

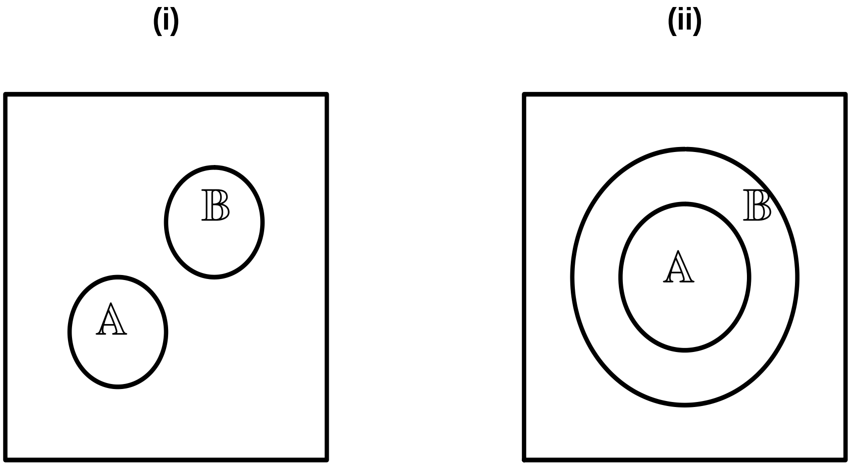

Figure 1.2 Diagrams showing special relationships among events. (i) $\mathbb{A}$ and $\mathbb{B}$ are mutually exclusive; (ii) $\mathbb{A}$ is contained in $\mathbb{B}$.

Property 1.2 If $\mathbb{A}_1, \ldots, \mathbb{A}_n$ are mutually exclusive, then

$$\Pr\left(\bigcup_{i=1}^{n} \mathbb{A}_i\right) = \sum_{i=1}^{n} \Pr(\mathbb{A}_i). \qquad \|$$

Proof. This is trivial, since $\cup_{i=1}^{n}\mathbb{A}_i = \mathbb{A}_1 \cup \cdots \cup \mathbb{A}_n \cup \emptyset \cup \emptyset \cup \cdots$. □

Property 1.3 $\Pr(\mathbb{A}') = 1 - \Pr(\mathbb{A})$. ‖

Property 1.4 If $\mathbb{A} \subseteq \mathbb{B}$, then $\Pr(\mathbb{A}) \leqslant \Pr(\mathbb{B})$. ‖

Proof. Note that $\mathbb{B} = \mathbb{A} \cup (\mathbb{A}' \cap \mathbb{B})$. Since $\mathbb{A}$ and $\mathbb{A}' \cap \mathbb{B}$ are mutually exclusive, then $\Pr(\mathbb{B}) = \Pr(\mathbb{A}) + \Pr(\mathbb{A}' \cap \mathbb{B}) \geqslant \Pr(\mathbb{A})$. □

Property 1.5 $0 \leqslant \Pr(\mathbb{A}) \leqslant 1$. ‖

Property 1.6 $\Pr(\mathbb{A} \cup \mathbb{B}) = \Pr(\mathbb{A}) + \Pr(\mathbb{B}) - \Pr(\mathbb{A} \cap \mathbb{B})$. ‖

Proof. We first partition $\mathbb{A} \cup \mathbb{B}$ into three mutually exclusive events $\mathbb{E}_1$, $\mathbb{E}_2$ and $\mathbb{E}_3$ as shown in Figure 1.3, then

$$\begin{aligned}
\Pr(\mathbb{A} \cup \mathbb{B}) &= \Pr(\mathbb{E}_1 \cup \mathbb{E}_2 \cup \mathbb{E}_3) \\
&= \Pr(\mathbb{E}_1) + \Pr(\mathbb{E}_2) + \Pr(\mathbb{E}_3) \\
&= \Pr(\mathbb{E}_1 \cup \mathbb{E}_3) + \Pr(\mathbb{E}_2 \cup \mathbb{E}_3) - \Pr(\mathbb{E}_3) \\
&= \Pr(\mathbb{A}) + \Pr(\mathbb{B}) - \Pr(\mathbb{A} \cap \mathbb{B}).
\end{aligned}$$

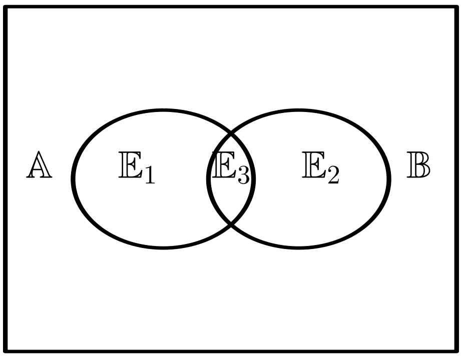

Figure 1.3 A partition of $\mathbb{A} \cup \mathbb{B}$. □

Property 1.7 $\Pr(\mathbb{A}\cup\mathbb{B}\cup\mathbb{C}) = \Pr(\mathbb{A})+\Pr(\mathbb{B})+\Pr(\mathbb{C})-\Pr(\mathbb{A}\cap\mathbb{B})-\Pr(\mathbb{B}\cap\mathbb{C})-\Pr(\mathbb{A}\cap\mathbb{C})+\Pr(\mathbb{A}\cap\mathbb{B}\cap\mathbb{C})$. In general, we have

$$\begin{aligned}\Pr\left(\bigcup_{i=1}^{n}\mathbb{A}_i\right) &= \sum_{i=1}^{n}\Pr(\mathbb{A}_i) - \sum_{i<j}\Pr(\mathbb{A}_i\cap\mathbb{A}_j)\\ &\quad + \sum_{i<j<k}\Pr(\mathbb{A}_i\cap\mathbb{A}_j\cap\mathbb{A}_k) - \cdots\\ &\quad + (-1)^{n+1}\Pr(\mathbb{A}_1\cap\mathbb{A}_2\cap\cdots\cap\mathbb{A}_n). \end{aligned}$$ ||

1.2 Conditional Probability

8• Definition of conditional probability

Definition 1.2 (Conditional probability of two sets). If $\mathbb{A}$ and $\mathbb{B}$ are two events in the sample space $\mathbb{S}$, the conditional probability of $\mathbb{B}$ given $\mathbb{A}$ is defined by

$$\Pr(\mathbb{B}|\mathbb{A}) = \frac{\Pr(\mathbb{A}\cap\mathbb{B})}{\Pr(\mathbb{A})}, \tag{1.8}$$

where $\Pr(\mathbb{A}) > 0$. ||

8.1• Equivalent formulae

— From (1.8), we immediately obtain

$$\Pr(\mathbb{A}\cap\mathbb{B}) = \Pr(\mathbb{A})\times\Pr(\mathbb{B}|\mathbb{A}), \quad \Pr(\mathbb{A})>0, \quad \text{and} \tag{1.9}$$

$$\Pr(\mathbb{A}\cap\mathbb{B}) = \Pr(\mathbb{B})\times\Pr(\mathbb{A}|\mathbb{B}), \quad \Pr(\mathbb{B})>0. \tag{1.10}$$

Example 1.2 (Car dealer service data). A research report of the services under warranty provided by 50 new-car dealers in a certain city is summarized in Table 1.1. Let $\mathbb{G}$ denote the selection of a dealer who provides good service under warranty, and let $\mathbb{T}$ denote the selection of a dealer who has been in business 10 years or more. The aim is to find $\Pr(\mathbb{G})$ and $\Pr(\mathbb{G}|\mathbb{T})$.

Table 1.1 Car dealer service data

Time in business	Service attitude		Total
	Good service ($\mathbb{G}$)	Poor service ($\mathbb{G}'$)	
$\geqslant 10$ years ($\mathbb{T}$)	16	4	20
< 10 years ($\mathbb{T}'$)	10	20	30
Total	26	24	50

Solution. Note that

$$\Pr(\mathbb{G}) = \frac{\#\text{ of favourable outcomes}}{\#\text{ of possible outcomes}} = \frac{26}{50} = \frac{13}{25}.$$

From (1.8), we have

$$\Pr(\mathbb{G}|\mathbb{T}) = \frac{\Pr(\mathbb{G}\cap\mathbb{T})}{\Pr(\mathbb{T})} = \frac{16/50}{20/50} = \frac{4}{5}. \qquad \|$$

9• Definition of independency of two events

Definition 1.3 (Independency of two events). Two events $\mathbb{A}$ and $\mathbb{B}$ are said to be *independent*, denoted by $\mathbb{A} \perp\!\!\!\perp \mathbb{B}$, if

$$\Pr(\mathbb{A}\cap\mathbb{B}) = \Pr(\mathbb{A}) \times \Pr(\mathbb{B}). \qquad \|$$

Theorem 1.1 (Independency). Let $\mathbb{A} \perp\!\!\!\perp \mathbb{B}$, then $\mathbb{A} \perp\!\!\!\perp \mathbb{B}'$ and $\mathbb{A}' \perp\!\!\!\perp \mathbb{B}'$. $\|$

Proof. Since $\mathbb{A} = (\mathbb{A}\cap\mathbb{B}) \cup (\mathbb{A}\cap\mathbb{B}')$, where $\mathbb{A}\cap\mathbb{B}$ and $\mathbb{A}\cap\mathbb{B}'$ are mutually exclusive, we obtain

$$\begin{aligned}
\Pr(\mathbb{A}\cap\mathbb{B}') &= \Pr(\mathbb{A}) - \Pr(\mathbb{A}\cap\mathbb{B}) \\
&= \Pr(\mathbb{A}) - \Pr(\mathbb{A}) \times \Pr(\mathbb{B}) \\
&= \Pr(\mathbb{A}) \times \Pr(\mathbb{B}'),
\end{aligned}$$

which indicates $\mathbb{A} \perp\!\!\!\perp \mathbb{B}'$. Using it again, we have $\mathbb{A}' \perp\!\!\!\perp \mathbb{B}'$. □

10• Mutual independency and pairwise independency

Definition 1.4 (Mutual independency). Events $\mathbb{A}_1, \ldots, \mathbb{A}_n$ are said to be mutually independent, if the probability of the intersection of any $2, 3, \ldots,$ or n of these events equals the product of their respective probabilities. ‖

10.1• Pairwise independency

— For $n = 3$, $\mathbb{A}_1, \mathbb{A}_2$ and $\mathbb{A}_3$ are mutually independent *if and only if* (iff)

$$\mathbb{A}_1 \perp\!\!\!\perp \mathbb{A}_2, \quad \mathbb{A}_1 \perp\!\!\!\perp \mathbb{A}_3, \quad \mathbb{A}_2 \perp\!\!\!\perp \mathbb{A}_3 \tag{1.11}$$

and

$$\Pr(\mathbb{A}_1 \cap \mathbb{A}_2 \cap \mathbb{A}_3) = \Pr(\mathbb{A}_1) \times \Pr(\mathbb{A}_2) \times \Pr(\mathbb{A}_3). \tag{1.12}$$

— Note that $\mathbb{A}_1, \mathbb{A}_2$ and $\mathbb{A}_3$ are called *pairwise* independent if (1.11) holds.

1.3 Bayes Theorem

11• Partition and bayes formula

Definition 1.5 (Partition). A partition of the sample space $\mathbb{S}$ is a collection of mutually exclusive sets $\mathbb{B}_1, \ldots, \mathbb{B}_n$ such that $\mathbb{S} = \cup_{i=1}^n \mathbb{B}_i$. ‖

Theorem 1.2 (Bayes formula). Let $\mathbb{B}_1, \ldots, \mathbb{B}_n$ be a partition of the sample space $\mathbb{S}$ and $\mathbb{A}$ be an event, then

(1) Law of total probability:

$$\Pr(\mathbb{A}) = \textstyle\sum_{i=1}^n \Pr(\mathbb{A}|\mathbb{B}_i) \Pr(\mathbb{B}_i). \tag{1.13}$$

(2) Bayes formulae:

$$\Pr(\mathbb{B}_j|\mathbb{A}) = \frac{\Pr(\mathbb{A}|\mathbb{B}_j) \Pr(\mathbb{B}_j)}{\sum_{i=1}^n \Pr(\mathbb{A}|\mathbb{B}_i) \Pr(\mathbb{B}_i)} \quad \text{for } j = 1, \ldots, n. \tag{1.14}$$

‖

Example 1.3 (Insurance data). An insurance company has three types of customers: high risk, medium risk, and low risk. Twenty percent of its customers are of high risk, 30% are of medium risk, and 50% are of low risk. The probability that a customer has at least one accident in the current year is 0.25 for high risk, 0.16 for medium risk, and 0.10 for low risk.

Let the events that a customer is high, medium, and low risk be $\mathbb{H}$, $\mathbb{M}$ and $\mathbb{L}$, respectively. Let the event that a customer has at least one accident in the current year be $\mathbb{A}$. Find $\Pr(\mathbb{A})$ and $\Pr(\mathbb{H}|\mathbb{A})$.

Solution. By the law of total probability, we have

$$\begin{aligned}\Pr(\mathbb{A}) &= \Pr(\mathbb{A}|\mathbb{H})\Pr(\mathbb{H}) + \Pr(\mathbb{A}|\mathbb{M})\Pr(\mathbb{M}) + \Pr(\mathbb{A}|\mathbb{L})\Pr(\mathbb{L}) \\ &= 0.25 \times 0.20 + 0.16 \times 0.30 + 0.10 \times 0.50 \\ &= 0.1480.\end{aligned}$$

By Bayes formula, we obtain

$$\Pr(\mathbb{H}|\mathbb{A}) = \frac{\Pr(\mathbb{A}|\mathbb{H})\Pr(\mathbb{H})}{\Pr(\mathbb{A})} = \frac{0.25 \times 0.20}{0.148} \approx 0.3378. \qquad \|$$

1.4 Probability Distributions

12• Discrete and continuous random variables

Definition 1.6 (Random variable). A *random variable* (r.v.) is a function from a sample space $\mathbb{S}$ into the real numbers. A r.v. is *discrete* if it takes values in a finite or countable set. A r.v. is *continuous* if it takes values over some interval. ‖

Definition 1.7 (Probability mass function). If X is a discrete r.v., the function defined by

$$p(x) = \Pr(X = x)$$

for each x within the range of X is called the *probability mass function* (pmf) of X. ‖

13• Basic properties of the pmf $p(x)$

- $p(x) > 0$.
- $\sum_x p(x) = 1$.

Definition 1.8 (Probability density function). Let X be a continuous r.v.. A non-negative function $f(x)$ is called the *probability density function* (pdf) of X, if

$$\Pr(X \in \mathbb{A}) = \int_{\mathbb{A}} f(x)\ \mathrm{d}x$$

for an arbitrary set $\mathbb{A}$ in the range of X. In particular, if the range of X is the real line (or one-dimensional Euclidean space) $\mathbb{R} = (-\infty, \infty)$, then

$$\Pr(a \leqslant X \leqslant b) = \int_a^b f(x)\ \mathrm{d}x. \qquad \|$$

14• Basic properties of the pdf $f(x)$

- $f(x) \geqslant 0$.
- $\int_{-\infty}^{\infty} f(x)\ \mathrm{d}x = 1$.

Definition 1.9 (Cumulative distribution function). The *cumulative distribution function* (cdf) of a r.v. X is defined by

$$F(x) = \Pr(X \leqslant x) = \begin{cases} \displaystyle\sum_{t \leqslant x} p(t), & \text{if } X \text{ is discrete,} \\ \displaystyle\int_{-\infty}^{x} f(t)\ \mathrm{d}t, & \text{if } X \text{ is continuous.} \end{cases} \qquad \|$$

15• Basic properties of the cdf $F(x)$

- $F(-\infty) = 0$ and $F(\infty) = 1$.
- If $a \leqslant b$, then $F(a) \leqslant F(b)$.
- If the range of a discrete r.v. X consists of the ordered values $x_1 < x_2 < \cdots < x_n$, then $p(x_1) = F(x_1)$ and $p(x_i) = F(x_i) - F(x_{i-1})$, $i = 2, 3, \ldots, n$.
- $f(x) = F'(x) \ \hat{=}\ \mathrm{d}F(x)/\,\mathrm{d}x$ if the density $f(x)$ exists.
- If X is continuous, then we have $\Pr(a < X < b) = \Pr(a \leqslant X < b) = \Pr(a < X \leqslant b) = \Pr(a \leqslant X \leqslant b) = F(b) - F(a) = \int_a^b f(x)\ \mathrm{d}x$.

Example 1.4 (Geometric distribution). Let X have the pmf $p(x) = 0.5^x$ for $x = 1, 2, \dots, \infty$. Find the cdf $F(x)$.

Solution. Let $S_n = \alpha + \alpha^2 + \cdots + \alpha^n$. First, we prove that

$$S_n = \begin{cases} \dfrac{\alpha(1-\alpha^n)}{1-\alpha}, & \text{if } \alpha \neq 1, \\ n, & \text{if } \alpha = 1. \end{cases} \tag{1.15}$$

$$\lim_{n\to\infty} S_n = \frac{\alpha}{1-\alpha}, \qquad \text{if } \alpha \in (0,1). \tag{1.16}$$

In fact, $\alpha S_n = \alpha^2 + \cdots + \alpha^n + \alpha^{n+1}$, then

$$S_n - \alpha S_n = \alpha - \alpha^{n+1}.$$

Thus, we immediately obtain (1.15) and (1.16).

Let $\alpha = 0.5$ in (1.16), we have $\sum_{x=1}^{\infty} p(x) = 1$, i.e., $p(x)$ is really a pmf. Furthermore, let $\alpha = 0.5$ in (1.15), we obtain

$$F(n) = \Pr(X \leqslant n) = \sum_{x=1}^{n} 0.5^x = 1 - 0.5^n,$$

$$F(x) = \begin{cases} 0, & \text{if } x < 1, \\ 1 - 0.5^n, & \text{if } n \leqslant x < n+1, \quad n = 1, 2, \dots. \end{cases}$$

Obviously, $F(\infty) = \lim_{n\to\infty}(1 - 0.5^n) = 1$. ‖

Example 1.5 (Exponential distribution). Let X have the pdf $f(x) = \lambda\,\mathrm{e}^{-3x}$ for $x \geqslant 0$, where λ is the normalizing constant.

1) Find the λ.

2) Find the cdf.

3) Evaluate $\Pr(0.5 \leqslant X < 1)$.

Solution. 1) Since $1 = \int_0^\infty f(x)\,\mathrm{d}x = \int_0^\infty \lambda\,\mathrm{e}^{-3x}\,\mathrm{d}x = \frac{\lambda}{3}$, we obtain $\lambda = 3$.

2) According to the relationship between the cdf and pdf, we have

$$F(x) = \int_0^x f(t)\,\mathrm{d}t = \int_0^x 3\,\mathrm{e}^{-3t}\,\mathrm{d}t = 1 - \mathrm{e}^{-3x}.$$

3) Therefore, we obtain

$$\Pr(0.5 \leqslant X < 1) = \int_{0.5}^{1} f(x)\,\mathrm{d}x = F(1) - F(0.5) = \mathrm{e}^{-1.5} - \mathrm{e}^{-3}. \qquad \|$$

1.5 Bivariate Distributions

1.5.1 Joint distribution

Definition 1.10 (Bivariate pmf). If X and Y are two discrete r.v.'s, the function defined by

$$p(x,y) = \Pr(X = x,\ Y = y)$$

for each pair of values (x,y) within the range of X and Y is called the joint pmf of X and Y. ‖

16• Basic properties of the joint pmf $p(x,y)$

- $p(x,y) > 0$.
- $\sum_x \sum_y p(x,y) = 1$.

Definition 1.11 (Bivariate pdf). A bivariate function $f(x,y)$ is called a joint pdf of the continuous r.v.'s X and Y if

$$\Pr\{(X,Y) \in \mathbb{A}\} = \int\int_{\mathbb{A}} f(x,y)\ \mathrm{d}x\ \mathrm{d}y$$

for a region $\mathbb{A}$ in the domain of (X,Y). ‖

17• Basic properties of the joint pdf $f(x,y)$

- $f(x,y) \geqslant 0$.
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\ \mathrm{d}x\ \mathrm{d}y = 1$.

Definition 1.12 (Bivariate cdf). The joint distribution (or joint cdf) of r.v.'s (X,Y) is defined by

$$\begin{aligned} F(x,y) &= \Pr(X \leqslant x,\ Y \leqslant y) \\ &= \begin{cases} \displaystyle\sum_{s \leqslant x}\sum_{t \leqslant y} p(s,t), & \text{if } X \text{ and } Y \text{ are discrete,} \\ \displaystyle\int_{-\infty}^{y}\int_{-\infty}^{x} f(s,t)\ \mathrm{d}s\ \mathrm{d}t, & \text{if } X \text{ and } Y \text{ are continuous.} \end{cases} \end{aligned}$$

For the continuous case, the joint pdf and cdf have the following relationship:

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} = \frac{\partial^2 F(x,y)}{\partial y \partial x}.$$ ‖

1.5.2 Marginal and conditional distributions

Definition 1.13 (Marginal pmfs and conditional pmfs). Let $p(x,y)$ be the joint pmf of discrete r.v.'s (X,Y). The *marginal* pmfs of X and Y are defined by

$$p(x) = \sum_y p(x,y) \quad \text{and} \quad p(y) = \sum_x p(x,y),$$

respectively. The *conditional* pmfs of X given $Y = y$ and Y given $X = x$ are defined by

$$p(x|y) = \frac{p(x,y)}{p(y)}, \quad p(y) \neq 0 \quad \text{and} \quad p(y|x) = \frac{p(x,y)}{p(x)}, \quad p(x) \neq 0,$$

respectively. ‖

Definition 1.14 (Marginal pdfs and conditional pdfs). Let $f(x,y)$ be the joint pdf of continuous r.v.'s (X,Y). The *marginal* pdfs of X and Y are defined by

$$f(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \quad \text{and} \quad f(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x,$$

respectively. The *conditional* pdfs of X given $Y = y$ and Y given $X = x$ are defined by

$$f(x|y) = \frac{f(x,y)}{f(y)}, \quad f(y) \neq 0 \quad \text{and} \quad f(y|x) = \frac{f(x,y)}{f(x)}, \quad f(x) \neq 0,$$

respectively. ‖

1.5.3 Independency of two random variables

Definition 1.15 (Independency of two r.v.'s). Let $f(x,y)$ denote the joint pdf of r.v.'s (X,Y), and $f(x)$ and $f(y)$ be their marginal pdfs. The r.v.'s X and Y are said to be *independent*, denoted by $X \perp\!\!\!\perp Y$, if

$$\begin{aligned} f(x,y) &= f(x) \times f(y), \quad \forall\,(x,y) \in \mathcal{S}_{(X,Y)}, \quad \text{or} \qquad (1.17) \\ F(x,y) &= F(x) \times F(y), \quad \forall\,(x,y) \in \mathcal{S}_{(X,Y)}, \qquad (1.18) \end{aligned}$$

where $\mathcal{S}_{(X,Y)} \hat{=} \{(x,y)\colon f(x,y) > 0\}$ denotes the joint *support* of (X,Y). In some cases, the joint pdf $f(x,y)$ may not exist, it is better to use (1.18) rather than (1.17) to define the independency of two r.v.'s. ‖

Example 1.6 (Uniform distribution in a circle). Let

$$f(x,y) = \begin{cases} \dfrac{1}{\pi r^2}, & \text{if } x^2+y^2 \leqslant r^2, \\ 0, & \text{otherwise.} \end{cases}$$

1) Find the marginal density $f(x)$.

2) Find the conditional density $f(y|x)$.

3) Calculate $\Pr(Y \geqslant 0.5r | X = 0.5r)$.

4) Are X and Y independent?

Solution. Let $I_{\mathbb{S}}(z)$ denote the *indicator function*, i.e., $I_{\mathbb{S}}(z) = 1$ if $z \in \mathbb{S}$ and $I_{\mathbb{S}}(z) = 0$ if $z \notin \mathbb{S}$. Note that the joint support of (X,Y) is $\mathcal{S}_{(X,Y)} = \{(x,y)\colon x^2+y^2 \leqslant r^2\}$, we can rewrite the joint pdf as

$$f(x,y) = \frac{1}{\pi r^2} \cdot I_{\mathcal{S}_{(X,Y)}}(x,y).$$

1) Since the marginal support of X is $\mathcal{S}_X = \{x\colon |x| \leqslant r\}$, then the marginal density of X is given by

$$f(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y = \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} f(x,y)\,\mathrm{d}y = \frac{2\sqrt{r^2-x^2}}{\pi r^2} \cdot I_{\mathcal{S}_X}(x).$$

By symmetry, we have $\mathcal{S}_Y = \{y\colon |y| \leqslant r\}$ and

$$f(y) = \frac{2\sqrt{r^2-y^2}}{\pi r^2} \cdot I_{\mathcal{S}_Y}(y).$$

2) Now, the conditional support $\mathcal{S}_{(Y|X=x)} = \{y\colon |y| \leqslant \sqrt{r^2-x^2}\,\}$. Thus, the conditional density is

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{1}{2\sqrt{r^2-x^2}} \cdot I_{\mathcal{S}_{(Y|X=x)}}(y).$$

3) When $X = 0.5r$, we have $\mathcal{S}_{(Y|X=0.5r)} = \{y\colon |y| \leqslant \sqrt{3}\,r/2\}$. Hence,

$$\begin{aligned} \Pr(Y \geqslant 0.5r|X=0.5r) &= \int_{0.5r}^{\sqrt{3}\,r/2} f(y|0.5r)\,\mathrm{d}y = \int_{0.5r}^{\sqrt{3}\,r/2} \frac{1}{\sqrt{3}\,r}\,\mathrm{d}y \\ &= \frac{1}{\sqrt{3}\,r}\left(\frac{\sqrt{3}\,r}{2} - \frac{r}{2}\right) = \frac{3-\sqrt{3}}{6}. \end{aligned}$$

4) Since $f(x,y) \neq f(x) \times f(y)$ for $(x,y) \in \mathcal{S}_{(X,Y)}$, X and Y are not independent. ‖

1.6 Expectation, Variance and Moments

1.6.1 Moments

18$^\bullet$ The general case

- Let X be a discrete (or continuous) r.v. with pmf $p(x)$ (or pdf $f(x)$).
- Let $g(\cdot)$ be an arbitrary function, then $g(X)$ itself is also a random variable.
- The expectation of $g(X)$ is defined by

$$E\{g(X)\} = \begin{cases} \displaystyle\sum_x g(x)p(x), & \text{if } X \text{ is discrete,} \\ \displaystyle\int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x, & \text{if } X \text{ is continuous,} \end{cases}$$

provided that $E\{|g(X)|\}$ exists, i.e., $E\{|g(X)|\} < +\infty$.

19$^\bullet$ Mean or expectation

- The expectation of X is defined as

$$\mu = E(X) = \begin{cases} \displaystyle\sum_x xp(x), & \text{if } X \text{ is discrete,} \\ \displaystyle\int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x, & \text{if } X \text{ is continuous,} \end{cases}$$

provided that $E(|X|) < +\infty$.

- It is a measure of the *central location* of the pdf of X.

19.1$^\bullet$ Some basic properties of expectation

— $E(c) = c$ for a constant c.

— $E\{cg(X)\} = cE\{g(X)\}$ for a constant c.

— $E\{\sum c_i g_i(X)\} = \sum c_i E\{g_i(X)\}$.

— $E\{g_1(X)\} \leqslant E\{g_2(X)\}$ if $g_1(x) \leqslant g_2(x) \quad \forall x$.

— If $X_1 \perp\!\!\!\perp X_2$, then $E(X_1X_2) = E(X_1)E(X_2)$.

20• Variance and standard deviation

- Let $\mu = E(X)$, then
$$\sigma^2 = \mathrm{Var}(X) = E(X-\mu)^2 = E(X^2) - \mu^2,$$
is a measure of the *dispersion* or variability of the pdf of X.

20.1• Some basic properties of variance

— $\mathrm{Var}(c) = 0$ for a constant c.

— $\mathrm{Var}\{cg(X)\} = c^2\mathrm{Var}\{g(X)\}$ for a constant c.

— $\mathrm{Var}(\sum c_i X_i) = \sum c_i^2 \mathrm{Var}(X_i) + 2\sum_{i<j} c_i c_j \mathrm{Cov}(X_i, X_j)$.

20.2• Standard deviation

— $\sigma = \sqrt{\mathrm{Var}(X)}$.

21• Covariance

- $\mathrm{Cov}(X_1, X_2) = E\{(X_1-\mu_1)(X_2-\mu_2)\}$, where $\mu_i = E(X_i)$, $i = 1, 2$.

21.1• Some basic properties of covariance

— $\mathrm{Cov}(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)$.

— If $X_1 \perp\!\!\!\perp X_2$, then $\mathrm{Cov}(X_1, X_2) = 0$.

22• Moments and central moments

- The r-th moment of the r.v. X is defined by $\mu_r' = E(X^r)$, which is also called the r-th *raw moment* or the r-th *moment about the origin*.

- The r-th central moment of the r.v. X is defined by $\mu_r = E(X-\mu)^r$, which is also called the r-th *central moment* or the r-th *moment about the mean*.

- It is easy to show that μ_r and μ_r' have the following relationship:
$$\mu_r = \sum_{i=0}^{r} (-1)^i \binom{r}{i} \mu_{r-i}' \mu^i. \tag{1.19}$$

23• Skewness and kurtosis

- The third central moment $\mu_3 = E(X-\mu)^3$ is a measure of *asymmetry* of the pdf of X.
- μ_3/σ^3 is called the *coefficient of skewness.*
- The fourth central moment $\mu_4 = E(X-\mu)^4$ is a measure of kurtosis, which is the *degree of flatness* of a density near its center.
- μ_4/σ^4, called the *coefficient of kurtosis*, is sometimes used to indicate that a density is more peaked around its center than the normal density.

24• Quantile and median

- The q-th quantile of a r.v. X with cdf $F(\cdot)$, denoted by ξ_q, is defined as the smallest real number ξ satisfying $F(\xi) = \Pr(X \leqslant \xi) \geqslant q$.
- If X is continuous, then the q-th quantile of X is defined as the smallest real number ξ satisfying $F(\xi) = \Pr(X \leqslant \xi) = q$.
- The 0.5-th quantile $\xi_{0.5}$ is defined as the median of X, denoted by $\text{med}(X)$.
- Alternatively, the median of X satisfies

$$\Pr\{X \leqslant \text{med}(X)\} \geqslant 0.5 \quad \text{and} \quad \Pr\{X \geqslant \text{med}(X)\} \geqslant 0.5.$$

- If X is a continuous r.v. with pdf $f(x)$, then the median of X satisfies

$$\int_{-\infty}^{\text{med}(X)} f(x)\,\mathrm{d}x = 0.5 = \int_{\text{med}(X)}^{\infty} f(x)\,\mathrm{d}x.$$

1.6.2 Some probability inequalities

25• A more general inequality

Theorem 1.3 (The general case). If X is a r.v. and $g(\cdot)$ is a non-negative function defined on the real line $\mathbb{R}$, then

$$\Pr\{g(X) \geqslant c\} \leqslant \frac{E\{g(X)\}}{c} \tag{1.20}$$

for any $c > 0$. ‖

Proof. We only consider the case that X is a continuous r.v. . Let the pdf of X be $f(x)$ and define $\mathbb{D} = \{x\colon\ g(x) \geqslant c\}$, we have

$$\begin{aligned} E\{g(X)\} &= \int_{-\infty}^{\infty} g(x)\cdot f(x)\,\mathrm{d}x \\ &= \int_{\mathbb{D}} g(x)\cdot f(x)\,\mathrm{d}x + \int_{\mathbb{D}'} g(x)\cdot f(x)\,\mathrm{d}x \\ &\geqslant c\int_{\mathbb{D}} f(x)\,\mathrm{d}x \\ &= c\,\Pr\{g(X)\geqslant c\}, \end{aligned}$$

so that (1.20) is proved. □

25.1• Chebyshev inequality

— Especially, in (1.20), by setting $g(X) = (X-\mu)^2$ and replacing c with $c^2\sigma^2$, we obtain the well–known Chebyshev inequality as follows.

— Let X be a r.v. and c be a positive constant, then

$$\Pr(|X-\mu| \geqslant c\,\sigma) \leqslant \frac{1}{c^2}, \tag{1.21}$$

where $\mu = E(X)$ and $\sigma^2 = \mathrm{Var}(X)$.

Example 1.7 (Three–point distribution). Let X be a discrete r.v. with pmf $\Pr(X=-1) = 1/(2c^2)$, $\Pr(X=0) = 1-1/c^2$, and $\Pr(X=1) = 1/(2c^2)$, where $c>1$. Then $E(X)=0$, $\mathrm{Var}(X) = 1/c^2$, and

$$\Pr(|X-\mu| \geqslant c\,\sigma) = \Pr(|X|\geqslant 1) = \Pr(X = 1 \text{ or } -1) = 1/c^2.$$

Thus equality holds in (1.21). ‖

25.2• Markov inequality

— Especially, in (1.20), by setting $g(X) = |X|^r$ and replacing c with c^r, we obtain Markov inequality as follows.

— Let X be a r.v. and c, r be two positive constants, then

$$\Pr(|X| \geqslant c) \leqslant \frac{E(|X|^r)}{c^r}, \tag{1.22}$$

provided that $E(|X|^r)$ exists.

26• Jensen's inequality

26.1• Convex function

— A continuous function $g(\cdot)$ defined on a subset $\mathbb{S}$ of the real line $\mathbb{R}$ is said to be *convex* if for any point $x_0 \in \mathbb{R}$, there exists a line which goes through the point $(x_0, g(x_0))$ and lies on or under the curve of the function $g(\cdot)$.

— A twice continuously differentiable function $g(x)$ is convex (or strictly convex) iff its second derivative $g''(x) \geqslant 0$ (or $g''(x) > 0$) for all $x \in \mathbb{S}$.

— A function $h(x)$ is *concave* (or strictly concave) iff $-h(x)$ is convex (or strictly convex).

Theorem 1.4 (Jensen's inequality). Let $g(\cdot)$ be a convex function. If X is a r.v. taking values in the domain of $g(\cdot)$, then

$$E\{g(X)\} \geqslant g(E(X)), \tag{1.23}$$

provided that both expectations $E(X)$ and $E\{g(X)\}$ exist. ‖

Proof. Since $g(x)$ is continuous and convex, there exists a straight line, say $\ell(x) = a + bx$, satisfying

$$\ell(x) = a + bx \leqslant g(x)$$

and

$$g(x_0) = g(E(X)) = \ell(E(X))$$

for any point $x_0 = E(X)$. Note that $E\{\ell(X)\} = E(a + bX) = a + bE(X) = \ell(E(X))$, we have

$$g(E(X)) = \ell(E(X)) = E\{\ell(X)\} \leqslant E\{g(X)\},$$

implying (1.23). □

Theorem 1.5 (Cauchy–Schwarz inequality). If two random variables X and Y have finite second moments, then

$$\{E(XY)\}^2 \leqslant E(X^2)E(Y^2), \tag{1.24}$$

with equality iff $\Pr(Y = cX) = 1$ for some constant c. ‖

Proof. Let

$$h(t) = E(tX - Y)^2 = E(X^2)t^2 - 2E(XY)t + E(Y^2),$$

then $h(t) \geqslant 0$ and it is a quadratic function of t.

If $h(t) > 0$, then the roots of $h(t)$ are not real; so

$$4\{E(XY)\}^2 - 4E(X^2)E(Y^2) < 0,$$

which means $\{E(XY)\}^2 < E(X^2)E(Y^2)$.

If $h(t) = 0$ for some t, say c, then $E(cX - Y)^2 = 0$, which implies $\Pr(Y = cX) = 1$. □

1.6.3 Conditional expectation

Definition 1.16 (Conditional expectation). Let X and Y be two r.v.'s and $p(x|y)$ (or $f(x|y)$) be the conditional pmf (or pdf) of X given $Y = y$, then the conditional expectation of $g(X)$ given $Y = y$ is

$$E\{g(X)|Y = y\} = \begin{cases} \sum_x g(x)p(x|y), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} g(x)f(x|y)\,\mathrm{d}x, & \text{if } X \text{ is continuous.} \end{cases} \qquad \|$$

27• Basic properties of $E\{g(X)|Y = y\}$

- $E\{g_1(X) + g_2(X)|Y = y\} = E\{g_1(X)|Y = y\} + E\{g_2(X)|Y = y\}$.
- $E\{g_1(X)g_2(Y)|Y = y\} = g_2(y)E\{g_1(X)|Y = y\}$.

28• Calculating $E(X)$ and $\mathrm{Var}(X)$ via $E(X|Y)$ and $\mathrm{Var}(X|Y)$

28.1• $E\{g(X)|Y\}$ is a random variable

— Note that $E\{g(X)|Y\}$ is a function of the r.v. Y, thus we can write

$$E\{g(X)|Y\} \mathrel{\hat{=}} h(Y),$$

which is also a random variable.

— That is, for any y in the range of Y, we have $E\{g(X)|Y = y\} = h(y)$.

28.2• Two general formulae

— For the continuous case, let $f(y)$ be the marginal density of Y, then

$$\begin{aligned} E[E\{g(X)|Y\}] &= E\{h(Y)\} = \int_{-\infty}^{\infty} h(y)f(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} E\{g(X)|Y=y\}f(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} g(x)f(x|y)\,\mathrm{d}x\right\}f(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= E\{g(X)\}. \end{aligned} \tag{1.25}$$

— Furthermore, we have

$$\begin{aligned} E\Big[\mathrm{Var}\{g(X)|Y\}\Big] &= E\Big[E\{g(X)g(X)|Y\} - E^2\{g(X)|Y\}\Big] \\ &\overset{(1.25)}{=} E\Big[\{g(X)\}^2\Big] - E\Big[\{h(Y)\}^2\Big] \\ &= E\Big[\{g(X)\}^2\Big] - \Big[\mathrm{Var}\{h(Y)\} + E^2\{h(Y)\}\Big] \\ &\overset{(1.25)}{=} E\{g(X)\}^2 - \mathrm{Var}\{h(Y)\} - E^2\{g(X)\} \\ &= \mathrm{Var}\{g(X)\} - \mathrm{Var}[E\{g(X)|Y\}]. \end{aligned}$$

That is,

$$\mathrm{Var}\{g(X)\} = E[\mathrm{Var}\{g(X)|Y\}] + \mathrm{Var}[E\{g(X)|Y\}]. \tag{1.26}$$

28.3• Two important formulae

— In particular, letting $g(X) = X$ in (1.25) and (1.26), we obtain

$$E(X) = E\{E(X|Y)\} = \int E(X|Y=y)\,f(y)\,\mathrm{d}y \quad \text{and} \tag{1.27}$$

$$\mathrm{Var}(X) = E\{\mathrm{Var}(X|Y)\} + \mathrm{Var}\{E(X|Y)\}. \tag{1.28}$$

1.6.4 Compound random variables

29• A compound random variable

- Let $\{X_1, X_2, \ldots\}$ be a sequence of *independent and identically distributed* (i.i.d.) r.v.'s that are independent of the non-negative integer-valued r.v. N.

- The random variable
$$S_N = \sum_{i=1}^{N} X_i \tag{1.29}$$
is called a *compound* random variable.

Theorem 1.6 (Expectation and variance of S_N). Let S_N be defined by (1.29), then

$$\begin{aligned} E(S_N) &= E(N)E(X) \quad \text{and} && (1.30) \\ \text{Var}(S_N) &= E(N)\text{Var}(X) + \text{Var}(N)\{E(X)\}^2, && (1.31) \end{aligned}$$

where the r.v. X has the same distribution with X_1. ‖

Proof. From (1.27), we have

$$E(S_N) = E\{E(S_N|N)\} = E\{NE(X)\} = E(N)E(X).$$

From (1.28), we obtain

$$\begin{aligned} \text{Var}(S_N) &= E\{\text{Var}(S_N|N)\} + \text{Var}\{E(S_N|N)\} \\ &= E\{N\text{Var}(X)\} + \text{Var}\{NE(X)\} \\ &= E(N)\text{Var}(X) + \text{Var}(N)\{E(X)\}^2, \end{aligned}$$

which means (1.31). □

1.6.5 Calculation of (conditional) probability via (conditional) expectation

30• Equivalence between probability and expectation

- Let $\mathbb{A}$ be an event and Y be a random variable. We would like to find probability $\Pr(\mathbb{A})$ and the conditional probability $\Pr(\mathbb{A}|Y)$.

- To do this, we first introduce an indicator r.v. associated with the event $\mathbb{A}$ as follows:

$$X = \begin{cases} 1, & \text{if the event } \mathbb{A} \text{ occurs with probability } \Pr(\mathbb{A}), \\ 0, & \text{if the event } \mathbb{A} \text{ does not occurs with probability } 1 - \Pr(\mathbb{A}). \end{cases}$$

- It is easy to see that X is a Bernoulli random variable.
- Hence, we have

$$E(X) = 1 \times \Pr(\mathbb{A}) + 0 \times \{1 - \Pr(\mathbb{A})\} = \Pr(\mathbb{A}) \tag{1.32}$$

and $E(X|Y) = \Pr(\mathbb{A}|Y)$.

- Moreover, we obtain

$$\begin{aligned} \Pr(\mathbb{A}) &= E(X) = E\{E(X|Y)\} = E\{\Pr(\mathbb{A}|Y)\} \\ &= \begin{cases} \sum_y \Pr(\mathbb{A}|Y = y)p(y), & \text{if } Y \text{ is discrete,} \\ \int \Pr(\mathbb{A}|Y = y)f(y)\,\mathrm{d}y, & \text{if } Y \text{ is continuous.} \end{cases} \end{aligned} \tag{1.33}$$

1.7 Moment Generating Function

31• Definition of mgf

Definition 1.17 (mgf). For a r.v. X, if $E(\mathrm{e}^{tX})$ exists for any $t \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$, then

$$M_X(t) = E(\mathrm{e}^{tX})$$

is called the *moment generating function* (mgf) of X. ‖

31.1• The relationship between mgf and the n-th moment

— By using Maclaurin's expansion, we have

$$M_X(t) = E\left\{\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right\} = \sum_{n=0}^{\infty} \frac{t^n}{n!}\,E(X^n) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\,\mu_n'.$$

— Thus, we obtain

$$\mu_n' = E(X^n) = \left.\frac{\mathrm{d}^n M_X(t)}{\mathrm{d}t^n}\right|_{t=0}. \tag{1.34}$$

31.2• Some basic properties of mgf

— $M_{a+bX}(t) = \mathrm{e}^{at} M_X(bt)$, where a and b are two constants.

— If $X \perp\!\!\!\perp Y$, then $M_{X+Y}(t) = M_X(t)M_Y(t)$.

— Jensen's inequality provides a lower bound: $M_X(t) \geqslant \exp\{tE(X)\}$.

Example 1.8 (Normal distribution). Assume that $X \sim N(\mu, \sigma^2)$, then

$$M_X(t) = \exp(\mu t + 0.5\sigma^2 t^2). \tag{1.35}$$

Solution. If $Z \sim N(0,1)$, then the mgf of Z is

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{\infty} \mathrm{e}^{tz} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-0.5z^2}\,\mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\{-0.5(z^2 - 2tz + t^2) + 0.5t^2\}\,\mathrm{d}z \\ &= \mathrm{e}^{0.5t^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\{-0.5(z-t)^2\}\,\mathrm{d}z = \mathrm{e}^{0.5t^2}. \end{aligned} \tag{1.36}$$

Since $X = \mu + \sigma Z$, we obtain

$$M_X(t) = M_{\mu+\sigma Z}(t) = \mathrm{e}^{\mu t} M_Z(\sigma t) = \mathrm{e}^{\mu t + 0.5\sigma^2 t^2}. \qquad \|$$

Table 1.2 Discrete probability distributions

Distribution	pmf $p(x)$	Parameter	mgf $M_X(t)$	$E(X)$	Var(X)
Binomial	$\binom{n}{x} p^x q^{n-x}$, $x = 0, 1, \ldots, n$	$0 < p < 1$	$(p\,\mathrm{e}^t + q)^n$	np	npq
Poisson	$\lambda^x\,\mathrm{e}^{-\lambda}/x!$, $x \in \mathbb{N}_0$	$\lambda > 0$	$\exp\{\lambda(\mathrm{e}^t - 1)\}$	λ	λ
Geometric	pq^{x-1}, $x \in \mathbb{N}_1$	$0 < p < 1$	$p\,\mathrm{e}^t/(1 - q\,\mathrm{e}^t)$	$1/p$	q/p^2
Negative binomial	$\binom{x-1}{r-1} p^r q^{x-r}$, $x \in \mathbb{N}_r$	$0 < p < 1$	$\{p\,\mathrm{e}^t/(1 - q\,\mathrm{e}^t)\}^r$	r/p	rq/p^2

NOTE: $q \mathrel{\hat=} 1 - p$ and $\mathbb{N}_r \mathrel{\hat=} \{r, r+1, \ldots, \infty\}$.

Table 1.3 Continuous probability distributions

Distribution	pdf $f(x)$	Parameter	mgf $M_X(t)$	$E(X)$	Var(X)
Uniform	$1/(b-a)$, $a \leqslant x \leqslant b$	$a < b$	$\dfrac{\mathrm{e}^{tb}-\mathrm{e}^{ta}}{t(b-a)}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Beta	Beta$(x\|a,b)$, $0 < x < 1$	$a > 0$, $b > 0$	–	$\dfrac{a}{a+b}$	$\dfrac{ab}{c^2(c+1)}$
Exponential	$\lambda\,\mathrm{e}^{-\lambda x}$, $x \geqslant 0$	$\lambda > 0$	$\lambda/(\lambda-t)$ $t < \lambda$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$
Gamma	Gamma$(x\|\alpha,\beta)$, $x > 0$	$\alpha > 0$, $\beta > 0$	$\{\beta/(\beta-t)\}^\alpha$ $t < \beta$	$\dfrac{\alpha}{\beta}$	$\dfrac{\alpha}{\beta^2}$
Normal	$N(x\|\mu,\sigma^2)$, $x \in \mathbb{R}$	$\mu \in \mathbb{R}$, $\sigma^2 > 0$	$\exp(\mu t+0.5\sigma^2 t^2)$	μ	σ^2

NOTE: $c \;\hat{=}\; a+b$.

Theorem 1.7 (Alternative to probability distribution). Two r.v.'s have the same mgf iff their distributions are identical. In other words, for all values of t, $M_X(t) = M_Y(t)$ iff $F_X(x) = F_Y(x)$ for all values of x (or equivalently $X \overset{\mathrm{d}}{=} Y$). ‖

31.3• A remark to Theorem 1.7

— The above statement is not equivalent to the statement "if two distributions have the same moments, then they are identical at all points."

— For some cases, the moments exist and yet the mgf does not, because

$$\sum_{n=0}^{\infty} \frac{t^n}{n!}\,\mu'_n = \lim_{n\to\infty} \sum_{i=0}^{n} \frac{t^i}{i!}\,\mu'_i$$

may not exist.

32• Relation to other functions

32.1• Characteristic function

— The mgf may not exist for some distributions.

— In general, we may apply the *characteristic function* (cf)

$$\varphi_X(t) = E(\mathrm{e}^{\mathrm{i}tX}) = M_{\mathrm{i}X}(t) = M_X(\mathrm{i}t),$$

which always exists, where $\mathrm{i}^2 = -1$.

32.2• Probability generating function

— If X is a discrete r.v. with support $\mathcal{S}_X$, then the *probability generating function* (pgf) of X is defined as

$$G_X(z) = E(z^X) = \sum_{x \in \mathcal{S}_X} z^x p_X(x),$$

where $p_X(\cdot)$ is the pmf of X.

— This immediately implies that $G_X(\mathrm{e}^t) = M_X(t)$.

1.8 Beta and Gamma Distributions

1.8.1 Beta distribution

33• Definition of beta distribution

Definition 1.18 (Beta density). A r.v. X is said to follow a beta distribution with parameters $a > 0$ and $b > 0$, if it has the pdf

$$\mathrm{Beta}(x|a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)}, \quad x \in \mathcal{S}_X, \tag{1.37}$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function and $\Gamma(\cdot)$ is the gamma function defined by (1.39). We will write $X \sim \mathrm{Beta}(a, b)$. $\|$

33.1• Support $\mathcal{S}_X$

— When $a = b = 1$, $\mathcal{S}_X = [0, 1]$. When $a \neq 1$ and $b \neq 1$, $\mathcal{S}_X = (0, 1)$.

— When $a \neq 1$ and $b = 1$, the pdf is ax^{a-1} so that $\mathcal{S}_X = (0, 1]$.

— When $a = 1$ and $b \neq 1$, the pdf is $b(1-x)^{b-1}$ so that $\mathcal{S}_X = [0, 1)$.

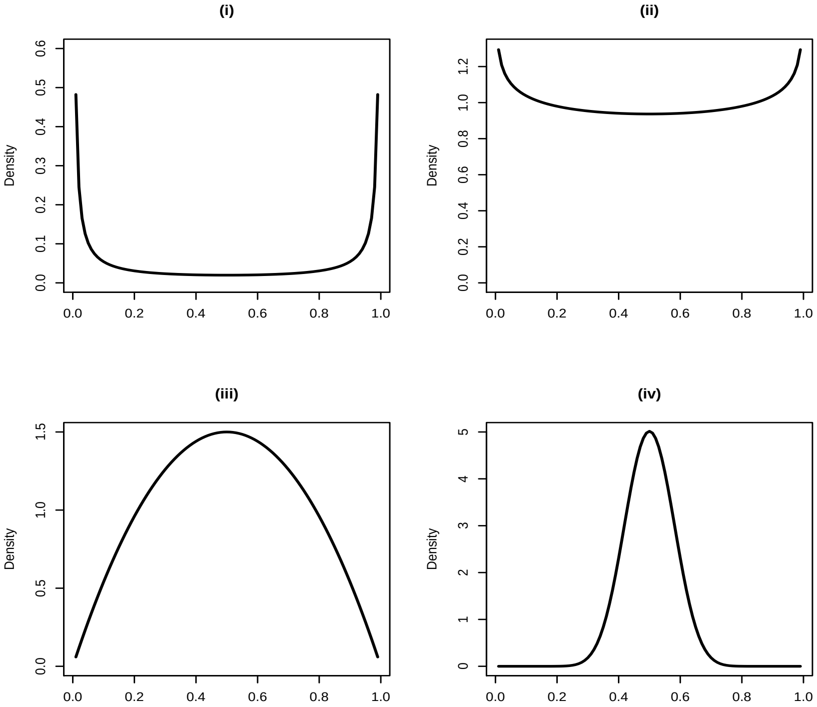

Figure 1.4 Plots of the densities of $X \sim \text{Beta}(a, a)$ with various parameter values. (i) $a = 0.01$; (ii) $a = 0.9$; (iii) $a = 2$; (iv) $a = 20$.

33.2• Densities

— Beta densities with various parameters are shown in Figure 1.4.

— Because $\int_0^1 \text{Beta}(x|a,b)\,\mathrm{d}x = 1$, we have the following identity:

$$\int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x = B(a,b).$$

33.3• Moments

— Let $X \sim \text{Beta}(a, b)$, then the r-th moment of X is given by

$$E(X^r) = \frac{B(a+r,b)}{B(a,b)} = \frac{\Gamma(a+r)}{\Gamma(a)} \cdot \frac{\Gamma(a+b)}{\Gamma(a+b+r)}. \qquad (1.38)$$

— Especially, we obtain

$$\begin{aligned} E(X) &= \frac{a}{a+b}, \quad E(X^2) = \frac{a(a+1)}{(a+b)(a+b+1)}, \quad \text{and} \\ \mathrm{Var}(X) &= \frac{ab}{(a+b)^2(a+b+1)}. \end{aligned}$$

33.4• Basic properties on $X \sim \mathrm{Beta}(a, b)$

— $\mathrm{Beta}(1, 1) = U[0, 1]$, and $1 - X \sim \mathrm{Beta}(b, a)$.

— The k-th order statistic (cf. Section 2.4) from a sample of n i.i.d. $U(0, 1)$ follows $\mathrm{Beta}(k,\ n - k + 1)$.

33.5• Usefulness

— The beta distribution is the conjugate prior for the binomial likelihood.

— A non-informative distribution is obtained as $a, b \to 0$.

— The built-in R function, `rbeta`(N, a, b), can be used to generate N i.i.d. samples from $\mathrm{Beta}(a, b)$.

1.8.2 Gamma distribution

34• The gamma function

- First of all, we briefly review the gamma function

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\, \mathrm{e}^{-x}\, \mathrm{d}x, \tag{1.39}$$

which is well defined for $\alpha > 0$.

34.1• Basic properties

— $\Gamma(1) = 1$, $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ and $\Gamma(1/2) = \sqrt{\pi}$.

— For a positive integer n, $\Gamma(n + 1) = n!$.

35• Definition of gamma distribution

Definition 1.19 (Gamma density). A r.v. X has a gamma distribution with shape parameter $\alpha > 0$ and rate parameter (1/rate is called scale parameter) $\beta > 0$, if its density function is given by

$$\mathrm{Gamma}(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\beta x}, \quad x \in \mathcal{S}_X. \tag{1.40}$$

When $\alpha \neq 1$, it is denoted by $X \sim \mathrm{Gamma}(\alpha, \beta)$ with $\mathcal{S}_X = (0, \infty)$. ‖

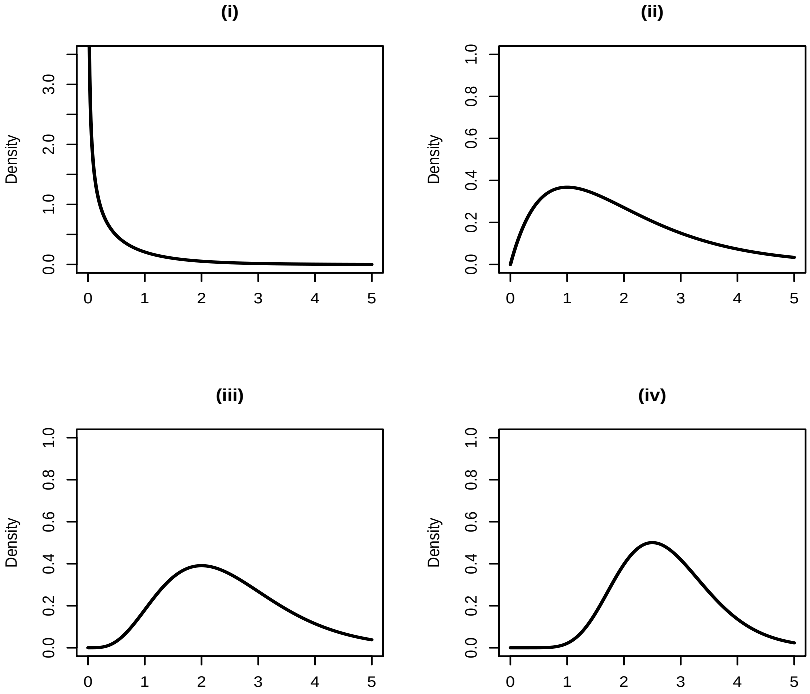

Figure 1.5 Plots of the densities of $X \sim \mathrm{Gamma}(\alpha, \beta)$ with various parameter values. (i) $\alpha = 0.5, \beta = 1$; (ii) $\alpha = 2, \beta = 1$; (iii) $\alpha = 5, \beta = 2$; (iv) $\alpha = 11, \beta = 4$.

35.1• Densities

— Gamma densities with various parameters are shown in Figure 1.5.

— By (1.39), we have

$$\int_0^\infty x^{\alpha-1} \mathrm{e}^{-\beta x} \, \mathrm{d}x = \frac{\Gamma(\alpha)}{\beta^\alpha}, \tag{1.41}$$

which implies $\int_0^\infty \mathrm{Gamma}(x|\alpha, \beta) \, \mathrm{d}x = 1$.

35.2• Moment generating function

— The mgf of X is given by

$$\begin{aligned} M_X(t) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \mathrm{e}^{tx} x^{\alpha-1} \mathrm{e}^{-\beta x}\,\mathrm{d}x = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} \mathrm{e}^{-(\beta-t)x}\,\mathrm{d}x \\ &\overset{(1.41)}{=} \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha)}{(\beta-t)^\alpha} = \left(\frac{\beta}{\beta-t}\right)^\alpha, \quad t < \beta. \end{aligned} \tag{1.42}$$

35.3• Moments

— Since

$$\begin{aligned} M_X'(t) &= \frac{\mathrm{d}M_X(t)}{\mathrm{d}t} = \alpha\beta^\alpha(\beta-t)^{-\alpha-1} \quad \text{and} \\ M_X''(t) &= \frac{\mathrm{d}^2 M_X(t)}{\mathrm{d}t^2} = \alpha(\alpha+1)\beta^\alpha(\beta-t)^{-\alpha-2}, \end{aligned}$$

we obtain

$$\begin{aligned} E(X) &= M_X'(0) = \frac{\alpha}{\beta}, \quad E(X^2) = M_X''(0) = \frac{\alpha(\alpha+1)}{\beta^2} \quad \text{and} \\ \mathrm{Var}(X) &= E(X^2) - \{E(X)\}^2 = \frac{\alpha}{\beta^2}. \end{aligned}$$

35.4• Other properties of $X \sim \mathrm{Gamma}(\alpha, \beta)$

— $\mathrm{Gamma}(1, \beta) = \mathrm{Exponential}(\beta)$ with pdf $\beta\,\mathrm{e}^{-\beta x}$ for $x \geqslant 0$.

— $\mathrm{Gamma}(n/2, 1/2) = \chi^2(n)$, the chi-squared distribution with n degrees of freedom.

— If $X \sim \mathrm{Gamma}(\alpha, \beta)$ and $c > 0$, then $Y = cX \sim \mathrm{Gamma}(\alpha, \beta/c)$.

— If $\{X_i\}_{i=1}^n \overset{\text{ind}}{\sim} \mathrm{Gamma}(\alpha_i, \beta)$, then $\sum_{i=1}^n X_i \sim \mathrm{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$.

— If $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} \mathrm{Exponential}(\beta)$, then $\sum_{i=1}^n X_i \sim \mathrm{Gamma}(n, \beta)$.

36• A useful formula

- Let $X \sim \mathrm{Gamma}(\alpha, \beta)$, then

$$E\{\log(X)\} = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log(\beta). \tag{1.43}$$

Proof. Let $Y = \beta X$, then $Y \sim \text{Gamma}(\alpha, 1)$ and

$$E\{\log(Y)\} = \log(\beta) + E\{\log(X)\}.$$

Differentiating both sides of the identity $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}\,\mathrm{e}^{-y}\,\mathrm{d}y$ with respect to α, we obtain

$$\Gamma'(\alpha) = \int_0^\infty y^{\alpha-1}\log(y)\cdot \mathrm{e}^{-y}\,\mathrm{d}y, \qquad [\because (a^x)' = a^x\log(a)]$$

or

$$\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \int_0^\infty \log(y)\cdot\frac{1}{\Gamma(\alpha)}y^{\alpha-1}\,\mathrm{e}^{-y}\,\mathrm{d}y = E\{\log(Y)\},$$

which indicates (1.43). □

1.9 Bivariate Normal Distribution

1.9.1 Univariate normal distribution

37• Definition of univariate normal distribution

- It is well known that X is normally distributed with mean μ and variance σ^2, denoted by $X \sim N(\mu, \sigma^2)$, if its pdf is

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty. \qquad (1.44)$$

37.1• Densities

— Normal densities with various parameters are shown in Figure 1.6.

— From (1.44), we have $I \hat{=} \int_{-\infty}^{+\infty}\mathrm{e}^{-0.5x^2}\,\mathrm{d}x = \sqrt{2\pi}$, which can be proved by calculating the following integral via polar transformation:

$$I^2 = \int \mathrm{e}^{-0.5x^2}\,\mathrm{d}x \cdot \int \mathrm{e}^{-0.5y^2}\,\mathrm{d}y = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\mathrm{e}^{-0.5(x^2+y^2)}\,\mathrm{d}x\,\mathrm{d}y.$$

37.2• Basic properties of $X \sim N(\mu, \sigma^2)$

— If $\mu = 0$ and $\sigma = 1$, then $E(X^{2r+1}) = 0$ and $E(X^{2r}) = (2r)!/(2^r r!)$.

Proof. On the one hand, if $g(\cdot)$ is an odd function defined in $(-\infty, \infty)$, then $\int_{-\infty}^{\infty} g(x)\,\mathrm{d}x = 0$. We immediately obtain the first formula. On the other hand, if $g(\cdot)$ is an even function defined in $(-\infty, \infty)$, then

$$\int_{-\infty}^{\infty} g(x)\,\mathrm{d}x = 2\int_{0}^{\infty} g(x)\,\mathrm{d}x.$$

By combining this fact with the identity (1.41), we obtain the second formula. □

— $a + bX \sim N(a + b\mu, b^2\sigma^2)$.

— In particular, $(X - \mu)/\sigma \sim N(0, 1)$.

— $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

— $M_X(t) = \exp(\mu t + 0.5\sigma^2 t^2)$.

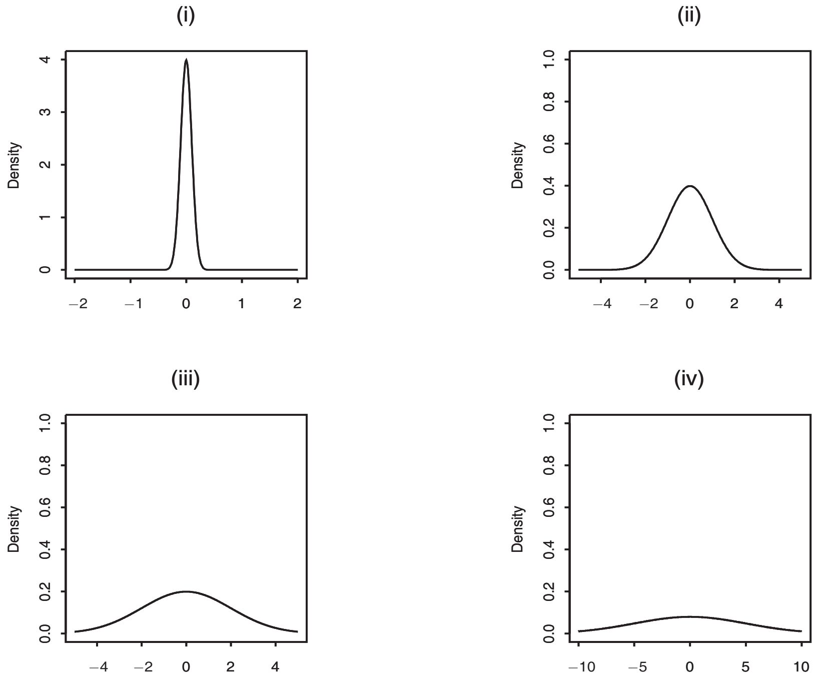

Figure 1.6 Plots of the densities of $X \sim N(0, \sigma^2)$ with various variances. (i) $\sigma = 0.1$; (ii) $\sigma = 1$; (iii) $\sigma = 2$; (iv) $\sigma = 5$.

1.9.2 Correlation coefficient

38• Definition of correlation coefficient

- To introduce the bivariate normal distribution, first of all, we introduce the concept of *correlation coefficient.*

- Given two r.v.'s X_1 and X_2 with $E(X_1) = \mu_1, E(X_2) = \mu_2, \mathrm{Var}(X_1) = \sigma_1^2$ and $\mathrm{Var}(X_2) = \sigma_2^2$, the covariance of X_1 and X_2 is

$$\mathrm{Cov}(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2).$$

- The correlation coefficient of X_1 and X_2 is defined by

$$\rho = \mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_1 \sigma_2}. \tag{1.45}$$

38.1• Using the Cauchy–Schwarz inequality to prove $|\rho| \leqslant 1$

— The correlation coefficient ρ defined by (1.45) can be rewritten as

$$\rho = \frac{E(X_1 - \mu_1)(X_2 - \mu_2)}{\sqrt{E(X_1 - \mu_1)^2 \cdot E(X_2 - \mu_2)^2}}.$$

— In (1.24), let $X = X_1 - \mu_1$ and $Y = X_2 - \mu_2$, we obtain

$$\frac{\{E(X_1 - \mu_1)(X_2 - \mu_2)\}^2}{E(X_1 - \mu_1)^2 \cdot E(X_2 - \mu_2)^2} \leqslant 1,$$

i.e., $\rho^2 \leqslant 1$ so that $-1 \leqslant \rho \leqslant 1$. □

1.9.3 Joint density

39• Multivariate case

- A random vector $\mathbf{x} = (X_1, \dots, X_d)^\top$ is said to follow a d-dimensional normal distribution, if its joint pdf is given by

$$N_d(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi}\,)^d |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\}, \tag{1.46}$$

for $\boldsymbol{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$, where the mean vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)^\top = (E(X_1), \dots, E(X_d))^\top \in \mathbb{R}^d$ and the variance–covariance matrix $\boldsymbol{\Sigma}$ is a positive definite matrix, denoted by $\boldsymbol{\Sigma} > 0$ (which is equivalent to $\boldsymbol{y}^\top \boldsymbol{\Sigma} \boldsymbol{y} > 0$ for any non-zero vector $\boldsymbol{y}$).

- We will write $\mathbf{x} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}_{d\times d} = (\sigma_{ij}) = (\text{Cov}(X_i, X_j))$ or

$$\boldsymbol{\Sigma} = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_d) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_d, X_1) & \text{Cov}(X_d, X_2) & \cdots & \text{Var}(X_d) \end{pmatrix},$$

  where $\text{Cov}(X_i, X_i) = \text{Var}(X_i)$ for $i = 1, \dots, d$.

39.1$^\bullet$ Two-dimensional case

— Especially, when $d = 2$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

we use $(X_1, X_2)^\top \sim N_2(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2; \rho)$ to represent the bivariate normal distribution.

— Its joint pdf is then given by

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{Q(x_1, x_2)}{2(1-\rho^2)} \right\}, \ x_1, x_2 \in \mathbb{R}, \quad (1.47)$$

where

$$Q(x_1, x_2) = \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1 - \mu_1}{\sigma_1}\right)\left(\frac{x_2 - \mu_2}{\sigma_2}\right) + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2.$$

39.2$^\bullet$ Marginal and conditional densities

— For any joint pdf $f(x_1, x_2)$, we have $f(x_1, x_2) = f(x_1) \times f(x_2|x_1)$.

— From (1.47), we can write

$$\begin{aligned} & \frac{1}{2(1-\rho^2)} Q(x_1, x_2) \\ = \ & \frac{1}{2(1-\rho^2)} \left\{ \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1 - \mu_1}{\sigma_1}\right)\left(\frac{x_2 - \mu_2}{\sigma_2}\right) + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 \right\} \\ = \ & \frac{(x_1 - \mu_1)^2}{2\sigma_1^2}\left(1 + \frac{\rho^2}{1-\rho^2}\right) + \frac{1}{2(1-\rho^2)}\left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 \end{aligned}$$

$$
\begin{aligned}
& + \frac{1}{2(1-\rho^2)} \left\{ -2\rho \left(\frac{x_1-\mu_1}{\sigma_1} \right) \left(\frac{x_2-\mu_2}{\sigma_2} \right) \right\} \\
= & \ \frac{(x_1-\mu_1)^2}{2\sigma_1^2} + \frac{(x_2-\mu_2)^2}{2\sigma_2^2(1-\rho^2)} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{2\sigma_1\sigma_2(1-\rho)^2} + \frac{\rho^2(x_1-\mu_1)^2}{2\sigma_1^2(1-\rho^2)} \\
= & \ \frac{(x_1-\mu_1)^2}{2\sigma_1^2} + \frac{\{x_2-\mu_2-\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)\}^2}{2\sigma_2^2(1-\rho^2)}.
\end{aligned}
$$

— Thus, we can decompose (1.47) into

$$
\begin{aligned}
f(x_1,x_2) &= \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\left\{ -\frac{(x_1-\mu_1)^2}{2\sigma_1^2} \right\} \\
&\quad \times \frac{1}{\sqrt{2\pi}\,\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{\{x_2-\mu_2-\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)\}^2}{2\sigma_2^2(1-\rho^2)} \right],
\end{aligned}
$$

which indicates

$$X_1 \sim N(\mu_1,\ \sigma_1^2) \quad \text{and} \tag{1.48}$$

$$X_2|(X_1=x_1) \sim N\left(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1),\ \sigma_2^2(1-\rho^2)\right). \tag{1.49}$$

— By symmetry, we also have

$$X_2 \sim N(\mu_2,\ \sigma_2^2) \quad \text{and} \tag{1.50}$$

$$X_1|(X_2=x_2) \sim N\left(\mu_1+\rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2),\ \sigma_1^2(1-\rho^2)\right). \tag{1.51}$$

39.3• Basic properties of $(X_1,X_2)^\top \sim N_2(\mu_1,\mu_2;\sigma_1^2,\sigma_2^2,\rho)$

— $E(X_i)=\mu_i$ and $\text{Var}(X_i)=\sigma_i^2$ for $i=1,2$.

— $E(X_1|X_2=x_2)=\mu_1+\rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2), \text{Var}(X_1|X_2=x_2)=\sigma_1^2(1-\rho^2)$,

$E(X_2|X_1=x_1)=\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1), \text{Var}(X_2|X_1=x_1)=\sigma_2^2(1-\rho^2)$.

— $\text{Cov}(X_1,X_2)=\rho\sigma_1\sigma_2$ and $\text{Corr}(X_1,X_2)=\rho$.

Proof. It is clear that

$$\begin{aligned}
\mathrm{Cov}(X_1, X_2) &= E\{(X_1-\mu_1)(X_2-\mu_2)\} \\
&\overset{(1.25)}{=} E\Big[E\{(X_1-\mu_1)(X_2-\mu_2)|X_2\}\Big] \\
&= E\Big[(X_2-\mu_2)E\{(X_1-\mu_1)|X_2\}\Big] \\
&= E\Big\{(X_2-\mu_2)\cdot\rho\frac{\sigma_1}{\sigma_2}(X_2-\mu_2)\Big\} \\
&= \rho\frac{\sigma_1}{\sigma_2}\cdot\mathrm{Var}(X_2) \\
&= \rho\sigma_1\sigma_2,
\end{aligned}$$

so that $\mathrm{Corr}(X_1, X_2) = \rho$. □

— $M_{(X_1, X_2)}(t_1, t_2) = \exp\Big\{\mu_1 t_1 + \mu_2 t_2 + 0.5(\sigma_1^2 t_1^2 + \sigma_2^2 t_2^2 + 2\rho\sigma_1\sigma_2 t_1 t_2)\Big\}$.

Proof. Let $t^* = \rho\frac{\sigma_1}{\sigma_2}t_1 + t_2$, we have

$$\begin{aligned}
& M_{(X_1, X_2)}(t_1, t_2) \\
=\ & E(\mathrm{e}^{t_1X_1+t_2X_2}) \\
\overset{(1.25)}{=}\ & E\Big\{E(\mathrm{e}^{t_1X_1+t_2X_2}|X_2)\Big\} \\
=\ & E\Big\{\mathrm{e}^{t_2X_2}E(\mathrm{e}^{t_1X_1}|X_2)\Big\} \\
\overset{(1.51)}{=}\ & E\Big[\mathrm{e}^{t_2X_2}\cdot\mathrm{e}^{\{\mu_1+\rho\frac{\sigma_1}{\sigma_2}(X_2-\mu_2)\}t_1+0.5\sigma_1^2(1-\rho^2)t_1^2}\Big] \\
=\ & \mathrm{e}^{\mu_1t_1-\rho\frac{\sigma_1}{\sigma_2}\mu_2t_1+0.5\sigma_1^2(1-\rho^2)t_1^2}\cdot E(\mathrm{e}^{t^*X_2}) \\
=\ & \mathrm{e}^{\mu_1t_1-\rho\frac{\sigma_1}{\sigma_2}\mu_2t_1+0.5\sigma_1^2(1-\rho^2)t_1^2}\cdot\mathrm{e}^{\mu_2t^*+0.5\sigma_2^2t^{*2}} \\
=\ & \exp\Big\{\mu_1t_1+\mu_2t_2+0.5(\sigma_1^2t_1^2+\sigma_2^2t_2^2+2\rho\sigma_1\sigma_2t_1t_2)\Big\},
\end{aligned}$$

which completes the proof. □

— $X_1 \perp\!\!\!\perp X_2$ iff $\rho = 0$.

— $a_1X_1 + a_2X_2 \sim N\Big(a_1\mu_1 + a_2\mu_2,\ a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1a_2\rho\sigma_1\sigma_2\Big)$.

1.9.4 Stochastic representation of random variables or random vectors

$40^\bullet$ Definition of stochastic representation

- Let X and $Y_1, \dots, Y_n$ be r.v.'s and $g(\cdot)$ a function.
- If X and $g(Y_1, \dots, Y_n)$ have the same distribution, denoted by
$$X \overset{\mathrm{d}}{=} g(Y_1, \dots, Y_n), \tag{1.52}$$
we say (1.52) is a one-to-many *stochastic representation* (SR) of the r.v. X.
- The symbol '$\overset{\mathrm{d}}{=}$' means that the r.v.'s on both sides of the equality have the same distribution.
- If $X = Y$, then we have $X \overset{\mathrm{d}}{=} Y$. However, conversely, it is not true. That is, $X \overset{\mathrm{d}}{=} Y \not\Rightarrow X = Y$.

$40.1^\bullet$ Several examples on the operator '$\overset{\mathrm{d}}{=}$'

— If $X \sim N(0,1)$, then $X \overset{\mathrm{d}}{=} -X$. However, $X \neq -X$.

— If $U \sim U(0,1)$, then $U \overset{\mathrm{d}}{=} 1 - U$. However, $U \neq 1 - U$.

— If $X \sim \text{Exponential}(\beta)$ and $U \sim U(0,1)$, then $X \overset{\mathrm{d}}{=} -\log(U)/\beta$.

— If $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} \text{Exponential}(\beta)$ and $Y \sim \text{Gamma}(n, \beta)$, then $Y \overset{\mathrm{d}}{=} \sum_{i=1}^n X_i$.

$41^\bullet$ Definition of a multivariate normal distribution via SR

- Let $Y_1, \dots, Y_d \overset{\text{iid}}{\sim} N(0,1)$ or $(Y_1, \dots, Y_d)^\top \sim N_d(\mathbf{0}, \boldsymbol{I}_d)$.
- Let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)^\top \in \mathbb{R}^d$ and $\boldsymbol{A}$ be a $d \times d$ square matrix. Define
$$\begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix} + \boldsymbol{A} \begin{pmatrix} Y_1 \\ \vdots \\ Y_d \end{pmatrix}, \tag{1.53}$$
we say that $\mathbf{x} = (X_1, \dots, X_d)^\top$ follows a d-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and variance–covariance matrix $\boldsymbol{A}\boldsymbol{A}^\top$, denoted by $\mathbf{x} \sim N_d(\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{A}^\top)$, where $\boldsymbol{A}\boldsymbol{A}^\top$ is not necessarily positive definite.

- It is clear that (1.53) implies $\mathbf{x} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + \boldsymbol{A}\mathbf{y}$ with $\mathbf{y} = (Y_1, \dots, Y_d)^{\top}$.

41.1• One advantage of (1.53) over (1.46)

— If $\boldsymbol{A}\boldsymbol{A}^{\top}$ is positive definite (i.e., $\boldsymbol{A}\boldsymbol{A}^{\top} > 0$), then $(\boldsymbol{A}\boldsymbol{A}^{\top})^{-1}$ exists, and the joint density of $\mathbf{x}$ exists.

— It can be shown by the transformation technique to be introduced in Section 2.1.2 that the joint pdf of $\mathbf{x}$ is exactly given by (1.46) with $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^{\top}$.

— It is clear that $\boldsymbol{A}\boldsymbol{A}^{\top}$ is always positive semi-definite, i.e., $\boldsymbol{A}\boldsymbol{A}^{\top} \geqslant 0$.

— If the determinant of $\boldsymbol{A}\boldsymbol{A}^{\top}$ is zero (i.e., $|\boldsymbol{A}\boldsymbol{A}^{\top}| = 0$), then the joint density of $\mathbf{x}$ does not exist but the distribution still exists.

— To define a multivariate normal distribution, we can see that using the SR (1.53) is better than using the joint density (1.46).

41.2• Other advantages of (1.53)

— By means of (1.53), it is much easier to show that $E(\mathbf{x}) = \boldsymbol{\mu}$,

— $\mathrm{Var}(\mathbf{x}) = \mathrm{Cov}(\mathbf{x}, \mathbf{x}) = \boldsymbol{A}\boldsymbol{A}^{\top}$, and

— $M_{\mathbf{x}}(\boldsymbol{t}) = E(\mathrm{e}^{\boldsymbol{t}^{\top}\mathbf{x}}) = \exp(\boldsymbol{t}^{\top}\boldsymbol{\mu} + 0.5\,\boldsymbol{t}^{\top}\boldsymbol{A}\boldsymbol{A}^{\top}\boldsymbol{t})$.

Proof. Let $\boldsymbol{s} = \boldsymbol{A}^{\top}\boldsymbol{t}$, we have

$$\begin{aligned} M_{\mathbf{x}}(\boldsymbol{t}) &= E(\mathrm{e}^{\boldsymbol{t}^{\top}\mathbf{x}}) = E[\exp\{\boldsymbol{t}^{\top}(\boldsymbol{\mu} + \boldsymbol{A}\mathbf{y})\}] \\ &= \exp(\boldsymbol{t}^{\top}\boldsymbol{\mu}) \cdot E\{\exp(\boldsymbol{s}^{\top}\mathbf{y})\} \\ &= \exp(\boldsymbol{t}^{\top}\boldsymbol{\mu}) \cdot E\{\exp(s_1 Y_1 + \cdots + s_d Y_d)\} \\ &= \exp(\boldsymbol{t}^{\top}\boldsymbol{\mu}) \cdot \prod_{i=1}^{d} E\{\exp(s_i Y_i)\} \\ &= \exp(\boldsymbol{t}^{\top}\boldsymbol{\mu}) \cdot \prod_{i=1}^{d} \exp(0.5\, s_i^2) \\ &= \exp(\boldsymbol{t}^{\top}\boldsymbol{\mu}) \cdot \exp(0.5\, \boldsymbol{s}^{\top}\boldsymbol{s}) \\ &= \exp(\boldsymbol{t}^{\top}\boldsymbol{\mu} + 0.5\, \boldsymbol{t}^{\top}\boldsymbol{A}\boldsymbol{A}^{\top}\boldsymbol{t}), \end{aligned}$$

which completes the proof. □

— Partition $\mathbf{x}$ into two parts

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{pmatrix}, \quad \text{where} \quad \mathbf{x}^{(1)} = \begin{pmatrix} X_1 \\ \vdots \\ X_r \end{pmatrix} \quad \text{and} \quad \mathbf{x}^{(2)} = \begin{pmatrix} X_{r+1} \\ \vdots \\ X_d \end{pmatrix}.$$

We can partition $\boldsymbol{\mu}$ and $\mathbf{y}$ in the same fashion. From (1.53), we obtain

$$\mathbf{x}^{(k)} \overset{\mathrm{d}}{=} \boldsymbol{\mu}^{(k)} + \boldsymbol{A}^{(k)}\mathbf{y}, \quad k = 1, 2,$$

indicating that $\mathbf{x}^{(k)} \sim N(\boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}\boldsymbol{A}^{(k)\top})$ for $k = 1, 2$, where

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}^{(1)} \\ \boldsymbol{A}^{(2)} \end{pmatrix}$$

with $\boldsymbol{A}^{(1)}$: $r \times d$ and $\boldsymbol{A}^{(2)}$: $(d - r) \times d$.

1.10 Inverse Bayes Formulae

1.10.1 Three inverse Bayes formulae

42• The issue

- Let two conditional densities $f_{(X|Y)}(x|y)$ and $f_{(Y|X)}(y|x)$ be known.
- Next, suppose that the assumption of $\mathcal{S}_{(X,Y)} = \mathcal{S}_X \times \mathcal{S}_Y$ is valid, where $\mathcal{S}_X$, $\mathcal{S}_Y$ and $\mathcal{S}_{(X,Y)}$ denote the marginal supports of X, Y and the joint support of (X, Y), respectively.
- What are the marginal density $f_X(x)$ and hence the joint density of (X, Y)?

42.1• What is the definition of support?

— Let $f_X(x)$ be the pdf of the r.v. X, then $\mathcal{S}_X = \{x\colon f_X(x) > 0\}$ is called the *support* of X.

— For example, if $X \sim U(0, 1)$, then the pdf of X is

$$f_X(x) = \begin{cases} 1, & \text{if } 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Hence, $\mathcal{S}_X = (0, 1)$.

— Let $f_{(X,Y)}(x,y)$ denote the joint density of (X,Y), then

$$\mathcal{S}_{(X,Y)} = \{(x,y)\colon\ f_{(X,Y)}(x,y) > 0\}$$

is called the *joint support* of (X,Y), see Definition 1.15 on page 14.

42.2• The product space and non-product space

— If $\mathcal{S}_{(X,Y)} = \mathcal{S}_X \times \mathcal{S}_Y$, we say $\mathcal{S}_{(X,Y)}$ is a *product space*; otherwise, it is called a *non-product space*.

— For example, let $(X,Y) \sim N_2(\mathbf{0}, \boldsymbol{I})$, then $\mathcal{S}_{(X,Y)} = \mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \mathcal{S}_X \times \mathcal{S}_Y$; that is, $\mathcal{S}_{(X,Y)}$ is a product space.

— If $(X,Y) \sim U(\mathbb{B}_2)$, where $\mathbb{B}_2 = \{(x,y)\colon\ x^2 + y^2 \leqslant 1\}$, then $\mathcal{S}_{(X,Y)} = \mathbb{B}_2 \neq [-1,1] \times [-1,1] = \mathcal{S}_X \times \mathcal{S}_Y$; that is, $\mathcal{S}_{(X,Y)}$ is a non-product space.

42.3• The point-wise formula

— We always have the following identity:

$$f_{(X|Y)}(x|y) f_Y(y) = f_{(Y|X)}(y|x) f_X(x), \quad (x,y) \in \mathcal{S}_{(X,Y)}. \tag{1.54}$$

— From (1.54), by division, we obtain

$$f_Y(y) = \frac{f_{(Y|X)}(y|x)}{f_{(X|Y)}(x|y)} \cdot f_X(x), \qquad x \in \mathcal{S}_X, \quad y \in \mathcal{S}_Y. \tag{1.55}$$

— Integrating this identity with respect to y on support $\mathcal{S}_Y$, i.e.,

$$\int_{\mathcal{S}_Y} f_Y(y)\ \mathrm{d}y = \int_{\mathcal{S}_Y} \frac{f_{(Y|X)}(y|x)}{f_{(X|Y)}(x|y)} \cdot f_X(x)\ \mathrm{d}y = f_X(x) \cdot \int_{\mathcal{S}_Y} \frac{f_{(Y|X)}(y|x)}{f_{(X|Y)}(x|y)}\ \mathrm{d}y,$$

we immediately have the following *point-wise formula*:

$$f_X(x) = \left\{ \int_{\mathcal{S}_Y} \frac{f_{(Y|X)}(y|x)}{f_{(X|Y)}(x|y)}\ \mathrm{d}y \right\}^{-1}, \quad \text{for any } x \in \mathcal{S}_X. \tag{1.56}$$

42.4• The function-wise formula

— Now substituting (1.56) into (1.55), we obtain the dual form of *inverse Bayes formula* (IBF) for $f_Y(y)$ and hence by symmetry we obtain the *function-wise formula* of $f_X(x)$:

$$f_X(x) = \left\{ \int_{\mathcal{S}_X} \frac{f_{(X|Y)}(x|y_0)}{f_{(Y|X)}(y_0|x)} \,\mathrm{d}x \right\}^{-1} \frac{f_{(X|Y)}(x|y_0)}{f_{(Y|X)}(y_0|x)}, \tag{1.57}$$

for all $x \in \mathcal{S}_X$ and an arbitrarily fixed $y_0 \in \mathcal{S}_Y$.

42.5• The sampling-wise formula

— By dropping the normalizing constant in (1.57), we obtain the so-called *sampling-wise formula*:

$$f_X(x) \propto \frac{f_{(X|Y)}(x|y_0)}{f_{(Y|X)}(y_0|x)}, \tag{1.58}$$

for all $x \in \mathcal{S}_X$ and an arbitrarily fixed $y_0 \in \mathcal{S}_Y$.

42.6• Discrete versions of (1.56) and (1.58)

— When both X and Y are discrete random variables, we have

$$\Pr(X = x) = \left\{ \sum_{y \in \mathcal{S}_Y} \frac{\Pr(Y = y|X = x)}{\Pr(X = x|Y = y)} \right\}^{-1}, \tag{1.59}$$

for any $x \in \mathcal{S}_X$, which is called the discrete version of the point-wise formula.

— The discrete version of the sampling-wise formula is

$$\Pr(X = x) \propto \frac{\Pr(X = x|Y = y_0)}{\Pr(Y = y_0|X = x)}, \tag{1.60}$$

for all $x \in \mathcal{S}_X$ and an arbitrarily fixed $y_0 \in \mathcal{S}_Y$.

1.10.2 Understanding the IBF

43• Why do they have the name of IBF?

- To answer this question, we first introduce Bayes formula or Bayes Theorem.

- From (1.55), we obtain

$$f_{(Y|X)}(y|x) = \frac{f_{(X|Y)}(x|y)f_Y(y)}{f_X(x)} = \frac{f_{(X|Y)}(x|y)f_Y(y)}{\int f_{(X|Y)}(x|y)f_Y(y)\,\mathrm{d}y}.$$

- Replacing y by θ and setting $f_Y(y) = \pi(\theta)$, the above identity becomes

$$p(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)} = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\,\mathrm{d}\theta}, \tag{1.61}$$

  which is called *Bayes formula*, where the parameter θ is treated as a random variable in Bayesian statistics, $\pi(\theta)$ is the prior density of θ, $p(\theta|x)$ is the posterior density of θ after the x was observed.

- The formula (1.61) states that given $f(x|\theta)$ and $\pi(\theta)$, the posterior density $p(\theta|x)$ can be determined uniquely.

- While (1.56)–(1.58) state that given both conditional densities, the marginal density can also be determined uniquely. This is why the name of IBF is taken.

44• Differences among the three IBF

- If we could obtain the *closed-form expression* for the integral in (1.56) or (1.57) (e.g., see Example 1.9 below), the three formulae (1.56)–(1.58) should have the same name.

- However, in practice, it may not be true.

- We need to evaluate the integral numerically by using Monte Carlo method.

44.1• What does the "point-wise" mean?

— To evaluate the integral in (1.56), we must *fix the value of* x, say $x = x_0 = 0.5$, because the integration is with respect to the variable y.

— In other words, given a point $x = x_0$, (1.56) can only be used to calculate the value of the density $f_X(x)$ evaluated at $x = x_0$, i.e., $f_X(x_0)$.

— For example, given $\{x_i\}_{i=1}^n$, say $n = 100$, we can calculate the values of $\{f_X(x_i)\}_{i=1}^n$ by performing n integrations.

— That is why (1.56) is called the point-wise formula.

44.2• What does the "function-wise" mean?

— The normalizing constant in (1.57) is $f_Y(y_0)$, which can be obtained by performing one integration like in (1.56).

— In other words, by only performing one integration, we can obtain the expression of the marginal density $f_X(x)$.

— That is why (1.57) is called the function-wise formula.

44.3• What does the "sampling-wise" mean?

— A density $f(x) = c \cdot g(x)$ can be factorized into two parts: c and $g(x)$, where $c = 1/\int g(x)\,\mathrm{d}x$ is called the *normalizing constant* and $g(x)$ is called the *kernel* of $f(x)$.

— If we would like to generate random samples from $f_X(x)$, there are many methods.

— For some methods (e.g., the acceptance–rejection algorithm, the grid method, the sampling/important resampling algorithm), it is not necessary to know the value of the normalizing constant.

— And it only needs to know the kernel. In (1.58), we only know the kernel of $f_X(x)$.

— That is why (1.58) is called the sampling-wise formula.

44.4• Remarks

— Often in practice, we know $f_{(X|Y)}(x|y)$ only up to a normalizing constant.

— In other words, $f_{(X|Y)}(x|y) = c(y) \cdot g(x|y)$, where $c(y)$ is unknown and $g(x|y)$ is completely known, then the function-wise IBF (1.57) and sampling-wise IBF (1.58) still hold if we replace $f_{(X|Y)}(x|y_0)$ by $g(x|y_0)$.

1.10.3 Two examples

Example 1.9 (Bivariate normal distribution). Assume that

$$\begin{aligned} X|(Y=y) &\sim N(\mu_1+\rho(y-\mu_2),\, 1-\rho^2) \quad \text{and} \\ Y|(X=x) &\sim N(\mu_2+\rho(x-\mu_1),\, 1-\rho^2). \end{aligned}$$

Find the marginal distribution of X and the joint distribution of (X, Y).

Solution. Note that $\mathcal{S}_{(X,Y)} = \mathcal{S}_{(X|Y)} \times \mathcal{S}_Y$. Since $\mathcal{S}_{(X|Y)} = \mathcal{S}_X = \mathbb{R}$, we have $\mathcal{S}_{(X,Y)} = \mathcal{S}_X \times \mathcal{S}_Y = \mathbb{R}^2$. From (1.56), we obtain

$$\{f_X(x)\}^{-1} = \sqrt{2\pi}\exp\{(x-\mu_1)^2/2\},$$

which means $X \sim N(\mu_1, 1)$. Therefore, the joint distribution of (X, Y) exists and is bivariate normal with means μ_1 and μ_2, unit variances and correlation coefficient ρ.

Alternative solution. When using (1.56), we need to evaluate an integral. In contrast, using (1.58), the integration can be avoided. In fact, let the arbitrary y_0 be μ_2, then

$$f_X(x) \propto \exp\{-(x-\mu_1)^2/2\}. \qquad \|$$

Example 1.10 (Bivariate discrete distribution). Let X be a discrete random variable with probability mass function (pmf) $p_i = \Pr(X = x_i)$ for $i = 1, 2, 3$ and Y be a discrete random variable with pmf $q_j = \Pr(Y = y_j)$ for $j = 1, 2, 3$. Given two conditional distribution matrices

$$\boldsymbol{A} = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} 1/6 & 0 & 3/14 \\ 0 & 1/4 & 4/14 \\ 5/6 & 3/4 & 7/14 \end{pmatrix}$$

and

$$\boldsymbol{B} = (b_{ij}) = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} = \begin{pmatrix} 1/4 & 0 & 3/4 \\ 0 & 1/3 & 2/3 \\ 5/18 & 6/18 & 7/18 \end{pmatrix},$$

where the (i, j) element of $\boldsymbol{A}$ is $a_{ij} = \Pr\{X = x_i|Y = y_j\}$ and the (i, j) element of $\boldsymbol{B}$ is $b_{ij} = \Pr\{Y = y_j|X = x_i\}$.

1) Find the marginal distributions of X and Y.

2) Find the joint distribution of (X, Y).

Solution. 1) The support of X and Y are $\mathcal{S}_X = \{x_1, x_2, x_3\}$ and $\mathcal{S}_Y = \{y_1, y_2, y_3\}$. By using (1.60) with $y_0 = y_3$, the X-marginal is given by

$$\begin{aligned}
p_1 &\hat{=} \Pr(X = x_1) = f_X(x_1) \\
&\propto \frac{f_{(X|Y)}(x_1|y_0)}{f_{(Y|X)}(y_0|x_1)} = \frac{\Pr(X = x_1|Y = y_3)}{\Pr(Y = y_3|X = x_1)} \\
&= \frac{a_{13}}{b_{13}} = \frac{3/14}{3/4} = \frac{4}{14}, \\
p_2 &\hat{=} \Pr(X = x_2) = f_X(x_2) \\
&\propto \frac{f_{(X|Y)}(x_2|y_0)}{f_{(Y|X)}(y_0|x_2)} = \frac{\Pr(X = x_2|Y = y_3)}{\Pr(Y = y_3|X = x_2)} \\
&= \frac{a_{23}}{b_{23}} = \frac{4/14}{2/3} = \frac{6}{14}, \\
p_3 &\hat{=} \Pr(X = x_3) = f_X(x_3) \\
&\propto \frac{f_{(X|Y)}(x_3|y_0)}{f_{(Y|X)}(y_0|x_3)} = \frac{\Pr(X = x_3|Y = y_3)}{\Pr(Y = y_3|X = x_3)} \\
&= \frac{a_{33}}{b_{33}} = \frac{7/14}{7/18} = \frac{18}{14}.
\end{aligned}$$

Note that $p_1 + p_2 + p_3 = 1$, we obtain

$$\begin{aligned}
p_1 &= \frac{4/14}{4/14 + 6/14 + 18/14} = \frac{4}{4 + 6 + 18} = \frac{4}{28} = \frac{2}{14}, \\
p_2 &= \frac{6/14}{4/14 + 6/14 + 18/14} = \frac{6}{4 + 6 + 18} = \frac{6}{28} = \frac{3}{14}, \\
p_3 &= \frac{18/14}{4/14 + 6/14 + 18/14} = \frac{18}{4 + 6 + 18} = \frac{18}{28} = \frac{9}{14},
\end{aligned}$$

which are summarized into

X	x_1	x_2	x_3
$p_i = \Pr(X = x_i)$	$2/14$	$3/14$	$9/14$

Similarly, letting $x_0 = x_3$ in (1.60) yields the following Y-marginal

Y	y_1	y_2	y_3
$q_j = \Pr(Y = y_j)$	$3/14$	$4/14$	$7/14$

2) The joint distribution of (X, Y) is given by

$$\boldsymbol{P} = \begin{pmatrix} 1/28 & 0 & 3/28 \\ 0 & 2/28 & 4/28 \\ 5/28 & 6/28 & 7/28 \end{pmatrix}. \qquad \|$$

1.11 Categorical Distribution

45• Finite discrete distribution

- Let the discrete r.v. X be defined as follows:

X	x_1	$\cdots$	x_i	$\cdots$	x_d
$\Pr(X = x_i)$	p_1	$\cdots$	p_i	$\cdots$	p_d

where the probabilities $p_i > 0$ and $\sum_{i=1}^d p_i = 1$.

- When d is finite, we say X follows a finite discrete distribution, denoted by $X \sim \text{FDiscrete}_d(\boldsymbol{x}, \boldsymbol{p})$, where $\boldsymbol{x} = (x_1, \dots, x_d)^\top$ and $\boldsymbol{p} = (p_1, \dots, p_d)^\top \in \mathbb{T}_d \hat{=} \{(p_1, \dots, p_d)^\top \colon p_i > 0, \ \sum_{i=1}^d p_i = 1\}$.

45.1• Basic features of $X \sim \textbf{FDiscrete}_d(\boldsymbol{x}, \boldsymbol{p})$

— X is a r.v., not a random vector with support being $\{x_1, \dots, x_d\}$.

— $\{x_i\}_{i=1}^d$ are *numeric* or real numbers.

— $E(X) = \sum_{i=1}^d x_i p_i$ is meaningful.

— Bernoulli, binomial, and hypergeometric distributions are special cases.

46• Categorical random variable

- Define a categorical random variable Y with d-category as follows:

Y	A_1	$\cdots$	A_i	$\cdots$	A_d
$\Pr(Y = A_i)$	p_1	$\cdots$	p_i	$\cdots$	p_d

where $\{A_i\}_{i=1}^d$ denote *characters, labels* or *symbols.*

46.1• Some examples

— Blood types can be classified as A, B, AB, and O.

— USA citizen can be classified as White, African-American, Asian or Pacific Islander, and Native America.

— The color can be classified as red, green, blue, and others.

46.2• A key feature

— Obviously, $E(Y)$ is *meaningless.*

— Therefore, we cannot define above distribution of Y as a categorial distribution.

47• Categorical distribution

- However, we can define a *one-to-one mapping* between a random vector $\mathbf{y} = (Y_1, \ldots, Y_d)^\top$ and the above categorical random variable Y.
- For the purpose of illustration, let $d = 4$, we define

$$\begin{aligned}
\mathbf{y} &= (1,0,0,0)^\top = \boldsymbol{e}_4^{(1)} \leftrightarrow Y = A_1 \text{ (if the blood type is A)},\\
\mathbf{y} &= (0,1,0,0)^\top = \boldsymbol{e}_4^{(2)} \leftrightarrow Y = A_2 \text{ (if the blood type is B)},\\
\mathbf{y} &= (0,0,1,0)^\top = \boldsymbol{e}_4^{(3)} \leftrightarrow Y = A_3 \text{ (if the blood type is AB)},\\
\mathbf{y} &= (0,0,0,1)^\top = \boldsymbol{e}_4^{(4)} \leftrightarrow Y = A_4 \text{ (if the blood type is O)}.
\end{aligned}$$

- Therefore, we can define the categorical distribution as follows:

$\mathbf{y}$	$\boldsymbol{e}_d^{(1)}$	$\cdots$	$\boldsymbol{e}_d^{(i)}$	$\cdots$	$\boldsymbol{e}_d^{(d)}$
$\Pr\{\mathbf{y} = \boldsymbol{e}_d^{(i)}\}$	p_1	$\cdots$	p_i	$\cdots$	p_d

and denote it by $\mathbf{y} = (Y_1, \ldots, Y_d)^\top \sim \text{Categorical}_d(\boldsymbol{p})$, which is a special case of the multinomial distribution with the first parameter being 1, i.e., $\mathbf{y} \sim \text{Multinomial}_d(1; \boldsymbol{p})$.

- The pmf of $\mathbf{y}$ is given by

$$\Pr(\mathbf{y} = \boldsymbol{y}) = \binom{1}{y_1, \ldots, y_d} \prod_{i=1}^{d} p_i^{y_i} = \prod_{i=1}^{d} p_i^{y_i}, \tag{1.62}$$

where $\boldsymbol{y} = (y_1, \ldots, y_d)^\top$, $y_i = 0$ or 1 and $\sum_{i=1}^d y_i = 1$. In other words, only one component of $\boldsymbol{y}$ is 1 and others are zeros.

47.1• Basic features of $\mathbf{y} \sim \text{Categorical}_d(\boldsymbol{p})$

— $\mathbf{y}$ is a random vector with support being $\{\boldsymbol{e}_d^{(1)}, \ldots, \boldsymbol{e}_d^{(d)}\}$.

— $\{\boldsymbol{e}_d^{(1)}, \ldots, \boldsymbol{e}_d^{(d)}\}$ are unit/base vectors.

— $E(\mathbf{y}) = \boldsymbol{p}$ is meaningful.

— It reduces to Bernoulli distribution when $d = 2$.

— There is a one-to-one mapping between $\mathbf{y} \sim \text{Categorical}_d(\boldsymbol{p})$ and the categorical random variable Y with $\Pr(Y = A_i) = p_i$ for $i = 1, \ldots, d$.

— If $\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{\text{iid}}{\sim} \text{Categorical}_d(\boldsymbol{p})$, then

$$\sum_{j=1}^{n} \mathbf{y}_j \sim \text{Multinomial}_d(n; \boldsymbol{p}). \tag{1.63}$$

1.12 Zero-inflated Poisson Distribution

48• Definition

- Let $Z \sim \text{Bernoulli}(1 - \phi)$, $X \sim \text{Poisson}(\lambda)$ and $Z \perp\!\!\!\perp X$.
- Let $Y \overset{\text{d}}{=} ZX$, we say Y follows the *zero-inflated Poisson* (ZIP) distribution, denoted by $Y \sim \text{ZIP}(\phi, \lambda)$.

48.1• The pmf of $Y \sim \text{ZIP}(\phi, \lambda)$

— The r.v. $Y \sim \text{ZIP}(\phi, \lambda)$ has the following stochastic representation:

$$Y \overset{\text{d}}{=} ZX = \begin{cases} 0, & \text{with probability } \phi, \\ X, & \text{with probability } 1 - \phi \end{cases} \tag{1.64}$$

with support $\mathcal{S}_Y = \{0, 1, 2, \ldots, \infty\}$.

— Since $\{Y = 0\} \Leftrightarrow \{Z = 0\}$ or $\{Z = 1, X = 0\}$, and for $y > 0$, $\{Y = y\} \Leftrightarrow \{Z = 1, X = y\}$, we obtain

$$\begin{aligned}
\Pr(Y = 0) &= \Pr(Z = 0) + \Pr(Z = 1, X = 0) = \phi + (1 - \phi)\,\mathrm{e}^{-\lambda}, \\
\Pr(Y = y) &= \Pr(Z = 1, X = y) = (1 - \phi)\,\Pr(X = y) \\
&= (1 - \phi)\frac{\lambda^y\,\mathrm{e}^{-\lambda}}{y!}, \qquad y > 0.
\end{aligned}$$

— The pmf of Y is given by

$$\begin{aligned}
& f(y|\phi, \lambda) = \Pr(Y = y) \\
&= \begin{cases} \phi + (1 - \phi)\,\mathrm{e}^{-\lambda}, & \text{if } \ y = 0, \\ (1 - \phi)\dfrac{\lambda^y\,\mathrm{e}^{-\lambda}}{y!}, & \text{if } \ y \geqslant 1 \end{cases} \\
&= \{\phi + (1 - \phi)\,\mathrm{e}^{-\lambda}\} I(y = 0) + \left\{(1 - \phi)\frac{\lambda^y\,\mathrm{e}^{-\lambda}}{y!}\right\} I(y > 0) \qquad (1.65) \\
&= \phi \cdot I(y = 0) + (1 - \phi)\,\Pr(X = y),
\end{aligned}$$

where the $\phi \in [0, 1)$ denotes the unknown proportion for incorporating more extra-zeros than those permitted by the ordinary Poisson distribution.

48.2• Comments

— The ZIP distribution may be viewed as a *mixture* of a degenerate distribution with all mass at zero (denoted by $\xi \sim \text{Degenerate}(0)$) and a Poisson($\lambda$) distribution.

— In particular, when $\phi = 0$, the ZIP distribution is reduced to the ordinary Poisson distribution.

— The ZIP distribution can be used to model count data with *extra zeros*.

48.3• Examples of count data with extra zeros

— The number of insurance claims within a population for a certain type of risk is very likely zero-inflated by those people who have not taken insurance against the risk and thus are unable to claim.

— The number of workdays missed due to sickness of a dependent in a four-week period.

— The number of papers published per year by a researcher.

49• Basic properties

49.1• Expectation and variance of Y

— From (1.64), we immediately obtain

$$\begin{aligned} E(Y) &= E(ZX) &= E(Z)E(X) &= (1-\phi)\lambda, \\ E(Y^2) &= E(Z^2X^2) &= E(Z)E(X^2) &= (1-\phi)(\lambda^2+\lambda), \end{aligned}$$

and $\mathrm{Var}(Y) = E(Y^2) - \{E(Y)\}^2 = (1-\phi)\lambda(1+\phi\lambda)$.

49.2• The mgf of Y

— We have

$$\begin{aligned} M_Y(t) &= E(\mathrm{e}^{tY}) = E(\mathrm{e}^{tZX}) \\ &= \phi + (1-\phi)E(\mathrm{e}^{tX}) \\ &= \phi + (1-\phi)\exp\{\lambda(\mathrm{e}^t - 1)\}. \end{aligned}$$

49.3• Show that $Y|(Z=z) \sim$ Poisson(λz)

— It is obvious. □

49.4• The conditional distribution of $Z|(Y=y)$

— The r.v. Z is a Bernoulli variable, only taking the value 0 or 1.

— Note that

$$\begin{aligned} \Pr(Z=1|Y=y) &= \frac{\Pr(Z=1, Y=y)}{\Pr(Y=y)} = \frac{\Pr(Z=1, X=y)}{f(y|\phi,\lambda)} \\ &= \frac{(1-\phi)\,\mathrm{e}^{-\lambda}\lambda^y/y!}{f(y|\phi,\lambda)} \hat{=} p_y, \end{aligned}$$

where $f(y|\phi,\lambda)$ is given by (1.65), then we have

$$p_0 = \frac{(1-\phi)\,\mathrm{e}^{-\lambda}}{\phi + (1-\phi)\,\mathrm{e}^{-\lambda}} \quad \text{and} \quad p_y = 1 \ \text{for} \ y > 0. \tag{1.66}$$

— Therefore,

$$Z|(Y=y) \sim \begin{cases} \text{Bernoulli}\,(p_0), & \text{if } y=0, \\ \text{Degenerate}(1), & \text{if } y>0. \end{cases} \tag{1.67}$$

49.5• The conditional distribution of $X|(Y=y)$

— We first find the conditional distribution of $X|(Y=y=0)$.

— Note that

$$\begin{aligned} \Pr(X=x|Y=0) &= \frac{\Pr(X=x, Y=0)}{\Pr(Y=0)} \\ &= \frac{\Pr(X=0, Y=0)}{f(0|\phi,\lambda)} I(x=0) + \frac{\Pr(X=x, Z=0)}{f(0|\phi,\lambda)} I(x>0) \\ &= \frac{\Pr(X=0)}{f(0|\phi,\lambda)} I(x=0) + \frac{\phi \Pr(X=x)}{f(0|\phi,\lambda)} I(x>0) \\ &\quad [\because \{X=0\} \subseteq \{Y=0\}] \\ &= \frac{\mathrm{e}^{-\lambda} I(x=0)}{\phi+(1-\phi)\,\mathrm{e}^{-\lambda}} + \frac{\phi}{\phi+(1-\phi)\,\mathrm{e}^{-\lambda}} \cdot \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} I(x>0) \\ &\overset{(1.66)}{=} \{p_0 + (1-p_0)\,\mathrm{e}^{-\lambda}\} I(x=0) \\ &\quad + \left\{(1-p_0)\frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}\right\} I(x>0). \end{aligned} \tag{1.68}$$

— By comparing (1.68) with (1.65), we have

$$X|(Y=0) \sim \text{ZIP}(p_0, \lambda). \tag{1.69}$$

— We then find the conditional distribution of $X|(Y=y>0)$.

— Note that

$$\begin{aligned} &\Pr(X=x|Y=y) \\ &= \frac{\Pr(X=x, Y=y)}{\Pr(Y=y)} \qquad [\because y>0 \Rightarrow x=y>0 \ \&\ Z=1] \\ &= \frac{\Pr(X=y, Z=1)}{f(y|\phi,\lambda)} \overset{(1.65)}{=} \frac{(1-\phi)\Pr(X=y)}{(1-\phi)\,\mathrm{e}^{-\lambda}\lambda^y/y!} = 1, \end{aligned}$$

implying that $X|(Y=y>0) \sim \text{Degenerate}(y)$.

Exercise 1

1.1 Let $X \sim \text{Binomial}(n, p)$, $Y \sim \text{Poisson}(\lambda)$, and $X \perp\!\!\!\perp Y$.

(a) Find the mgf of X.

(b) Use the formula (1.34) to find the variance of X.

(c) Find the distribution of $X + Y$.

1.2 The joint pmf of X and Y is given by

(X, Y)	$(1,1)$	$(1,2)$	$(1,3)$	$(1,4)$	$(2,2)$
Probability	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{2}{16}$
(X, Y)	$(2,3)$	$(2,4)$	$(3,3)$	$(3,4)$	$(4,4)$
Probability	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{4}{16}$

(a) Find the marginal distribution of X.

(b) Find the pmf of $X + Y$.

1.3 Let two conditional distributions be exponential restricted to the interval $[0, b)$; that is,

$$\begin{aligned} f_{(X|Y)}(x|y) &= \frac{y\exp(-yx)}{1-\exp(-by)}, \quad 0 \leqslant x < b < +\infty, \\ f_{(Y|X)}(y|x) &= \frac{x\exp(-xy)}{1-\exp(-bx)}, \quad 0 \leqslant y < b < +\infty. \end{aligned}$$

(a) Find the marginal distribution of X.

(b) If $b = +\infty$, please discuss the existence of $f_X(x)$.

1.4 Let X be a discrete r.v. with pmf $p_i = \Pr(X = x_i)$ for $i = 1, 2, 3$ and Y be another discrete r.v. with pmf $q_j = \Pr(Y = y_j)$ for $j = 1, 2, 3, 4$. Given two conditional distribution matrices

$$\boldsymbol{A} = \begin{pmatrix} 1/7 & 1/4 & 3/7 & 1/7 \\ 2/7 & 1/2 & 1/7 & 2/7 \\ 4/7 & 1/4 & 3/7 & 4/7 \end{pmatrix}$$

and

$$\boldsymbol{B} = \begin{pmatrix} 1/6 & 1/6 & 1/2 & 1/6 \\ 2/7 & 2/7 & 1/7 & 2/7 \\ 1/3 & 1/12 & 1/4 & 1/3 \end{pmatrix},$$

where the (i, j) element of $\boldsymbol{A}$ is $a_{ij} = \Pr(X = x_i|Y = y_j)$ and the (i, j) element of $\boldsymbol{B}$ is $b_{ij} = \Pr(Y = y_j|X = x_i)$.

(a) Find the marginal distributions of X and Y.

(b) Find the joint distribution of (X, Y).

1.5 Let X be a continuous r.v. with pdf $f(x)$. If m is the unique median of the distribution of X and b is a real constant.

(a) Show that

$$E(|X - b|) = E(|X - m|) + 2\int_m^b (b - x)f(x)\,\mathrm{d}x,$$

provided that the expectation exists.

(b) Find the value of b such that $E(|X - b|)$ is minimized.

1.6 Let X be a r.v. having the following cdf

$$F(x) = \begin{cases} 0, & x < 0, \\ 2x^2, & 0 \leqslant x < 1/2, \\ 1 - 2(1 - x)^2, & 1/2 \leqslant x < 1, \\ 1, & 1 \leqslant x. \end{cases}$$

(a) Calculate $\Pr(1/4 < X < 5/8)$.

(b) Find the variance of X.

1.7 Let $\mathbf{x} = (X_1, \ldots, X_d)^\top$ be a random vector, the joint mgf of $\mathbf{x}$ is defined by $M_{\mathbf{x}}(\boldsymbol{t}) = E\{\exp(t_1X_1 + \cdots + t_dX_d)\}$, where $\boldsymbol{t} = (t_1, \ldots, t_d)^\top$ be a real vector. For either the discrete case or the continuous case, prove that

(a) For a fixed i $(i = 1, \ldots, d)$, the partial derivative of the joint mgf with respect to t_i evaluated at $t_1 = \cdots = t_d = 0$ is $E(X_i)$.

(b) For two fixed i, j $(i \neq j,\ i, j = 1, \ldots, d)$, the second derivative of the joint mgf with respect to t_i and t_j evaluated at $t_1 = \cdots = t_d = 0$ is $E(X_iX_j)$.

(c) If two r.v.'s have the following joint density

$$f(x, y) = \begin{cases} \exp(-x - y), & \text{for } x \geqslant 0,\ y \geqslant 0, \\ 0, & \text{elsewhere,} \end{cases}$$

find the joint mgf and use it to derive $E(XY)$, $E(X)$, $E(Y)$ and $\mathrm{Cov}(X,Y)$.

1.8 A discrete r.v. X is said to follow a *zero-truncated Poisson* (ZTP) distribution if its pmf is

$$\Pr(X=x) = c \cdot \frac{\lambda^x \, \mathrm{e}^{-\lambda}}{x!}, \quad x = 1, 2, \ldots,$$

where $\lambda > 0$ is an unknown parameter, and c is the normalizing constant such that $\sum_{x=1}^{\infty} \Pr(X=x) = 1$. We will write $X \sim \mathrm{ZTP}(\lambda)$.

A. Let $X \sim \mathrm{ZTP}(\lambda)$.

(a) Find the normalizing constant c.

(b) Find $E(X)$, $E(X^2)$ and $\mathrm{Var}(X)$.

(c) Find the moment generating function of X.

B. Let $X_1 \sim \mathrm{ZTP}(\lambda_1)$, $X_2 \sim \mathrm{ZTP}(\lambda_2)$, and they are independent.

(d) Find the distribution of $X_1 + X_2$.

(e) Find the conditional distribution of $X_1|(X_1+X_2 = x)$, where $x \geqslant 2$ and x is an integer.

1.9 Let $U \sim \mathrm{Poisson}(\lambda_0)$, $V \sim \mathrm{Poisson}(\lambda)$, $W \sim \mathrm{Poisson}(\beta\lambda)$, $Z \sim \mathrm{Bernoulli}(1-\phi)$, and U, V, W, Z are mutually independent. Define

$$X = U + V \quad \text{and} \quad Y = Z(U+W).$$

(a) Find $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$ and $\mathrm{Cov}(X,Y)$.

(b) Find the joint distribution of X and Y.

1.10 Suppose that the mgfs of two independent random variables X and Y are given by

$$M_X(t) = \exp(t^2/2) \quad \text{and} \quad M_Y(t) = \exp(2t^2 - t),$$

respectively. Define $W = 3X + 2Y$.

(a) Calculate the probability $\Pr(-12 < W < 3)$.

(b) Calculate $E(W^2)$.

1.11 Suppose that the random variable $X \sim N(0,1)$ and define $Y = |X|$.

(a) Calculate the expectation and variance of Y.

(b) Find the cdf and pdf of Y.

1.12 Let X be a random variable with cdf $F(x)$. Note that $F(\cdot)$ is a non-decreasing function, the inverse function $F^{-1}(\cdot)$ can be defined by

$$F^{-1}(u) = \text{infimum}\{x\colon\ F(x) \geqslant u\}, \quad u \in (0,1).$$

Let $U \sim U(0,1)$, show that $F(X) \overset{\mathrm{d}}{=} U$ or equivalently $X \overset{\mathrm{d}}{=} F^{-1}(U)$.