UNIVERSITY OF LONDON

Advanced statistics: distribution theory

J.S. Abdey

ST2133

2021

Undergraduate study in
Economics, Management, Finance and the Social Sciences

This subject guide is for a 200 course offered as part of the University of London undergraduate study in Economics, Management, Finance and the Social Sciences. This is equivalent to Level 5 within the Framework for Higher Education Qualifications in England, Wales and Northern Ireland (FHEQ). For more information, see: www.london.ac.uk

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

This guide was prepared for the University of London by:
James S. Abdey, BA (Hons), MSc, PGCertHE, PhD, Department of Statistics, London School of Economics and Political Science.
This is one of a series of subject guides published by the University. We regret that due to pressure of work the author is unable to enter into any correspondence relating to, or arising from, the guide.

University of London
Publications Office
Stewart House
32 Russell Square
London WC1B 5DN
United Kingdom
www.london.ac.uk

Published by: University of London

Contents

Chapter 1
Introduction

1.1 Route map to the subject guide

This subject guide provides you with a framework for covering the syllabus of the **ST2133 Advanced statistics: distribution theory** half course.

Uncertainty is part of life. If the outcomes of our actions were never in any doubt, life would be incredibly dull! Each of us possesses an intuitive sense that some events are more likely to occur than others. For example, if I were to knock over a cup of coffee, I can be pretty sure the contents will spill out; if I were to buy a lottery ticket I might hope to win the jackpot and become a millionaire by matching the numbers drawn, however there are many other (more likely) outcomes.

Science requires us to go beyond making purely vague statements such as 'I can be pretty sure' and 'there are more likely outcomes'. We need to *quantify* uncertainty by assigning numbers to possible events. Probability is the mechanism by which we quantify uncertainty, and *distribution theory* provides the tools which allow us to build probabilistic models of real-world phenomena.

Appreciate that distribution theory provides a foundation on which questions of statistical inference rely. Consider the following examples.

- In the UK, 'stop and search' is the practice by police of stopping people in the street if there are 'reasonable grounds' to suspect an individual is carrying illegal drugs, weapons, stolen property, or items which could be used to commit a crime. Each month the police collect information (*statistics*) on the number of criminal offences reported, as well as the number of people stopped and searched. A natural hypothesis is that the more people who are stopped and searched, crime levels decrease. To test this hypothesis we would need to decide which is an appropriate probability distribution to *model* the amount of crime. Also, we would wish to investigate the nature of any *association* between the numbers of crimes and searches, and whether there is any trend in crime over time.
- In the business world, probability distributions can be extremely useful for modelling future returns and profitability of a business. For example, following a marketing campaign, senior management may wish to model the change in sales. Also, distribution theory is helpful for quantifying *risk* – see **ST2187 Business analytics, applied modelling and prediction** where Monte Carlo simulation models require probability distributions for a system's random input variables.

Being able to quantify uncertainty is advantageous in many industries, in particular investment banking, insurance and market research (by no means an exhaustive list). There is also great demand for skilled statisticians in the public sector, such as

government statistical services, government departments and law enforcement agencies. Finally, distribution theory is essential to scientific modelling – medical researchers, climate researchers, social scientists, psychologists and biologists (among others) all rely heavily on statistical inference methods which draw heavily on distribution theory.

Distribution theory is a subject which rewards practice and repetition. Understanding it well is the key to understanding the whole of statistics!

1.2 Aims and learning outcomes

The principal aim of this course is to provide a thorough theoretical grounding in probability distributions. The half course teaches fundamental material which is required for specialised courses in statistics, actuarial science and econometrics.

At the end of this half course and having completed the recommended reading and activities you should be able to:

- recall a large number of probability distributions and be a competent user of their mass/density and distribution functions and moment generating functions
- explain relationships between variables, conditioning, independence and correlation
- relate the theory and method taught in the half course to solve practical problems.

1.3 How to use this subject guide

The subject guide is intended to help you interpret the syllabus. Detailed exposition is provided for each of the areas along with some suggestions for reading. Precisely how you use the subject guide is clearly a matter of personal preference. A recommended approach is as follows.

- Read each chapter in full reasonably quickly. Do not worry too much if some of it does not make sense initially.
- Go through each chapter section by section by reading, making notes, completing activities and checking that you understand every aspect of the section.
- Go through the learning outcomes and make sure that you are comfortable with all of the points.
- Complete the end-of-chapter sample examination questions, referring to the subject guide as necessary.

The topics in **ST2133 Advanced statistics: distribution theory** build on each other incrementally and sequentially. You should be aware that, although sections are fairly self-contained, examination questions may contain material from many different parts of the subject guide. Finally, remember that you will be assessed on your ability to answer questions. Practice is the *only* way to make sure you can do this successfully.

1.4 Syllabus

The syllabus of **ST2133 Advanced statistics: distribution theory** is as follows.

- **Probability:** Probability measure. Conditional probability. Bayes' theorem.
- **Distribution theory:** Distribution function. Mass and density. Expectation operator. Moments, moment generating functions, cumulant generating functions. Convergence concepts.
- **Multivariate distributions:** Joint distributions. Conditional distributions, conditional moments. Functions of random variables.

The subject guide is comprised of the following chapters.

- Chapter 1 is this introduction.
- Chapter 2 introduces probability fundamentals which are essential for the remaining chapters.
- Chapter 3 presents the key ideas of distribution theory, namely random variables, distributions and expectation.
- Chapter 4 extends Chapter 3 to the case where there are many variables of interest, with a particular emphasis on bivariate cases.
- Chapter 5 deals with conditional distributions and introduces some important classes of practical problems.

Basic notation

We often use the symbol ■ to denote the end of a proof, where we have finished explaining why a particular result is true. This is just to make it clear where the proof ends and the following text begins.

Calculators

A calculator may be used when answering questions on the examination paper for **ST2133 Advanced statistics: distribution theory**. It must comply in all respects with the specification given in the Regulations. You should also refer to the admission notice you will receive when entering the examination and the 'Notice on permitted materials'.

1.4.1 Recommended reading

In principle, this subject guide acts as your essential reading and is sufficient to perform well in the final examination – provided you practise many problems! However, textbooks can be beneficial since they provide:

- an alternative (and in some instances more detailed) perspective on material in the subject guide
- additional illustrative examples
- more exercises to practise.

In looking at textbooks, you will encounter much which will be familiar from looking at the subject guide. However, you will also come across topics which are not covered and questions requiring a deeper level of understanding than is provided here. Remember, this subject guide is the *definitive* syllabus and that the level of examination questions will be similar to the level of questions given at the end of each chapter.

If you do wish to consult a textbook for this half course, we recommend:

- Casella, G. and R.L. Berger, *Statistical Inference*. (Duxbury, 2008) second edition [ISBN 9788131503942].

Indicative reading in this subject guide refers to the above edition of the textbook. Newer editions may have been published by the time you study this half course. You can use a more recent edition by using the detailed chapter and section headings and the index to identify relevant readings. Also, check the virtual learning environment (VLE) regularly for updated guidance on readings.

1.4.2 Further reading

Please note that you are free to read around the subject area in any text, paper or online resource. You can support your learning by reading as widely as possible and by thinking about how these principles apply in the real world. To help you read extensively, you have free access to the virtual learning environment (VLE) and University of London Online Library (see below).

Other useful textbooks for this course include:

- Bartoszynski, R. and M. Niewiadomska–Bugaj, *Probability and Statistical Inference*. (Wiley, 1996) [ISBN 9780471310730].
- Grimmett, G. and D. Stirzaker, *Probability and Random Processes*. (Oxford University Press, 2001) third edition [ISBN 9780198572220].
- Hogg, R.V. and E.A. Tanis, *Probability and Statistical Inference*. (Pearson, 2014) ninth edition [ISBN 9781292062358].
- Larsen R.J. and M.L. Marx, *Introduction to Mathematical Statistics and Its Applications*. (Pearson, 2013) fifth edition [ISBN 9781292023557].
- Miller, I. and M. Miller, *John E. Freund's Mathematical Statistics with Applications*. (Pearson, 2012) eighth edition [ISBN 9780321904409].

1.4.3 Online study resources (the Online Library and the VLE)

In addition to the subject guide and the recommended reading, it is crucial that you take advantage of the study resources which are available online for this course, including the virtual learning environment (VLE) and the Online Library.

You can access the VLE, the Online Library and your University of London email account via the Student Portal at:

`http://my.london.ac.uk`

You should have received your login details for the Student Portal with your official offer, which was emailed to the address that you gave on your application form. You have probably already logged in to the Student Portal in order to register! As soon as you registered, you will automatically have been granted access to the VLE, Online Library and your fully functional University of London email account.

If you forget your login details, please click on the 'Forgotten your password' link on the login page.

The VLE

The VLE, which complements this subject guide, has been designed to enhance your learning experience, providing additional support and a sense of community. It forms an important part of your study experience with the University of London and you should access it regularly.

The VLE provides a range of resources for EMFSS courses:

- **Course materials:** Subject guides and other course materials available for download. In some courses, the content of the subject guide is transferred into the VLE and additional resources and activities are integrated with the text.
- **Readings:** Direct links, wherever possible, to essential readings in the Online Library, including journal articles and ebooks.
- **Video content:** Including introductions to courses and topics within courses, interviews, lessons and debates.
- **Screencasts:** Videos of PowerPoint presentations, animated podcasts and on-screen worked examples.
- **External material:** Links out to carefully selected third-party resources.
- **Self-test activities:** Multiple-choice, numerical and algebraic quizzes to check your understanding.
- **Collaborative activities:** Work with fellow students to build a body of knowledge.
- **Discussion forums:** A space where you can share your thoughts and questions with fellow students. Many forums will be supported by a 'course moderator', a subject expert employed by LSE to facilitate the discussion and clarify difficult topics.

- **Past examination papers:** We provide up to three years of past examinations alongside *Examiners' commentaries* that provide guidance on how to approach the questions.

- **Study skills:** Expert advice on getting started with your studies, preparing for examinations and developing your digital literacy skills.

Some of these resources are available for certain courses only, but we are expanding our provision all the time and you should check the VLE regularly for updates.

Note: Students registered for Laws courses also receive access to the dedicated Laws VLE.

Making use of the Online Library

The Online Library (`http://onlinelibrary.london.ac.uk`) contains a huge array of journal articles and other resources to help you read widely and extensively.

To access the majority of resources via the Online Library you will either need to use your University of London Student Portal login details, or you will be required to register and use an Athens login.

The easiest way to locate relevant content and journal articles in the Online Library is to use the **Summon** search engine.

If you are having trouble finding an article listed in a reading list, try removing any punctuation from the title, such as single quotation marks, question marks and colons.

For further advice, please use the online help pages (`https://onlinelibrary.london.ac.uk/resources/summon`) or contact the Online Library team: onlinelibrary@shl.london.ac.uk

Additional material

There is a lot of computer-based teaching material available freely over the web. A fairly comprehensive list can be found in the 'Books & Manuals' section of:

`http://statpages.info`

Unless otherwise stated, all websites in this subject guide were accessed in August 2020. We cannot guarantee, however, that they will stay current and you may need to perform an internet search to find the relevant pages.

1.5 Examination advice

Important: the information and advice given here are based on the examination structure used at the time this subject guide was written. Please note that subject guides may be used for several years. Because of this we strongly advise you to always check both the current Programme regulations for relevant information about the examination, and the VLE where you should be advised of any forthcoming changes.

You should also carefully check the rubric/instructions on the paper you actually sit and follow those instructions.

The examination paper for this half course is two hours long. There will be *four* questions. You should answer all four questions. The examiners attempt to ensure that the examination provides broad coverage of the syllabus. Some questions may cover more than one topic. A Sample examination paper appears in Appendix D, along with solutions in Appendix E.

There are three reasons why there is no choice of questions in the examination paper:

i. to make sure you study all parts of the course thoroughly

ii. to avoid the risk that candidates make bad choices when selecting questions

iii. to assess you on your ability to use distribution theory, not your ability to choose examination questions.

The format of the paper is as follows.

- **Section A** (40%): One question divided into several parts. A useful mark allocation for each part will be provided. This question will broadly cover the material in the syllabus.
- **Section B** (60%): Three questions each worth (approximately) 20%. Each question will attempt to test the depth of your knowledge of the material in the syllabus.

You should divide your time in proportion to the marks available for each subquestion. Roughly, this works out as follows: five minutes to read the paper, forty-five minutes on Section A, twenty minutes on each question in Section B, leaving ten minutes for checking.

Remember, it is important to check the VLE for:

- up-to-date information on examination and assessment arrangements for this course
- where available, past examination papers and *Examiners' commentaries* for the course which give advice on how each question might best be answered.

1. Introduction

Chapter 2
Probability space

2.1 Recommended reading

Casella, G. and R.L. Berger, *Statistical Inference*. Chapter 1, Sections 1.1–1.3.

2.2 Learning outcomes

On completion of this chapter, you should be able to:

- calculate probabilities for simple situations by counting
- provide the definition of a σ-algebra
- derive the properties of probability measure from axioms
- calculate probabilities for problems with large spaces using combinations and permutations
- explain the association between the number of combinations and multinomial coefficients
- derive and demonstrate conditional probability
- prove conditional probability defines a valid probability measure
- apply Bayes' theorem to solve problems involving reverse conditioning
- define independent events and exploit independence to calculate probabilities.

2.3 Introduction

This chapter extends the discussion of probability theory introduced in **ST104b Statistics 2**, so it is advisable to review that material first before proceeding.

2.4 Probability fundamentals

Recall that probabilities are real numbers in the interval $[0, 1]$, quantifying how likely an event, such as A, is to occur. If $P(A) = 0$ then A is an impossible event, while if

$P(A) = 1$ then A is a certain event, and for $0 < P(A) < 1$ the event A may, or may not, occur with A being increasingly likely as $P(A) \to 1$.

Definitions

1. An **experiment** is a repeatable procedure which has a well-defined set of possible outcomes.
2. The **sample space**, denoted by Ω, (previously denoted by S in **ST104b Statistics 2**), is the set of all possible outcomes of an experiment. Therefore, any sample outcome ω is a member of the sample space, i.e. $\omega \in \Omega$.
3. An **event** A is a set of outcomes which is of interest to us. An event is a subset of the sample space, i.e. $A \subseteq \Omega$, or $A \subset \Omega$ if A is a strict subset of Ω (that is, there are elementary outcomes in Ω which are not in A).

When all of the outcomes in the sample space are *equally likely* and the sample space is finite, an intuitive definition of the probability of the event A is:

$$P(A) = \frac{n(A)}{n(\Omega)}$$

where $n(A)$ is the number of elementary outcomes which are in A, and $n(\Omega)$ is the total number of (equally likely) possible outcomes. The statement of the sample space being finite means that there is a finite number of possible outcomes of the experiment, i.e. $n(\Omega) < \infty$. This concept of 'classical probability' was first introduced in **ST104b Statistics 2**.

At this point, it is worth remembering that probability is a *mathematical construct*. Whenever we apply the ideas of probability to real-world situations we always make *assumptions*. Therefore, probability statements are statements about a mathematical model, *not* statements about reality.

Example 2.1 Consider rolling two fair dice. The sample space is the set of all possible outcomes of the experiment, given by:

$$\begin{aligned}\Omega = \{&(1,1), \quad (1,2), \quad (1,3), \quad (1,4), \quad (1,5), \quad (1,6),\\ &(2,1), \quad (2,2), \quad (2,3), \quad (2,4), \quad (2,5), \quad (2,6),\\ &(3,1), \quad (3,2), \quad (3,3), \quad (3,4), \quad (3,5), \quad (3,6),\\ &(4,1), \quad (4,2), \quad (4,3), \quad (4,4), \quad (4,5), \quad (4,6),\\ &(5,1), \quad (5,2), \quad (5,3), \quad (5,4), \quad (5,5), \quad (5,6),\\ &(6,1), \quad (6,2), \quad (6,3), \quad (6,4), \quad (6,5), \quad (6,6)\}.\end{aligned}$$

Here the elementary outcomes are of the form (value on first die, value on second die). Suppose our interest is in the event that the sum of the two values is greater than 10, hence $A = \{(5,6), (6,5), (6,6)\}$, alternatively $A = \{(5,6) \cup (6,5) \cup (6,6)\}$ when expressed as the *union* of the elementary outcomes. Immediately, it is seen that $A \subset \Omega$. Assuming *fair* dice then all members of the sample space are equally likely. Hence by counting outcomes we have that $n(A) = 3$ and $n(\Omega) = 36$, and so $P(A) = 3/36 = 1/12$.

Note that the probability calculated in Example 2.1 is a statement about a *model* for rolling two dice in which each of the possible outcomes is equally likely. For this model to be valid, the dice must be fair such that the six outcomes from rolling each die are equally likely. This is a reasonable assumption, although this may not hold in practice – imperfections in the dice may make some values slightly more likely than others. However, for the sake of simplicity, we may choose to ignore any slight differences (which are likely to be negligible differences in practice anyway).

Example 2.2 Suppose that a fair coin is tossed repeatedly until both a head and a tail have appeared at least once.

(a) Describe the sample space.

(b) What is the probability that three tosses will be required?

Solution

(a) We have $\Omega = \Omega_H \cup \Omega_T$, where:

$$\Omega_H = \{i \text{ tails followed by 1 head} : i = 1, 2, \ldots\}$$

and:

$$\Omega_T = \{i \text{ heads followed by 1 tail} : i = 1, 2, \ldots\}.$$

(b) Using independence, the required probability is:

$$P(TTH) + P(HHT) = 2 \times \left(\frac{1}{2}\right)^3 = \frac{1}{4}.$$

Example 2.3 The probability that a child has blue eyes is 0.25. Assume independence between children. Consider a family with three children.

(a) If it is known that at least one child has blue eyes, what is the probability that at least two children have blue eyes?

(b) If it is known that the youngest child has blue eyes, what is the probability that at least two children have blue eyes?

Solution

Let $A_i = 1$ denote the event that the ith child in the family has blue eyes, and $A_i = 0$ denote the event that the ith child has not got blue eyes. Hence we have $P(A_i = 1) = 0.25$ and $P(A_i = 0) = 0.75$, for $i = 1, 2, 3$. The sample space is:

$$\Omega = \{\{A_1, A_2, A_3\} : A_i = 1 \text{ or } 0\}.$$

(a) We have:

$$P(\{A_1, A_2, A_3\} : \text{at least two of } A_i \text{ are } 1 \,|\, \{A_1, A_2, A_3\} : \text{at least one of } A_i \text{ is } 1)$$

$$
\begin{aligned}
&= \frac{P(\{A_1, A_2, A_3\} : \text{at least two of } A_i \text{ are } 1)}{P(\{A_1, A_2, A_3\} : \text{at least one of } A_i \text{ is } 1)} \\
&= \frac{P(\{1,1,1\} \cup \{1,1,0\} \cup \{1,0,1\} \cup \{0,1,1\})}{1 - P(\{0,0,0\})} \\
&= \frac{(0.25)^3 + 3 \times (0.25)^2 \times 0.75}{1 - (0.75)^3} \\
&= \frac{0.15625}{0.578125} \\
&= 0.2703.
\end{aligned}
$$

(b) We have:

$$
\begin{aligned}
&P(\{A_1, A_2, 1\} : \text{at least one of } A_1 \text{ and } A_2 \text{ are } 1 \,|\, A_3 = 1) \\
&= \frac{P(\{A_1, A_2, 1\} : \text{at least one of } A_1 \text{ and } A_2 \text{ are } 1)}{P(A_3 = 1)} \\
&= \frac{P(\{1,1,1\} \cup \{1,0,1\} \cup \{0,1,1\})}{0.25} \\
&= \frac{(0.25)^3 + 2 \times (0.25)^2 \times 0.75}{0.25} \\
&= 0.4375.
\end{aligned}
$$

Alternatively:

$$
\begin{aligned}
&P(\{A_1, A_2, 1\} : \text{at least one of } A_1 \text{ and } A_2 \text{ are } 1 \,|\, A_3 = 1) \\
&= P(\{A_1, A_2\} : \text{at least one of } A_1 \text{ and } A_2 \text{ is } 1) \\
&= 1 - P(\{A_1 = 0\} \cap \{A_2 = 0\}) \\
&= 1 - (0.75)^2 \\
&= 0.4375.
\end{aligned}
$$

Activity 2.1 An urn contains five balls numbered 1 to 5. We select a random sample of two balls with replacement.

(a) Draw a diagram to represent the sample space for this experiment. Mark the following events on your diagram:

- E_1 = first ball drawn is a 5
- E_2 = second ball drawn has a value less than 4
- E_3 = sum of the values of the two draws is greater than or equal to 8.

(b) Evaluate $P(E_1)$, $P(E_2)$ and $P(E_3)$.

(c) Repeat parts (a) and (b) assuming no replacement.

Activity 2.2 Six people stand in a row. We choose two people at random, say A and B. What is the probability that there will be exactly r people between A and B? If they stand in a ring, show that the probability of exactly r people between A and B in the clockwise direction is $1/5$. Try replacing 6 by n for general problems of the same sort.

Note solutions to all activities can be found in Appendix B.

2.5 Mathematical probability

We now proceed to develop a more rigorous framework for probability by considering the mathematical topic of **measure theory**. While this content may seem somewhat abstract, the effort to master it will be worth it!

2.5.1 Measure

Consider a set Ψ, and a subset $A \subset \Psi$. To determine the *size* of A, if A is finite we can simply count the number of outcomes in A. A **measure** is a function acting on subsets which gives us an idea of their size and generalises the notion of counting the number of outcomes in a set. A measure acts on subsets of the sample space, and so the domain of a measure will be a collection of subsets. We require this collection to have certain properties so that the measure can be defined.

Definition of a σ-algebra

Let Ψ be a set and let $\mathcal{B}$ be a collection of subsets of Ψ. $\mathcal{B}$ is called a **σ-algebra**, or **Borel field**, if it satisfies the following properties.

i. $\emptyset \in \mathcal{B}$, i.e. the empty set is a member of $\mathcal{B}$.

ii. If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$, i.e. $\mathcal{B}$ is closed under complementation.

iii. If $A_1, A_2, \ldots \in \mathcal{B}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$, i.e. $\mathcal{B}$ is closed under countable unions.

We now consider examples of two σ-algebras which may be constructed for any set which has a non-trivial subset.

Example 2.4 Consider a set Ψ together with a non-trivial subset $A \subset \Psi$. Two examples of σ-algebras defined on Ψ are as follows.

1. The smallest non-degenerate σ-algebra contains four elements:

$$\mathcal{B} = \{\emptyset, A, A^c, \Psi\}$$

where $A \subset \Psi$.

2. The σ-algebra with the largest number of members is given by including every subset of Ψ. This can be written as:

$$\mathcal{B} = \{A : A \subseteq \Psi\}.$$

This is referred to as the **power set** of Ψ, sometimes writen $\mathcal{P}(\Psi)$ or $\{0,1\}^{\Psi}$.

The pair consisting of a set and a σ-algebra defined on that set, $(\Psi, \mathcal{B})$, is referred to as a **measurable space**. We define measure on $(\Psi, \mathcal{B})$.

Definition of measure

Given a measurable space $(\Psi, \mathcal{B})$, a **measure** on $(\Psi, \mathcal{B})$ is a function, m, where $m : \mathcal{B} \to \mathbb{R}^+$, such that:

i. $m(A) \geq 0$ for all $A \in \mathcal{B}$

ii. $m(\emptyset) = 0$

iii. if $A_1, A_2, \ldots \in \mathcal{B}$ are mutually exclusive then:

$$m\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} m(A_i).$$

The triple $(\Psi, \mathcal{B}, m)$ consisting of a set, a σ-algebra defined on that set, and a measure is referred to as a **measure space**. Property i. says that size is a non-negative quantity. Property ii. says that if the set is empty then its size is zero. Property iii. says that for mutually exclusive sets (i.e. sets with no elements in common) then the combined set formed from their union has size equal to the sum of the sizes of the individual constituent sets.

We now proceed to define an indicator function.

Definition of an indicator function

The **indicator function** for a set $\Omega \subseteq \mathbb{R}$ is $I_\Omega(x)$ where:

$$I_\Omega(x) = \begin{cases} 1 & \text{for } x \in \Omega \\ 0 & \text{otherwise.} \end{cases}$$

In applications, the set Ω in the indicator function definition will be the sample space of a random variable which, for continuous random variables, takes the form of an interval such as $[0, 1]$ or $[0, \infty)$. The indicator function $I_\Omega(x)$ is a function which takes two values; 1 if $x \in \Omega$ and 0 if $x \notin \Omega$.

Example 2.5 If $\Omega = [0, \infty)$, we have:

$$I_\Omega(x) = I_{[0,\infty)}(x) = \begin{cases} 1 & \text{for } x \geq 0 \\ 0 & \text{for } x < 0. \end{cases}$$

Counts and indicators as measures

Suppose that $(\Psi, \mathcal{B})$ is a measurable space.

i. If Ψ is finite and we define m by $m(A) = n(A)$ for all $A \in \mathcal{B}$, then m is a measure.

ii. If I_A is the indicator function for set A, then the function defined by $m(A) = I_A(x)$ for given $x \in \Psi$ is a measure.

2.5.2 Probability measure

Using Section 2.5.1, we proceed to develop a rigorous definition of probability. Recall that *measure* allows us to express the *size* of sets, while *probability* expresses *how likely* an event is to occur. Combining these ideas allows us to define probability as a measure.

In order to define a measure, a measurable space is required, i.e. a set and a σ-algebra defined on the set. We have already defined the sample space, Ω, to be the set of all possible outcomes in an experiment, and events to be subsets of Ω which are of particular interest. We now generate a measurable space, $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is a σ-algebra defined on Ω. Here, $\mathcal{F}$ is a collection of subsets of Ω, with the elements of $\mathcal{F}$ interpreted as events. Therefore, if $A \in \mathcal{F}$, then A is an event. Since probability is always associated with events, so $\mathcal{F}$ will be the domain for probability measure.

Definition of probability measure

Given a measurable space $(\Omega, \mathcal{F})$, a **probability measure** on $(\Omega, \mathcal{F})$ is a measure $P : \mathcal{F} \to [0, 1]$ with the property that $P(\Omega) = 1$.

(Note that the codomain of P is the unit interval, $[0, 1]$, as we would wish since probabilities are always in $[0, 1]$.)

The triple $(\Omega, \mathcal{F}, P)$ consisting of a sample space, a collection of events (forming a σ-algebra on the sample space) and a probability measure is referred to as a **probability space**.

> **Example 2.6** Consider the measurable space $(\Omega, \mathcal{F})$. Two previously-encountered functions satisfy the properties of probability measures.
>
> 1. $P(A) = n(A)/n(\Omega)$ defines a probability measure P for $A \in \mathcal{F}$.
> 2. If we define $P(A) = I_A(\omega)$ for some $\omega \in \Omega$ this is a perfectly valid probability measure. It takes the value 1 for any event containing ω, and 0 for all other events.

We have now derived a definition of probability based on the set-theoretic notion of

measure. This allows us to intepret various probability statements for events A and B as follows.

- $P(A)$ means the probability that the outcome of the experiment is in A.
- $P(A^c)$ means the probability that the outcome of the experiment is not in A.
- $P(A \cap B)$ means the probability that the outcome of the experiment is in both A and B.
- $P(A \cup B)$ means the probability that the outcome of the experiment is in A or B or in both (i.e. inclusive or).
- $P(A \setminus B)$ means the probability that the outcome of the experiment is in A but not in B.

Remember to distinguish between *outcomes* and *events*. An *outcome* is an element of the sample space, whereas an *event* is a subset of the sample space.

> **Example 2.7** Suppose $\Omega = \{\omega_1, \omega_2, \ldots\}$, then $\omega_1, \omega_2, \ldots$ are outcomes, and subsets such as $\{\omega_1, \omega_2, \omega_3\}$ and $\{\omega_5, \omega_7\}$ are events. $\{\omega_1\}$ is also an event.

In all practical examples we will consider, the collection of events is taken to be the power set $\mathcal{F} = \{0, 1\}^\Omega$, i.e. the collection of all possible subsets of Ω. This means sets containing single outcomes, such as $\{\omega_1\}$ and $\{\omega_2\}$, are events. Note that events containing different single outcomes are mutually exclusive, by definition.

Events form a σ-algebra on the sample space, and probability is a measure on the resulting measurable space. Hence events form a collection which satisfies the properties of a σ-algebra, and probability is a function which satisfies the properties of a measure.

Empty set and sample space

Consider a probability space $(\Omega, \mathcal{F}, P)$. By property i. of a σ-algebra, $\emptyset \in \mathcal{F}$. Since $\Omega = \emptyset^c$, by property ii. of a σ-algebra, $\Omega \in \mathcal{F}$. Hence both the empty set, $\emptyset$, and the sample space, Ω, are events.

By property i. of a measure, $P(\emptyset) = 0$. A probability measure has the additional property that $P(\Omega) = 1$.

Note you should *not* think of the event $\emptyset$ as 'nothing happening'. 'Nothing happening' could well be an outcome with strictly positive probability in some experiments. Rather, remember that since $P(\Omega) = 1$ the sample space includes all possible outcomes of the experiment, so Ω is often referred to as a **certain event**, and $\emptyset$ as an **impossible event**.

Complement

Consider a probability space $(\Omega, \mathcal{F}, P)$. Property ii. of a σ-algebra says that if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$. Hence if A is an event, then A^c is also an event. By being interested in outcomes in A, by default we are also interested in outcomes which are not in A, hence we can speak about the probability of events not happening. Therefore, $A^c \in \mathcal{F}$ means that $P(A^c)$ is well-defined.

Union

Consider a probability space $(\Omega, \mathcal{F}, P)$. Property iii. of a σ-algebra says that if $A_1, A_2, \ldots \in \mathcal{F}$, then:

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

Therefore, the union of a collection of events is also an event. The property could be expressed as 'the collection of events is closed under union'. It ensures that if A and B are events, then $P(A \cup B)$ is well-defined.

In addition, property iii. of a measure tells us that if A and B do not have any outcomes in common, i.e. they are mutually exclusive events, then:

$$P(A \cup B) = P(A) + P(B).$$

A number of useful results can be derived now that probability has been defined as a measure. We now have at our disposal tools which allow us to establish some probability properties from first principles.

Note that a *lemma* is a proven proposition which is used as a stepping stone to a more important result, and a *corollary* is a statement which follows readily from a previous statement.

Basic probability properties

Lemma: Consider a probability space $(\Omega, \mathcal{F}, P)$, with $A \in \mathcal{F}$ and $B \in \mathcal{F}$, i.e. A and B are events.

i. $P(A^c) = 1 - P(A)$.

ii. If $A \subseteq B$, then $P(B \setminus A) = P(B) - P(A)$.

iii. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Corollary: If $A \subseteq B$, then $P(A) \leq P(B)$.

Appendix A provides non-examinable proofs.

Example 2.8 Suppose A and B are events with probabilities $P(A) = 3/4$ and $P(B) = 1/3$. Show that:

$$\frac{1}{12} \leq P(A \cap B) \leq \frac{1}{3}$$

and give examples to show that both bounds are possible.

Solution

We have:

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1 = \frac{1}{12}.$$

Also, since $A \cap B \subseteq A$ and $A \cap B \subseteq B$, then:

$$P(A \cap B) \leq \min(P(A), P(B)) = \frac{1}{3}.$$

Example: choose $\Omega = \{1, 2, \ldots, 12\}$ and $A = \{1, 2, \ldots, 9\}$. Taking $B = \{9, 10, 11, 12\}$ gives the lower bound while $B = \{1, 2, 3, 4\}$ gives the upper bound.

We now consider the general addition law.

General addition law

Consider a probability space $(\Omega, \mathcal{F}, P)$ with $A_1, A_2, \ldots, A_n \in \mathcal{F}$, then:

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n).$$

One of the properties inherited from measure is that the probability of a union of disjoint events is the sum of the individual probabilities. When the events are not mutually exclusive, the relationship becomes an inequality, known as the **Boole inequality**.

Boole inequality

Consider a probability space $(\Omega, \mathcal{F}, P)$ with $A_1, A_2, \ldots \in \mathcal{F}$, then:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i).$$

Appendix A provides a non-examinable proof.

We now state some propositions which establish the connection between limits of probabilities and sequences of sets. The concept of a sequence of sets will be seen in Chapter 3 when considering results associated with random variables.

Probability limit of a sequence of sets

i. If $\{A_i : i = 1, 2, \ldots\}$ is an increasing sequence of sets $A_1 \subseteq A_2 \subseteq \cdots$, then:

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcup_{i=1}^{\infty} A_i\right).$$

ii. If $\{A_i : i = 1, 2, \ldots\}$ is a decreasing sequence of sets $A_1 \supseteq A_2 \supseteq \cdots$, then:

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcap_{i=1}^{\infty} A_i\right).$$

Appendix A provides non-examinable proofs.

Activity 2.3 This question concerns various facts about σ-algebras.

(a) What is the smallest possible σ-algebra on Ψ?

(b) Show that $\{\emptyset, A, A^c, \Psi\}$, where $A \subset \Psi$, is a σ-algebra on Ψ.

(c) If $|\Psi| < \infty$, how many elements are there in the power set $\{0, 1\}^{\Psi}$? Can you think of an explanation for the notation $\{0, 1\}^{\Psi}$?

(d) Use de Morgan's laws (see **ST104b Statistics 2**) to show that a σ-algebra is closed under intersection, i.e. if $A_1, A_2, \ldots, \in \mathcal{B}$ then $\bigcap_{i=1}^{\infty} A_i \in \mathcal{B}$.

Activity 2.4 Consider a set Ψ and two subsets $A, B \subset \Psi$, where $A \cap B \neq \emptyset$ but $A \cap B \neq A$ and $A \cap B \neq B$. What is the size of the smallest σ-algebra containing both A and B?

Activity 2.5 Consider a measurable space $(\Psi, \mathcal{B})$. Show that, if $A \in \mathcal{B}$, both of the following are measures:

(a) $m(A) = n(A)$

(b) $m(A) = I_A(x)$ for given $x \in \Psi$.

Activity 2.6 Consider the sample space Ω and event A. Show that both of the functions P defined below satisfy the conditions for a probability measure.

(a) $P(A) = n(A)/n(\Omega)$ where $n(\Omega) < \infty$.

(b) $P(A) = I_A(\omega)$ for some fixed $\omega \in \Omega$.

Activity 2.7 Show that if $A_1, A_2, \ldots \in \mathcal{F}$ and we define $B_1 = A_1$ and:

$$B_i = A_i \setminus \bigcup_{j=1}^{i-1} B_j$$

for $i > 1$, then $B_i \in \mathcal{F}$ and:

$$\bigcup_{j=1}^{i} B_j = \bigcup_{j=1}^{i} A_j$$

for all i. Also, show $B_i \cap B_j = \emptyset$ for any $i \neq j$.

2.6 Methods for counting outcomes

Consider a finite sample space Ω made up of $n(\Omega)$ equally likely outcomes, and a probability measure $P(A) = n(A)/n(\Omega)$ for all $A \in \mathcal{F}$. In such classical probability problems, it is necessary to count outcomes in order to determine $n(A)$ and $n(\Omega)$.

For 'small' problems, brute force can be used by just listing all possibilities, such as in the following example.

Example 2.9 Consider a group of four people, where each pair of people is either connected (i.e. are friends) or not. How many different *patterns* of connections are there (ignoring the identities of who is friends with whom)?

The answer is 11. See the patterns in Figure 2.1.

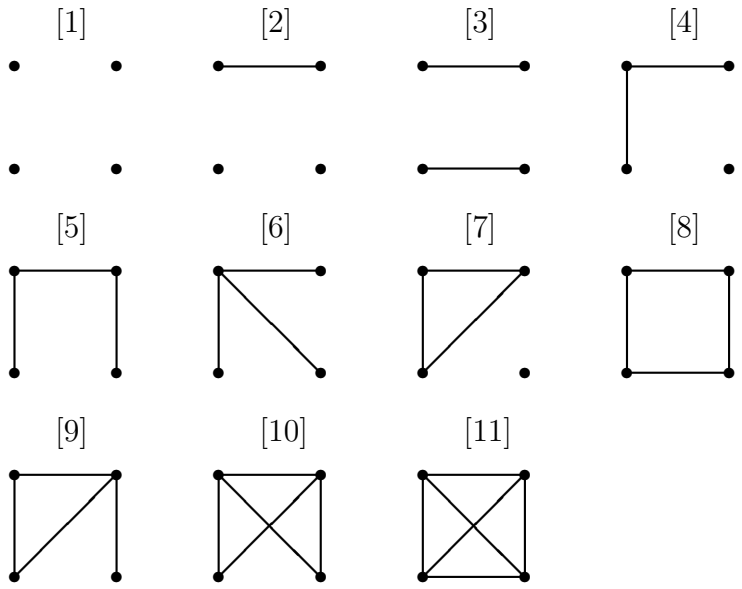

Figure 2.1: Friendship patterns in a four-person network.

In many practical experiments, the sample space is too large to consider such 'brute force' methods hence we may choose to combine counts using rules of sum and product.

Rule of sum

If an element can be selected in n_1 ways from set 1, *or* n_2 ways from set 2, ... *or* n_k ways from set k, the total number of possible selections is:

$$n_1 + n_2 + \cdots + n_k.$$

Equivalently, let $E_1, E_2, \ldots, E_k$ be experiments with, respectively, n_i possible outcomes, for $i = 1, 2, \ldots, k$, then the total number of possible outcomes is:

$$\sum_{i=1}^{k} n_i.$$

Rule of product

If, in an ordered sequence of k elements, element 1 can be selected in n_1 ways, *and* then element 2 in n_2 ways, ... *and* then element k in m_k ways, the total number of possible sequences is:

$$n_1 \times n_2 \times \cdots \times n_k.$$

Equivalently, let $E_1, E_2, \ldots, E_k$ be experiments with, respectively, n_i possible outcomes, for $i = 1, 2, \ldots, k$, then the experiment consisting of the ordered sequence $(E_1, E_2, \ldots, E_k)$ has:

$$\prod_{i=1}^{k} n_i$$

possible outcomes.

Before reviewing permutations and combinations, recall that the number of possible ordered sequences of k objects selected with replacement from n objects is:

$$\overbrace{n \times n \times \cdots \times n}^{k \text{ times}} = n^k.$$

2.6.1 Permutations and combinations

Recall the definitions of permutations and combinations introduced in **ST104b Statistics 2**. Here, these will be used as general counting rules to determine $n(A)$ and $n(\Omega)$, and hence $P(A) = n(A)/n(\Omega)$, when the outcomes in Ω are equally likely.

Permutations

The number of **permutations**, i.e. the number of possible *ordered* sequences, of k objects selected without replacement from n objects is:

$${}^{n}P_k = \frac{n!}{(n-k)!}.$$

Combinations

The number of **combinations**, i.e. the number of *unordered* sequences, of k objects selected without replacement from n objects is:

$$^nC_k = \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

which is also known as the **binomial coefficient**.

Note that the number of permutations is a direct consequence of the rule of product, since:

$$^nP_k = \frac{n!}{(n-k)!} = n \times (n-1) \times \cdots \times (n-k+1).$$

Example 2.10 Consider a set Ω consisting of the first 9 integers, i.e. we have $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. An example of a permutation of four integers from Ω is $\omega_1 = \{5, 8, 2, 6\}$. The permutation $\omega_2 = \{2, 6, 8, 5\}$ is distinct from ω_1 because although the elements are the same, the order is different. In total, the number of permutations of four integers from Ω is:

$$^9P_4 = \frac{9!}{(9-4)!} = \frac{9!}{5!} = 9 \times 8 \times 7 \times 6 = 3{,}024.$$

The number of combinations of four integers from Ω is:

$$^9C_4 = \binom{9}{4} = \frac{9!}{4!\,(9-4)!} = \frac{^9P_4}{4!} = \frac{3{,}024}{24} = 126.$$

Example 2.11 Five playing cards are drawn from a well-shuffled deck of 52 playing cards. What is the probability that the cards form a hand which is higher than 'a flush'? The cards in a hand are treated as an unordered set.

First, we determine the size of the sample space which is all unordered subsets of 5 cards selected from 52. So the size of the sample space is:

$$\binom{52}{5} = \frac{52!}{5! \times 47!} = \frac{52 \times 51 \times 50 \times 49 \times 48}{5 \times 4 \times 3 \times 2 \times 1} = 2{,}598{,}960.$$

The hand is higher than a flush if it is a:

'straight flush' *or* 'four-of-a-kind' *or* 'full house'.

The rule of sum says that the number of hands better than a flush is:

$$\begin{aligned}
&\text{number of straight flushes} + \text{number of four-of-a-kinds} \\
&+ \text{number of full houses} \\
&= 40 + 624 + 3{,}744 \\
&= 4{,}408.
\end{aligned}$$

Therefore, the probability we want is:

$$\frac{4{,}408}{2{,}598{,}960} \approx 0.0017.$$

How did we get the counts above?

- For full houses, shown next.
- For the others, see Appendix A.

A 'full house' is three cards of the same rank *and* two cards of another rank, for example:

$$\diamondsuit 2\ \spadesuit 2\ \clubsuit 2\ \diamondsuit 4\ \spadesuit 4.$$

We can break the number of ways of choosing these into two steps.

- The total number of ways of selecting the three: the rank of these can be any of the 13 ranks. There are four cards of this rank, so the three of that rank can be chosen in $\binom{4}{3} = 4$ ways. So the total number of different triplets is $13 \times 4 = 52$.
- The total number of ways of selecting the two: the rank of these can be any of the remaining 12 ranks, and the two cards of that rank can be chosen in $\binom{4}{2} = 6$ ways. So the total number of different pairs (with a different rank than the triplet) is $12 \times 6 = 72$.

The rule of product then says that the total number of full houses is:

$$52 \times 72 = 3{,}744.$$

The following is a summary of the numbers of all types of 5-card hands, and their probabilities (see Appendix A for determination of all the probabilities). **Note you do not need to memorise the different types of hand!**

Hand	Number	Probability
Straight flush	40	0.000015
Four-of-a-kind	624	0.00024
Full house	3,744	0.00144
Flush	5,108	0.0020
Straight	10,200	0.0039
Three-of-a-kind	54,912	0.0211
Two pairs	123,552	0.0475
One pair	1,098,240	0.4226
High card	1,302,540	0.5012
Total	2,598,960	1.0

Summary of the combinatorial counting rules

The number of k outcomes from n distinct possible outcomes can be summarised as follows:

	With replacement	Without replacement
Ordered	n^k	$\dfrac{n!}{(n-k)!}$
Unordered	$\binom{n+k-1}{k}$	$\binom{n}{k} = \dfrac{n!}{k!\,(n-k)!}$

We have not yet discussed the unordered, with replacement case. It is provided here through an illustration given in Example 2.12.

Example 2.12 We consider an outline of the proof, using $n = 5$ and $k = 3$ for illustration. Half-graphically, let `x` denote selected values and `|` the 'walls' between different distinct values. For example:

- `x|xx|||` denotes the selection of set $(1, 2, 2)$
- `x||x||x` denotes the set $(1, 3, 5)$
- `||||xxx` denotes the set $(5, 5, 5)$.

In general, we have a sequence of $n + k - 1$ symbols, i.e. $n - 1$ walls (`|`) and k selections (`x`). The number of different unordered sets of k objects selected with replacement from n objects is the number of different ways of choosing the locations of the `x`s in this, that is:

$$\binom{n+k-1}{k}.$$

2.6.2 Combinations and multinomial coefficients

From the definition of combinations, we can think of the quantity $\binom{n}{k}$ as the number of ways of *choosing* k objects from n.

Proposition

Consider a collection of n objects, k of which are of type a and $(n - k)$ of which are of type b. The number of ways of arranging these objects into sequences of type a and type b is $\binom{n}{k}$.

Proof: View this as a problem of positioning k objects of type a into n slots, with the remaining $(n - k)$ slots filled with type b objects. Suppose the slots are labelled

$1, 2, \ldots, n$, then the problem is equivalent to selecting a set of k numbers from $\{1, 2, \ldots, n\}$, with each number chosen giving us the position occupied by a type a object, hence order is unimportant. The number of ways of doing this is $\binom{n}{k}$. ■

Expansions of expressions of the form $(a+b)^n$ also make use of combinations. Consider a few cases:

$$\begin{aligned}(a+b) &= a+b\\ (a+b)^2 &= (a+b)(a+b) = a^2+ab+ba+b^2 = a^2+2ab+b^2\\ (a+b)^3 &= (a+b)(a+b)(a+b) = \cdots = a^3+3a^2b+3ab^2+b^3\\ &\vdots\end{aligned}$$

The above proposition allows us to write down a general closed form for $(a+b)^n$. In general, $\binom{n}{k}$ will be the coefficient of $a^k b^{n-k}$ in the expansion of $(a+b)^n$, such that:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

This is why quantities of the form $\binom{n}{k}$ are sometimes referred to as *binomial coefficients*.

Basic properties of binomial coefficients

i. We have:

$$\binom{n}{k} = \binom{n}{n-k}.$$

In particular:

$$\binom{n}{0} = \binom{n}{n} = 1 \quad \text{and} \quad \binom{n}{1} = \binom{n}{n-1} = n.$$

ii. We have:

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}.$$

iii. We have:

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$

Property ii. of binomial coefficients states that each binomial coefficient is the sum of two coefficients of lower order, which forms the basis of **Pascal's triangle**[1]. Starting with 1s down the outside of the triangle, we sum adjacent entries to generate the next row, as shown in Figure 2.2. If we label the top row of the triangle as row 0 then, for $k = 0, 1, 2, \ldots, n$, the binomial coefficient $\binom{n}{k}$ is the kth entry in the nth row.

Property iii. of binomial coefficients can be considered in the context of the binary representation of integers. Suppose we have n binary digits, i.e. n bits. Each bit can

[1] Technically, this should be called **Yanghui's triangle** because it was discovered five centuries before Pascal by the Chinese mathematician Yanghui.

$$
\begin{array}{ccccccccccccccccccc}
 & & & & & & & & 1 & & & & & & & & & & \\
 & & & & & & & 1 & & 1 & & & & & & & & & \\
 & & & & & & 1 & & 2 & & 1 & & & & & & & & \\
 & & & & & 1 & & 3 & & 3 & & 1 & & & & & & & \\
 & & & & 1 & & 4 & & 6 & & 4 & & 1 & & & & & & \\
 & & & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & & & & & \\
 & & 1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 & & & & \\
 & 1 & & 7 & & 21 & & 35 & & 35 & & 21 & & 7 & & 1 & & & \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots
\end{array}
$$

Figure 2.2: Pascal's (Yanghui's) triangle.

take the value 0 or 1, so by the rule of product we can represent 2^n integers using n bits. If we start at 0, then these numbers will be $0, 1, 2, \ldots, 2^n - 1$.

We can generalise the notion of a binomial coefficient to the multinomial case.

Multinomial coefficient

Consider n objects where n_1 are of type 1, n_2 are of type 2, ..., and n_r are of type r. The number of ways of arranging these n objects is:

$$(n_1, n_2, \ldots, n_r)! = \frac{n!}{n_1!\, n_2! \cdots n_r!}$$

where we have $\sum_{i=1}^{r} n_i = n$. Quantities of the form $(n_1, n_2, \ldots, n_r)!$ are referred to as **multinomial coefficients**.

Appendix A provides a non-examinable proof.

Note in the binomial case we have:

$$(k, n-k)! = \binom{n}{k}.$$

The multinomial coefficients feature in the multinomial expansion:

$$(a_1 + a_2 + \cdots + a_r)^n = \sum_{n_1, n_2, \ldots, n_r \geq 0} (n_1, n_2, \ldots, n_r)!\, a_1^{n_1} a_2^{n_2} \cdots a_r^{n_r}$$

where $\sum_{i=1}^{r} n_i = n$.

Example 2.13 Consider a series of n independent trials, each of which can result in one of three possible outcomes. Let n_1 and n_2 denote the number of these trials which result in the first and second types of outcome, respectively, with corresponding probabilities of π_1 and π_2. We derive an expression which computes the probability of n_1 and n_2 occurrences of the first and second outcomes, respectively.

If there are n_1 occurrences of the first type of outcome, and n_2 of the second type, then there must be $n - n_1 - n_2$ occurrences of the third type each with probability $1 - \pi_1 - \pi_2$. By independence, any such sequence occurs with probability:

$$\pi_1^{n_1}\pi_2^{n_2}(1 - \pi_1 - \pi_2)^{n-n_1-n_2}.$$

There are $\binom{n}{n_1}$ positions for the first type of outcome and subsequently there are $\binom{n-n_1}{n_2}$ positions for the second type, with the remaining $n - n_1 - n_2$ positions having to be the $n - n_1 - n_2$ outcomes of the third type. Hence the probability of n_1 and n_2 occurrences of the first and second outcomes is:

$$\begin{aligned}
&\binom{n}{n_1}\binom{n-n_1}{n_2}\pi_1^{n_1}\pi_2^{n_2}(1 - \pi_1 - \pi_2)^{n-n_1-n_2} \\
&= \frac{n!}{n_1!\,(n-n_1)!}\,\frac{(n-n_1)!}{n_2!\,(n-n_1-n_2)!}\,\pi_1^{n_1}\pi_2^{n_2}(1 - \pi_1 - \pi_2)^{n-n_1-n_2} \\
&= \frac{n!}{n_1!\,n_2!\,(n-n_1-n_2)!}\,\pi_1^{n_1}\pi_2^{n_2}(1 - \pi_1 - \pi_2)^{n-n_1-n_2}.
\end{aligned}$$

Example 2.14 Patients are treated for one of three types of virus (types 1, 2 and 3), each requiring a different drug treatment. The procurement officer in a hospital needs to decide on how many of each drug treatment to stock. From experience, the probability that a patient is admitted with each of the virus types is 0.6, 0.3 and 0.1, respectively. Using the result of Example 2.13 we calculate the probability that for the next 20 patients that 10 will need the drug treatment for the first type of virus, 7 will need treatment for the second type, with the other 3 being admitted for the third type.

The probability is:

$$\frac{n!}{n_1!\,n_2!\,(n-n_1-n_2)!}\,\pi_1^{n_1}\pi_2^{n_2}(1 - \pi_1 - \pi_2)^{n-n_1-n_2}$$

with $n = 20$, $k_1 = 10$, $k_2 = 7$, $\pi_1 = 0.6$ and $\pi_2 = 0.3$. Hence we have:

$$\frac{20!}{10! \times 7! \times 3!} \times (0.6)^{10} \times (0.3)^7 \times (0.1)^3 = 0.0293.$$

Activity 2.8 A regular insurance claimant is trying to hide three fraudulent claims among seven genuine claims. The claimant knows that the insurance company processes claims in batches of five or in batches of ten. For batches of five, the insurance company will investigate one claim at random to check for fraud; for batches of 10, two of the claims are randomly selected for investigation. The claimant has three possible strategies:

(a) submit all ten claims in a single batch

(b) submit two batches of five, one containing two fraudulent claims, the other containing one

(c) submit two batches of five, one containing three fraudulent claims, the other containing none.

What is the probability that all three fraudulent claims will go undetected in each case? What is the optimal strategy for the fraudster?

2.7 Conditional probability and independence

In light of our probability space $(\Omega, \mathcal{F}, P)$, we review the definition of conditional probability which was introduced in **ST104b Statistics 2**. Conditional probability allows us to update our probabilistic belief in some event A occurring, given that another event, such as B, has occurred. In effect, as we move from the unconditional probability $P(A)$ to the conditional probability $P(A \mid B)$, the sample space shrinks from Ω to B.

Conditional probability

Consider the probability space $(\Omega, \mathcal{F}, P)$ and events $A, B \in \mathcal{F}$, with $P(B) > 0$. The **conditional probability** of A given B is:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Note $P(B) > 0$ is required for $P(A \mid B)$ to be defined. We cannot condition on events with zero probability.

A brief word on conditioning notation. It is important to distinguish between $P(A \mid B)$ and $P(A \setminus B)$. Despite similar notation, the meanings are completely different.

- $P(A \mid B) = P(A \cap B)/P(B)$, i.e. the conditional probability of A given B.
- $P(A \setminus B) = P(A \cap B^c)$, i.e. the probability of A occurring but not B.

Example 2.15 Let G be the event that a defendant is guilty, and T be the event that some testimony is true. Some lawyers have argued along the lines of the assumption $P(G \mid T) = P(T \mid G)$. When does this hold?

Solution

It is true only when $P(G) = P(T)$. Indeed:

$$P(G \mid T) = \frac{P(G \cap T)}{P(T)} \quad \text{and} \quad P(T \mid G) = \frac{P(G \cap T)}{P(G)}.$$

Example 2.16 Consider Example 2.1 again where $P(A) = 1/12$. Now suppose that event B is the event that the value on the second die is a 6, as illustrated in Figure

2.3. We see that $n(B) = 6$ and $n(A \cap B) = 2$, hence $P(B) = 1/6$ and $P(A \cap B) = 1/18$. Therefore:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/18}{1/6} = \frac{1}{3}.$$

		Second die roll					
		1	2	3	4	5	6
	1	.	.	.	.	.	B
	2	.	.	.	.	.	B
First die roll	3	.	.	.	.	.	B
	4	.	.	.	.	.	B
	5	.	.	.	.	.	A, B
	6	.	.	.	.	A	A, B

Figure 2.3: Sample space for the experiment of throwing two fair dice with events A and B denoted.

Note that when all the outcomes in Ω are equally likely, we have:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{n(A \cap B)/n(\Omega)}{n(B)/n(\Omega)} = \frac{n(A \cap B)}{n(B)}.$$

By conditioning on B, we effectively assign B the role of being the sample space in our probability calculations. This is reasonable because we know that the outcome of the experiment is in B.

We note that the function $P(\cdot \mid B)$ is a probability measure on $(\Omega, \mathcal{F})$.

Simple properties of conditional probability

i. $P(A \cap B) = P(A \mid B)\, P(B)$.

ii. If A and B are mutually exclusive, then $P(A \mid B) = 0$.

iii. If $B \subseteq A$, then $P(A \mid B) = 1$.

Property i. above was introduced in **ST104b Statistics 2** as the **chain rule** of conditional probabilities. Consider the three-way intersection $A_1 \cap A_2 \cap A_3$. By the transitivity of intersections, $A_1 \cap A_2 \cap A_3 = A_1 \cap (A_2 \cap A_3)$, hence:

$$P(A_1 \cap A_2 \cap A_3) = P(A_1 \mid A_2 \cap A_3)\, P(A_2 \cap A_3) = P(A_1 \mid A_2 \cap A_3)\, P(A_2 \mid A_3)\, P(A_3).$$

This generalises as follows.

Chain rule of conditional probabilities

Consider the probability space $(\Omega, \mathcal{F}, P)$ with $A_1, A_2, \ldots, A_n \in \mathcal{F}$ and where:

$$P\left(\bigcap_{j=1}^{n} A_j\right) \neq 0$$

then:

$$P\left(\bigcap_{i=1}^{n} A_j\right) = \prod_{j=1}^{n} P\left(A_j \,\middle|\, \bigcap_{i=0}^{j-1} A_i\right)$$

where we define $A_0 = \Omega$.

Example 2.17 Show that for events A, B and C with positive probability it holds that:

$$P(A \cap B \cap C) = P(A \,|\, B \cap C)\, P(B \,|\, C)\, P(C).$$

Solution

Note that:

$$\begin{aligned} P(A \,|\, B \cap C)\, P(B \,|\, C)\, P(C) &= \frac{P(A \cap B \cap C)}{P(B \cap C)} \frac{P(B \cap C)}{P(C)} P(C) \\ &= P(A \cap B \cap C). \end{aligned}$$

2.7.1 Total probability formula and Bayes' theorem

Recall that Bayes' theorem is a mechanism for reversing the order of conditioning.

Example 2.18 An insurance company processes claims for car theft on a monthly basis. An individual's probability of having their car stolen in any given month is 0.005, and 99% of those who have their car stolen make a claim. A small proportion of customers (1 in 10,000) will put in a claim even though their car has not been stolen (i.e. they commit insurance fraud).

- What is the probability that an individual customer makes a claim?
- What is the probability that a claim made by a customer is fraudulent?

We will return to this example later.

As first seen in **ST104b Statistics 2**, many probability problems can be made more tractable by dividing the sample space into a partition.

Partition

Consider a probability space $(\Omega, \mathcal{F}, P)$. A set of events $\{B_1, B_2, \ldots, B_n\}$ is called a **partition** if it has the following properties.

i. Exhaustive: $\bigcup_{i=1}^{n} B_i = \Omega$.

ii. Mutually exclusive: $B_i \cap B_j = \emptyset$ for all $i \neq j$.

iii. Non-zero probability: $P(B_i) > 0$ for $i = 1, 2, \ldots, n$.

Since the elements of the partition are events, we have $B_i \in \mathcal{F}$ for all i. The exhaustive property means that the members of a partition cover the whole sample space, i.e. for any $\omega \in \Omega$ we have $\omega \in B_i$ for some i.

Example 2.19 Consider the mutually exclusive and exhaustive events A, B and C. Is it possible to have the following probabilities?

$$P(A \cup B) = \frac{1}{2}, \quad P(B \cup C) = \frac{1}{2} \quad \text{and} \quad P(C \cup A) = \frac{2}{3}.$$

Solution

First note that all of these are probabilities of the sum of mutually exclusive events so:

$$P(A) + P(B) = \frac{1}{2}, \quad P(B) + P(C) = \frac{1}{2} \quad \text{and} \quad P(C) + P(A) = \frac{2}{3}.$$

We may add up these three equations to give:

$$2 \times (P(A) + P(B) + P(C)) = 1 + \frac{2}{3} \quad \Leftrightarrow \quad P(A) + P(B) + P(C) = \frac{5}{6}.$$

However, we are told A, B and C are exhaustive so $P(A) + P(B) + P(C) = 1$. Hence it is not possible to have the suggested probabilities.

Recall that, for mutually exclusive events, the probability of the union is the sum of the probabilities. This is applied to derive the **total probability formula**.

Total probability formula

Consider the probability space $(\Omega, \mathcal{F}, P)$ with a partition $\{B_1, B_2, \ldots, B_n\}$ of Ω. For all $A \in \mathcal{F}$, then:

$$P(A) = \sum_{i=1}^{n} P(A \,|\, B_i)\, P(B_i).$$

Proof: We have:

$$\bigcup_{i=1}^{n} (A \cap B_i) = A \cap \bigcup_{i=1}^{n} B_i.$$

Since $\{B_1, B_2, \ldots, B_n\}$ is a partition, we know that $\bigcup_{i=1}^{n} B_i = \Omega$ by the exhaustive property. Therefore, $\bigcup_{i=1}^{n} (A \cap B_i) = A$. The mutually exclusive property of the partition tells us that:

$$(A \cap B_i) \cap (A \cap B_j) = A \cap (B_i \cap B_j) = A \cap \emptyset = \emptyset$$

hence $(A \cap B_i)$ and $(A \cap B_j)$ are also mutually exclusive. Hence we have:

$$P(A) = P\left(\bigcup_{i=1}^{n} (A \cap B_i)\right) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A \mid B_i)\, P(B_i)$$

where the conditional probability is guaranteed well-defined since $P(B_i) > 0$ for all $i = 1, 2, \ldots, n$. ■

Bayes' theorem – different forms

Consider the probability space $(\Omega, \mathcal{F}, P)$ with $A, B \in \mathcal{F}$ and $\{B_1, B_2, \ldots, B_n\}$ is a partition of Ω.

i. First form:

$$P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A)}.$$

ii. Second form:

$$P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A \mid B)\, P(B) + P(A \mid B^c)\, P(B^c)}.$$

iii. Third form:

$$P(B_j \mid A) = \frac{P(A \mid B_j)\, P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)\, P(B_i)}.$$

Example 2.20 Continuing Example 2.18, we address the first question using the total probability formula. Let A denote the event that a customer makes a claim, B_1 denote the event that their car is stolen, and B_2 denote the event their car is not stolen. We know:

$$P(B_1) = 0.005, \quad P(B_2) = 0.995, \quad P(A \mid B_1) = 0.99 \quad \text{and} \quad P(A \mid B_2) = 0.0001.$$

Note that $B_2 = B_1^c$ so $\{B_1, B_2\}$ is a partition of Ω. The probability that a customer makes a claim is then:

$$P(A) = \sum_{i=1}^{2} P(A \mid B_i)\, P(B_i) = (0.99 \times 0.005) + (0.0001 \times 0.995) = 0.0050495.$$

As the insurance company, we observe whether or not a customer has made a claim, i.e. we know whether or not event A has occurred. However, we do not observe

whether or not a claim is fraudulent (i.e. was the car stolen?) so we have to determine this probability using Bayes' theorem, such that:

$$P(B_2 \mid A) = \frac{P(A \mid B_2)\, P(B_2)}{P(A)} = \frac{0.0001 \times 0.995}{0.0050495} = 0.0197.$$

Example 2.21 Suppose that 30% of computer owners use a Mac, 50% use Windows, and 20% use Linux. Suppose that 65% of the Mac users have succumbed to a computer virus, 82% of the Windows users get the virus, and 30% of the Linux users get the virus. We select a person at random. What is the probability that:

(a) the person's computer has been infected with the virus?

(b) the person is a Windows user, given that their system has already been infected with the virus?

Solution

Let B = a user's computer has the virus, and A_i, for $i = 1, 2, 3$, denote a Mac, Windows and Linux user, respectively.

(a) Using the total probability formula:

$$\begin{aligned} P(B) &= \sum_{i=1}^{3} P(B \cap A_i) \\ &= \sum_{i=1}^{3} P(B \mid A_i)\, P(A_i) \\ &= 0.65 \times 0.30 + 0.82 \times 0.50 + 0.30 \times 0.20 \\ &= 0.665. \end{aligned}$$

(b) We have:

$$P(A_2 \mid B) = \frac{P(A_2 \cap B)}{P(B)} = \frac{0.82 \times 0.50}{0.665} = 0.6165.$$

Example 2.22 There are three cards: one is green on both sides, one is red on both sides, and one is green on one side and red on the other side. We choose a card at random and we see one side (also chosen at random). If the side we see is green, what is the probability that the other side is also green? Many people intuitively answer 1/2. Show that the correct answer is 2/3.

Solution

Let S_r and S_g denote the side we see is red and green, respectively. Let C_r, C_g and C_{rg} denote the card we have is two-sided red, two-sided green and the one with different colours, respectively.

The required probability is:

$$P(C_g \mid S_g) = \frac{P(C_g \cap S_g)}{P(S_g)} = \frac{P(S_g \cap C_g)\, P(C_g)}{P(S_g \cap C_{rg}) + P(S_g \cap C_g) + P(S_g \cap C_r)}$$
$$= \frac{1 \times 1/3}{(1/2 + 1 + 0) \times 1/3}$$
$$= \frac{2}{3}.$$

Alternatively, treat each side as a separate 'subject', we then have in total 6 subjects: three red, and three green. Furthermore, two in red are 'linked', two in green are 'linked', and the other two (one green and one red) are 'linked'. Now we have randomly selected one, which happens to be green. Of course, the probability that this green was from the 'linked' pair is 2/3.

Example 2.23 There are n urns of which the kth contains $k-1$ red balls and $n-k$ black balls. You pick an urn at random and remove two balls at random without replacement. Find the probability that:

(a) the first ball is black

(b) the second ball is black, given that the first is black.

Solution

Let C_i denote the colour of the ith ball picked, for $i = 1, 2$.

(a) Each urn contains $n-1$ balls. The second ball is equally likely to be any of the $n(n-1)$ total balls. Of these 1/2 are black and 1/2 are red. Indeed there are $\sum_{k=0}^{n-1} k$ black balls and the same number of red balls. Furthermore:

$$\sum_{k=0}^{n-1} k = \frac{n(n-1)}{2} \quad \Rightarrow \quad P(C_1 = B) = \frac{1}{2}.$$

(b) We have to compute:

$$P(C_2 = B \mid C_1 = B) = \frac{P(C_1 = C_2 = B)}{P(C_1 = B)}.$$

The numerator can be computed by conditioning on the urn chosen:

$$P(C_1 = C_2 = B) = \sum_{k=1}^{n} P(C_1 = C_2 = B \mid k)\, P(k)$$
$$= \sum_{k=1}^{n} \left(\frac{n-k}{n-1}\right)\left(\frac{n-k-1}{n-2}\right)\frac{1}{n}.$$

2.7.2 Independence

Recall from **ST104b Statistics 2** that events A and B are independent if the probability that A occurs is unaffected by whether or not B has occurred. Assuming $P(B) > 0$, we could write this as $P(A \mid B) = P(A)$.

Independence

Consider a probability space $(\Omega, \mathcal{F}, P)$ with $A, B \in \mathcal{F}$. Events A and B are said to be **independent**, denoted $A \perp\!\!\!\perp B$, if and only if:

$$P(A \cap B) = P(A)\,P(B).$$

Also, remember not to confuse *mutually exclusive* events (for which $P(A \cap B) = 0$) and *independent* events (for which $P(A \cap B) = P(A)\,P(B)$). Mutually exclusive events will not be independent unless $P(A) = 0$ or $P(B) = 0$, or both.

Consider the case where the sample space, Ω, is made up of a finite number of equally likely outcomes. Assuming $P(B) > 0$, if A is independent of B then $P(A \mid B) = P(A)$ and so:

$$\frac{n(A \cap B)}{n(B)} = \frac{n(A)}{n(\Omega)}.$$

This means that if A and B are independent then the *proportion* of outcomes in B which are also in A is equal to the *proportion* of outcomes in Ω (that is all outcomes) which are also in A.

Basic properties of independent events

i. If $A \perp\!\!\!\perp B$, then $B \perp\!\!\!\perp A$.

ii. If $P(A) > 0$, then $P(B \mid A) = P(B) \Leftrightarrow A \perp\!\!\!\perp B$. If $P(B) > 0$, then $P(A \mid B) = P(A) \Leftrightarrow A \perp\!\!\!\perp B$.

iii. If $A \perp\!\!\!\perp B$, then $A^c \perp\!\!\!\perp B^c$, $A^c \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp B^c$.

For more than two events we define independence recursively.

Mutual independence

Consider a probability space $(\Omega, \mathcal{F}, P)$ and a set of events $\{A_1, A_2, \ldots, A_n\}$. We say that $\{A_1, A_2, \ldots, A_n\}$ are **mutually independent** if every subset of two or more elements is mutually independent and:

$$P\left(\bigcap_{i=1}^{k} A_i\right) = \prod_{i=1}^{k} P(A_i).$$

Example 2.24 Suppose that the event A is independent of itself. What value(s) can $P(A)$ take? Explain your answer.

Solution

If A is independent of itself, then $P(A) = P(A \cap A) = (P(A))^2$, hence $P(A) \in \{0, 1\}$.

Example 2.25 For two events A and B, show that if $P(B) = 1$ then A and B are independent.

Solution

If $P(B) = 1$, it follows that:

$$\begin{aligned} P(A \cap B) &= P(A) + P(B) - P(A \cup B) \\ &= P(A) + 1 - 1 \\ &= P(A) \end{aligned}$$

since $1 \geq P(A \cup B) \geq P(B) = 1$. If follows that:

$$P(A \cap B) = P(A) = P(A)\, P(B).$$

Independence is often inferred from the physics of an experiment. An assumption of independence can be used to construct an intuitively appealing probability measure in situations where we do not have a sample space of equally likely outcomes.

Example 2.26 We consider two simple examples of independence.

1. *Rolling two fair dice.* When rolling two fair dice, we assume the dice are rolled in such a way that the value shown on the first die does not have any impact on the value shown on the second. Events which are purely associated with the roll of die 1 will be independent of those purely associated with the roll of die 2. For example, let A denote the event that the value on the first die is odd and B denote the event that the value of the second die is larger than 4. Immediately, $P(A) = 1/2$, $P(B) = 1/3$ and $P(A \cap B) = 1/6$, hence A and B are independent.

2. *Flipping a coin three times.* Suppose we flip a coin three times in succession where the probability of a head on any given flip is π, and hence the probability of a tail is $(1 - \pi)$. If we assume independence of the flips then we use multiplication to determine probabilities of particular outcomes and hence probabilities associated with events. For example, if we are interested in the number of heads in the sequence of three tosses, then:

$$P(\#\text{ heads} = 0) = P(\{TTT\}) = (1 - \pi) \times (1 - \pi) \times (1 - \pi) = (1 - \pi)^3$$

also:

$$\begin{aligned} P(\#\text{ heads} = 1) &= P(\{TTH\} \cup \{THT\} \cup \{HTT\}) \\ &= P(\{TTH\}) + P(\{THT\}) + P(\{HTT\}) \\ &= (1-\pi) \times (1-\pi) \times \pi + (1-\pi) \times \pi \times (1-\pi) \\ &\quad + \pi \times (1-\pi) \times (1-\pi) \\ &= 3\pi(1-\pi)^2 \end{aligned}$$

and:

$$\begin{aligned} P(\#\text{ heads} = 2) &= P(\{THH\} \cup \{HTH\} \cup \{HHT\}) \\ &= P(\{THH\}) + P(\{HTH\}) + P(\{HHT\}) \\ &= (1-\pi) \times \pi \times \pi + \pi \times (1-\pi) \times \pi + \pi \times \pi \times (1-\pi) \\ &= 3\pi^2(1-\pi). \end{aligned}$$

Finally:

$$P(\#\text{ heads} = 3) = P(\{HHH\}) = \pi \times \pi \times \pi = \pi^3.$$

Each one of these probabilities is a term in the expansion of $(\pi + (1-\pi))^3$.

In the second case of Example 2.26 we are not really interested in the outcomes themselves, rather we are interested in a *property* of the outcomes, i.e. the number of heads. Quantities which associate real numbers with outcomes are referred to as **random variables**, discussed in the next chapter.

Activity 2.9 Consider events A, B and C where $A \subseteq C$, $B \subseteq C$ and $P(B) > 0$. Show that the relative probabilities of A and B are unchanged by conditioning on C, that is:

$$\frac{P(A \,|\, C)}{P(B \,|\, C)} = \frac{P(A)}{P(B)}.$$

Activity 2.10 Prove the chain rule of conditional probabilities using the definition of conditional probability. Take care to make sure any events which are being conditioned on have positive probability.

Activity 2.11 Consider a set of three events $\{A, B, C\}$. In order to establish independence we need to check that every subset of two or more elements is independent. This reduces to three conditions:

$$P(A \cap B) = P(A)\,P(B), \quad P(A \cap C) = P(A)\,P(C) \quad \text{and} \quad P(B \cap C) = P(B)\,P(C).$$

With the additional condition that $P(A \cap B \cap C) = P(A)\,P(B)\,P(C)$, we have a total of four conditions to check. Show that in order to establish that a set of four events is independent we need to check eleven conditions. Generalise to a set of n events.

Activity 2.12 Consider a probability space $(\Omega, \mathcal{F}, P)$ with $A, B \in \mathcal{F}$. We define the conditional probability of A given B as $P(A \mid B) = P(A \cap B)/P(B)$, provided $P(B) > 0$.

(a) Show that $P(\cdot \mid B)$ defines a probability measure on $(\Omega, \mathcal{F})$.

(b) Define $\mathcal{B} = \{A \cap B \mid A \in \mathcal{F}\}$. Show that $\mathcal{B}$ is a σ-algebra in B. Hence show that, if $m(C) = P(C)/P(B)$ for all $C \in \mathcal{B}$, then m is a probability measure on $\{B, \mathcal{B}\}$.

Activity 2.13 Consider the mutually exclusive and collectively exhaustive events A_0, A_1 and A_2.

(a) Is it possible to have $P(A_0 \cup A_1) = 1/2$, $P(A_1 \cup A_2) = 1/2$ and $P(A_2 \cup A_0) = 2/3$?

(b) Suppose:

$$P(A_0 \cup A_1) = \pi_0, \quad P(A_1 \cup A_2) = \pi_1 \quad \text{and} \quad P(A_2 \cup A_0) = \pi_2.$$

What condition on π_0, π_1 and π_2 must hold? Now generalise to n mutually exclusive and collectively exhaustive events $A_0, \ldots, A_{n-1}$, where:

$$P\left(\bigcup_{i=r}^{r+k-1} A_{i(\text{mod } n)}\right) = \pi_r$$

for $r = 0, 1, 2, \ldots, n-1$ and $0 < k < n$. What condition on $\pi_0, \pi_1, \pi_2, \ldots, \pi_{n-1}$ must hold? Note that $i(\text{mod } n)$ is the remainder when i is divided by n.

Activity 2.14 Let $\Omega = \{abc, acb, cab, cba, bca, bac, aaa, bbb, ccc\}$ and let each element of Ω have equal probability. Let A_k be the event that the kth letter in the outcome is a. Find $P(A_1 \mid A_2)$ and $P(A_1 \mid A_2^c)$. Show that the events $\{A_1, A_2, A_3\}$ are pairwise independent and determine whether they are mutually independent.

Activity 2.15 Suppose that the sexes of all children in a family are independent and that boys and girls are equally likely, i.e. both have probability 0.5.

(a) For families of three children calculate the probabilities of the events A, B and $A \cap B$ where $A =$ 'there are children of both sexes' and $B =$ 'not more than one child is a girl'.

(b) Repeat (a) for families of four children.

(c) Are A and B independent events in (a) and (b)?

(d) Let the probability, π_n, that a family has exactly n children be $(1-\pi)\pi^n$ for $n = 0, 1, 2, \ldots$, with π such that $0 < \pi < 1$. Show that the probability that a family contains exactly k boys is given by:

$$\frac{2(1-\pi)\pi^k}{(2-\pi)^{k+1}}.$$

(You might like to try for $k = 0$ and $k = 1$ before attempting for general k. You will need to use a form of the negative binomial expansion for general k.)

2.8 A reminder of your learning outcomes

On completion of this chapter, you should be able to:

- calculate probabilities for simple situations by counting
- provide the definition of a σ-algebra
- derive the properties of probability measure from axioms
- calculate probabilities for problems with large spaces using combinations and permutations
- explain the association between the number of combinations and multinomial coefficients
- derive and demonstrate conditional probability
- prove conditional probability defines a valid probability measure
- apply Bayes' theorem to solve problems involving reverse conditioning
- define independent events and exploit independence to calculate probabilities.

2.9 Sample examination questions

Solutions can be found in Appendix C.

1. You are given three events A, B and C with:

$$P(A \mid B^c \cup C) = 0.4, \quad P(B \mid C^c) = 0.3 \quad \text{and} \quad P(C) = 0.1.$$

   (a) Find $P(A \cap (B^c \cup C))$. You can use the identity $(A \cup B)^c = A^c \cap B^c$ for any events A and B without proof.

   (b) Show that for any events A_1, A_2 and A_3 and $P(A_1) > 0$, we have:

$$P(A_2 \cup A_3 \mid A_1) = P(A_2 \mid A_1) + P(A_3 \mid A_1) - P(A_2 \cap A_3 \mid A_1).$$

   You can use the identity $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any events A and B without proof.

   (c) For the same events A, B and C as in part (a), suppose you are also told that

$$P(B \mid A) = 0.4, \quad P(C \mid A) = 0.2 \quad \text{and} \quad P(A \cup C) = 0.5.$$

   Find $P(A)$ and $P(B \cap C^c \mid A)$. Hence show that $P(B \cup C^c \mid A) = 0.784$.

2. Consider the probability space $(\Omega, \mathcal{F}, P)$ and events A and B. Recall that A^c is the event such that $A \cap A^c = \emptyset$ and $A \cup A^c = \Omega$. By partitioning appropriate sets into disjoint subsets, carefully prove:

   (a) $P(A^c) = 1 - P(A)$

   (b) $P(A^c \cap B) = P(B) = P(A \cap B)$

   (c) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

   (d) $P((A \cup B) \cap (A^c \cup B^c)) = P(A) + P(B) - 2 \times P(A \cap B)$.

   How would you interpret the event $(A \cup B) \cap (A^c \cup B^c)$?

3. A fair coin is tossed repeatedly and it is decided to stop tossing the coin as soon as three heads have been obtained. Show that the probability, π_n, that exactly n tosses are required is given by:

$$\pi_n = (n-1)(n-2)2^{-n-1}.$$

   Hence show that:

$$\sum_{n=3}^{\infty} (n-1)(n-2)2^{-n} = 2.$$

Chapter 3
Random variables and univariate distributions

3.1 Recommended reading

Casella, G. and R.L. Berger, *Statistical Inference*. Chapter 1, Sections 1.4–1.6; Chapter 2, Sections 2.1–2.3; Chapter 3, Sections 3.1–3.3; Chapter 5, Section 5.5.

3.2 Learning outcomes

On completion of this chapter, you should be able to:

- provide both formal and informal definitions of a random variable
- formulate problems in terms of random variables
- explain the characteristics of distribution functions
- explain the distinction between discrete and continuous random variables
- provide the probability mass function (pmf) and support for some common discrete distributions
- provide the probability density function (pdf) and support for some common continuous distributions
- explain whether a function defines a valid mass or density
- calculate moments for discrete and continuous distributions
- prove and manipulate inequalities involving the expectation operator
- derive moment generating functions for discrete and continuous distributions
- calculate moments from a moment generating function
- calculate cumulants from a cumulant generating function
- determine the distribution of a function of a random variable
- summarise scale/location and probability integral transformations.

3.3 Introduction

This chapter extends the discussion of (univariate) random variables and common distributions of random variables introduced in **ST104b Statistics 2**, so it is advisable to review that material first before proceeding. Here, we continue with the study of random variables (which, recall, associate real numbers – the sample space – with experimental outcomes), probability distributions (which, recall, associate probability with the sample space), and expectations (which, recall, provide the central tendency, or location, of a distribution). Fully appreciating random variables, and the associated statistical notation, is a core part of understanding distribution theory.

3.4 Mapping outcomes to real numbers

A random variable is used to assign real numbers to the outcomes of an experiment. The outcomes themselves may be non-numeric (in which case a *mapping* of outcomes to real numbers is required), or numeric (when no such mapping is necessary). A probability distribution then shows how likely different outcomes are.

Example 3.1 Continuing Example 2.26 (flipping a coin three times), suppose we define the random variable X to denote the number of heads. We assume the outcomes of the flips are independent (which is a reasonable assumption), with a constant probability of success (which is also reasonable, as it is the same coin). Hence X can take the values 0, 1, 2 and 3, and so is a discrete random variable such that $\Omega = \{0, 1, 2, 3\}$. For completeness, when writing out the probability (mass) function we should specify the probability of X for all real values, not just those in Ω. This is easily achieved with '0 otherwise'. Hence:

$$P(X = x) = p_X(x) = \begin{cases} (1-\pi)^3 & \text{for } x = 0 \\ 3\pi(1-\pi)^2 & \text{for } x = 1 \\ 3\pi^2(1-\pi) & \text{for } x = 2 \\ \pi^3 & \text{for } x = 3 \\ 0 & \text{otherwise.} \end{cases}$$

The cumulative distribution, $F_X(x) = P(X \leq x)$, in this case is:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ (1-\pi)^3 & \text{for } 0 \leq x < 1 \\ (1-\pi)^2(1+2\pi) & \text{for } 1 \leq x < 2 \\ 1-\pi^3 & \text{for } 2 \leq x < 3 \\ 1 & \text{for } x \geq 3. \end{cases}$$

Formally, we can define a random variable drawing on our study of probability space in Chapter 2. Recall that we define probability as a measure which maps events to the unit interval, i.e. $[0, 1]$. Hence in the expression '$P(A)$', the argument 'A' must represent

an event. So $P(X = x)$ denotes the probability of the event $\{X = x\}$, similarly $P(X \leq x)$ denotes the probability of the event $\{X \leq x\}$.

Random variable

A **random variable** is a function $X : \Omega \to \mathbb{R}$ with the property that if:

$$A_x = \{\omega \in \Omega : X(\omega) \leq x\}$$

then:

$$A_x \in \mathcal{F} \quad \text{for all } x \in \mathbb{R}.$$

Therefore, A_x is an event for every real-valued x.

The above definition tells us that random variables *map* experimental outcomes to real numbers, i.e. $X : \Omega \to \mathbb{R}$. Hence the random variable X is a function, such that if ω is an experimental outcome, then $X(\omega)$ is a real number. The remainder of the definition allows us to discuss quantities such as $P(X \leq x)$. Technically, we should write in full $P(\{\omega \in \Omega : X(\omega) \leq x\})$, but will use $P(X \leq x)$ for brevity.

3.4.1 Functions of random variables

Our interest usually extends beyond a random variable X, such that we may wish to consider *functions* of a random variable.

Example 3.2 Continuing Example 3.1, let Y denote the number of tails. Hence $Y(\omega) = 3 - X(\omega)$, for any outcome ω. More concisely, this can be written as $Y = 3 - X$, i.e. as a linear transformation of X.

Function of a random variable

If $g : \mathbb{R} \to \mathbb{R}$ is a well-behaved[1] function, $X : \Omega \to \mathbb{R}$ is a random variable and $Y = g(X)$, then Y is a random variable, $Y : \Omega \to \mathbb{R}$ with $Y(\omega) = g(X(\omega))$ for all $\omega \in \Omega$.

3.4.2 Positive random variables

Many observed phenomena relate to quantities which cannot be negative, by definition. Examples include workers' incomes, examination marks etc. When modelling such real-world phenomena we need to capture this 'stylised fact' by using random variables with sample spaces restricted to non-negative real values.

[1] All functions considered in this course are well-behaved, and so we will omit a technical discussion.

Positive random variable

A random variable X is **positive**, denoted $X \geq 0$, if it takes a positive value for every possible outcome, i.e. $X(\omega) \geq 0$ for all $\omega \in \Omega$.

Example 3.3 We extend Examples 3.1 and 3.2. When flipping a coin three times, the sample space is:

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

Let $\mathcal{F}$ be the collection of all possible subsets of Ω, i.e. $\mathcal{F} = \{0,1\}^{\Omega}$, such that any set of outcomes is an event. As the random variable X denotes the number of heads, its full specification is:

$$X(TTT) = 0$$

$$X(TTH) = X(THT) = X(HTT) = 1$$

$$X(HHT) = X(HTH) = X(THH) = 2$$

$$\text{and} \quad X(HHH) = 3.$$

A few examples of *cumulative* probabilities are:

$$P(X \leq -1) = P(\{\omega \in \Omega : X(\omega) \leq -1\}) = P(\emptyset) = 0$$

$$P(X \leq 0) = P(\{\omega \in \Omega : X(\omega) \leq 0\}) = P(\{TTT\}) = (1-\pi)^3$$

$$P(X \leq 1) = P(\{\omega \in \Omega : X(\omega) \leq 1\}) = P(\{TTT, TTH, THT, HTT\}) = (1-\pi)^2(1+2\pi)$$

$$P(X \leq 1.5) = P(\{\omega \in \Omega : X(\omega) \leq 1.5\}) = P(\{TTT, TTH, THT, HTT\}) = (1-\pi)^2(1+2\pi)$$

$$P(X \leq 4) = P(\{\omega \in \Omega : X(\omega) \leq 4\}) = P(\Omega) = 1.$$

Note that $P(X \leq 1) = P(X \leq 1.5)$.

Similarly, as Y denotes the number of tails, we have:

$$Y(TTT) = 3 - X(TTT) = 3, \quad Y(TTH) = 3 - X(TTH) = 2 \quad \text{etc.}$$

and the following illustrative probabilities:

$$P(Y = 2) = P(3 - X = 2) = P(X = 1) = 3\pi(1-\pi)^2$$

and:

$$P(Y \leq 2) = P(X > 0) = 1 - P(X \leq 0) = 1 - (1-\pi)^3.$$

Since X and Y are simple counts, i.e. the number of heads and tails, respectively, these are positive random variables.

Activity 3.1 Two fair dice are rolled. Let X denote the absolute value of the difference between the values shown on the top face of the dice. Express each of the following in words.

(a) $\{X \leq 2\}$.

(b) $\{X = 0\}$.

(c) $P(X \leq 2)$.

(d) $P(X = 0)$.

Activity 3.2 A die is rolled and a coin is tossed. Defining the random variable X to be the value shown on the die, and the random variable Y to represent the coin outcome such that:

$$Y = \begin{cases} 1 & \text{if heads} \\ 0 & \text{if tails} \end{cases}$$

write down concise mathematical expressions for each of the following.

(a) The value on the die is less than 3.

(b) The probability that the value on the die is less than 3.

(c) The coin shows a head.

(d) The probability that the number of heads shown is less than 1.

(e) The die roll is a 6 and there are no heads.

(f) The probability that the number of heads is less than the value on the die.

3.5 Distribution functions

Although a random variable maps outcomes to real numbers, our interest usually lies in probabilities associated with the random variable. The (cumulative) distribution function fully characterises the probability distribution associated with a random variable.

Distribution function

The **distribution function**, or **cumulative distribution function (cdf)**, of a random variable X is the function $F_X : \mathbb{R} \to [0, 1]$ given by:

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

Some remarks on distribution functions are the following.

i. The terms ‘distribution function’, ‘cumulative distribution function’ and ‘cdf’ can

be used synonymously.

ii. F_X denotes the cdf of the random variable X, similarly F_Y denotes the cdf of the random variable Y etc. If an application has only one random variable, such as X, where it is unambiguous what the random variable is, we may simply write F.

We now define right continuity which is required to establish the properties of distribution functions.

Right continuity

A function $g : \mathbb{R} \to \mathbb{R}$ is **right continuous** if $g(x+) = g(x)$ for all $x \in \mathbb{R}$, where:

$$g(x+) = \lim_{h \downarrow 0} g(x+h).$$

(Note that $g(x+)$ refers to the limit of the values given by g as we approach point x from the right.[2])

We now consider properties of distribution functions. Non-examinable proofs can be found in Appendix A.

Properties of distribution functions

A distribution function F_X has the following properties.

i. F_X is a non-decreasing function, i.e. if $x < y$ then $F_X(x) \leq F_X(y)$.

ii. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.

iii. F_X is right continuous, i.e. $F_X(x+) = F_X(x)$ for all $x \in \mathbb{R}$.

Since distribution functions return cumulative probabilities, we can use them to determine probabilities of specific events of interest. Non-examinable proofs can be found in Appendix A.

Probabilities from distribution functions

For real numbers x and y, with $x < y$, we have the following.

i. $P(X > x) = 1 - F_X(x)$.

ii. $P(x < X \leq y) = F_X(y) - F_X(x)$.

iii. $P(X < x) = \lim_{h \downarrow 0} F_X(x-h) = F_X(x-)$.

iv. $P(X = x) = F_X(x) - F_X(x-)$.

[2]Similarly, we may define *left continuity* as $g(x-) = g(x)$ for all $x \in \mathbb{R}$, where $g(x-) = \lim_{h \downarrow 0} g(x-h)$.

Example 3.4 Find the cumulative distribution functions corresponding to the following density functions.

(a) Standard Cauchy:

$$f_X(x) = \frac{1}{\pi(1+x^2)} \quad \text{for } -\infty < x < \infty.$$

(b) Logistic:

$$f_X(x) = \frac{\mathrm{e}^{-x}}{(1+\mathrm{e}^{-x})^2} \quad \text{for } -\infty < x < \infty.$$

(c) Pareto:

$$f_X(x) = \frac{a-1}{(1+x)^a} \quad \text{for } 0 < x < \infty.$$

(d) Weibull:

$$f_X(x) = c\tau x^{\tau-1}\mathrm{e}^{-cx^\tau} \quad \text{for } x \geq 0,\ c > 0 \text{ and } \tau > 0.$$

Solution

(a) We have:

$$\begin{aligned} F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = \int_{-\infty}^{x} \frac{1}{\pi(1+t^2)}\,\mathrm{d}t &= \left[\frac{1}{\pi}\arctan t\right]_{-\infty}^{x} \\ &= \frac{1}{\pi}\arctan x - \frac{1}{\pi}\frac{-\pi}{2} \\ &= \frac{1}{\pi}\arctan x + \frac{1}{2}. \end{aligned}$$

(b) We have:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = \int_{-\infty}^{x} \frac{\mathrm{e}^{-t}}{(1+\mathrm{e}^{-t})^2}\,\mathrm{d}t = \left[\frac{1}{1+\mathrm{e}^{-t}}\right]_{-\infty}^{x} = \frac{1}{1+\mathrm{e}^{-x}}.$$

(c) We have:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = \int_{0}^{x} \frac{a-1}{(1+t)^a}\,\mathrm{d}t = \left[-\frac{1}{(1+y)^{a-1}}\right]_{0}^{x} = 1 - \frac{1}{(1+x)^{a-1}}.$$

For $x < 0$ it is obvious that $F_X(x) = 0$, so in full:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - \dfrac{1}{(1+x)^{a-1}} & \text{for } x \geq 0. \end{cases}$$

(d) We have:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = \int_{0}^{x} c\tau t^{\tau-1}\mathrm{e}^{-cy^\tau}\,\mathrm{d}t = \left[-\mathrm{e}^{-cy^\tau}\right]_{0}^{x} = 1 - \mathrm{e}^{-cx^\tau}.$$

For $x < 0$ it is obvious that $F_X(x) = 0$, so in full:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - \mathrm{e}^{-cx^\tau} & \text{for } x \geq 0. \end{cases}$$

Example 3.5 Suppose X is a discrete random variable with distribution function:

$$F_X(x) = \frac{x(x+1)}{42}$$

over the support $\{1, 2, \ldots, 6\}$. Determine the mass function of X, i.e. p_X.

Solution

We have:

$$p_X(x) = F_X(x) - F_X(x-1) = \frac{x(x+1)}{42} - \frac{(x-1)x}{42} = \frac{x}{21}.$$

In full:

$$p_X(x) = \begin{cases} x/21 & \text{for } x = 1, 2, \ldots, 6 \\ 0 & \text{otherwise.} \end{cases}$$

Activity 3.3 Let X be a random variable which models the value of claims received at an insurance company. Suppose that only claims greater than k are paid. Write an expression for the distribution functions of claims paid and claims not paid in terms of the distribution function of X.

While the distribution function returns $P(X \leq x)$, there are occasions when we are interested in $P(X > x)$, i.e. the probability that X is larger than x. In models of lifetime, then the event $\{X > x\}$ represents *survival* beyond time x. This gives rise to the survival function.

Survival function

If X is a random variable with distribution function F_X, the **survival function** $\bar{F}_X$ is defined as:

$$\bar{F}_X(x) = P(X > x) = 1 - F_X(x).$$

Example 3.6 If X is a non-negative continuous random variable, then $\bar{F}_X(x) = P(X > x)$ is called the *survival function*, and $h_X(x) = f_X(x)/\bar{F}_X(x)$ is the *hazard function*.

Find the hazard function for the following distributions.

(a) Pareto:

$$f_X(x) = \frac{a-1}{(1+x)^a}$$

for $0 < x < 1$ and $a > 1$.

(b) Weibull:

$$f_X(x) = c\tau x^{\tau-1} \mathrm{e}^{-cx^\tau}$$

for $x \geq 0$, $c > 0$ and $\tau > 0$.

Solution

First notice that the hazard function, often denoted by $\lambda(x)$, is (applying the chain rule):

$$\lambda(x) = -\frac{\mathrm{d}}{\mathrm{d}x} \ln \bar{F}_X(x) = -\left(\frac{-f_X(x)}{\bar{F}_X(x)}\right) = \frac{f_X(x)}{\bar{F}_X(x)}.$$

(a) For the Pareto distribution, using Example 3.4:

$$\bar{F}_X(x) = 1 - F_X(x) = \frac{1}{(1+x)^{a-1}}$$

hence:

$$\lambda(x) = -\frac{\mathrm{d}(-(a-1)\ln(1+x))}{\mathrm{d}x} = \frac{a-1}{1+x}.$$

(b) For the Weibull distribution, using Example 3.4:

$$\bar{F}_X(x) = 1 - F_X(x) = \mathrm{e}^{-cx^\tau}$$

hence:

$$\lambda(x) = -\frac{\mathrm{d}(-cx^\tau)}{\mathrm{d}x} = c\tau x^{\tau-1}.$$

Example 3.7 Show that, in general, the hazard function of a non-negative continuous random variable X does not decrease as x increases if:

$$\frac{\bar{F}_X(x+y)}{\bar{F}_X(x)}$$

does not increase as x increases, for all $y \geq 0$.

Hint: Differentiate the logarithm of the given expression with respect to x.

Solution

We are asked to show that $\lambda(x) \leq \lambda(x+y)$ when $y \geq 0$. We are told that $\bar{F}_X(x+y)/\bar{F}_X(x)$ does not increase as x increases, and so $\ln \bar{F}_X(x+y) - \ln \bar{F}_X(x)$ has a non-positive derivative with respect to x. Differentiating with respect to x, we have:

$$-\lambda(x+y) + \lambda(x) \leq 0$$

which is the required result.

3.6 Discrete vs. continuous random variables

As in **ST104b Statistics 2**, we focus on two classes of random variables and their associated probability distributions:

- *discrete* random variables – typically assumed for variables which we can *count*

- *continuous* random variables – typically assumed for variables which we can *measure*.

Example 3.8 We consider some examples to demonstrate the distinction between discreteness and continuity.

1. **Discrete model:** Any real-world variable which can be counted, taking natural numbers 0, 1, 2, ..., is a candidate for being modelled with a discrete random variable. Examples include the number of children in a household, the number of passengers on a flight etc.

2. **Continuous model:** Any real-world variables which can be measured on a continuous scale would be a candidate for being modelled with a continuous random variable. In practice, measurement is limited by the accuracy of the measuring device (for example, think how accurately you can read off a ruler). Examples include height and weight of people, duration of a flight etc.

3. **Continuous model for a discrete situation:** Consider the value of claims received at an insurance company. All values will be monetary amounts, which can be expressed in terms of the smallest unit of currency. For example, in the UK, a claim could be £100 (equivalently 10,000 pence, since £1 = 100 pence) or £10,000 (equivalently 1,000,000 pence). Since pence are the smallest unit of the currency, the value of claims must be a positive integer number of pence. Hence the value of claims is a discrete variable. However, due to the (very) large number of distinct possible values, we may consider using a continuous random variable as an *approximating model* for the value of claims.

4. **Neither discrete nor continuous model:** A random variable can also be a *mixture* of discrete and continuous parts. For example, consider the value of payments which an insurance company needs to make on *all* insurance policies of a particular type. Most policies result in no claims, so the payment for them is 0. For those policies which do result in a claim, the size of each claim is some number greater than 0. The resulting model has some discrete characteristics (a policy either results in a claim or it does not) and some continuous characteristics (treating the value of the claim as being measured on a continuous scale). Therefore, we may choose to model this situation using a random variable which is neither discrete nor continuous, i.e. with a *mixture* distribution.

In **ST104b Statistics 2**, the possible values which a random variable could take was referred to as the *sample space*, which for many distributions is a subset of $\mathbb{R}$, the real line. In this course, we will use the term *support*.

Support of a function

The **support** of a positive real-valued function, f, is the subset of the real line where f takes values strictly greater than zero:

$$\{x \in \mathbb{R} : f(x) > 0\}.$$

3.7 Discrete random variables

Discrete random variables have supports which are in some countable subset $\{x_1, x_2, \ldots\}$ of $\mathbb{R}$. This means that the probability that a discrete random variable X can take a value outside $\{x_1, x_2, \ldots\}$ is zero, i.e. $P(X = x) = 0$ for $x \notin \{x_1, x_2, \ldots\}$.

Probability distributions of discrete random variables can, of course, be represented by distribution functions, but we may also consider probability at a point using the probability mass function (pmf). We have the following definition, along with the properties of a pmf.

Probability mass function

The **probability mass function** of a discrete random variable X is the function $p_X : \mathbb{R} \to [0, 1]$ given by:

$$p_X(x) = P(X = x).$$

For brevity, we may refer to these simply as 'mass functions'. If p_X is a mass function then:

i. $0 \leq p_X(x) \leq 1$ for all $x \in \mathbb{R}$

ii. $p_X(x) = 0$ for $x \notin \{x_1, x_2, \ldots\}$

iii. $\sum_x p_X(x) = 1$.

Since the support of a mass function is the countable set $\{x_1, x_2, \ldots\}$, in summations such as in property iii. above, where the limits of summation are not explicitly provided, the sum will be assumed to be over the support of the mass function.

The probability distribution of a discrete random variable may be represented either by its distribution function or its mass function. Unsurprisingly, these two types of function are related.

Relationship between mass and distribution functions

If X is a discrete random variable such that p_X is its mass function and F_X is its distribution function, then:

i. $p_X(x) = F_X(x) - F_X(x-)$

ii. $F_X(x) = \sum_{x_i \leq x} p_X(x_i)$.

From i. above we can deduce that $F_X(x) = F_X(x-) + p_X(x)$. Since $p_X(x) = 0$ for $x \notin \{x_1, x_2, \ldots\}$, we have that $F_X(x) = F_X(x-)$ for $x \notin \{x_1, x_2, \ldots\}$. This means that the distribution function is a *step function*, i.e. flat except for discontinuities at the points $\{x_1, x_2, \ldots\}$ which represent the non-zero probabilities at the given points in the support.

We now consider some common discrete probability distributions, several of which were

first mentioned in **ST104b Statistics 2**. We refer to 'families' of probability distributions, with different members of each family distinguished by one or more *parameters*.

3.7.1 Degenerate distribution

A **degenerate distribution** concentrates all probability at a single point. If X is a degenerate random variable, its support is $\{a\}$, for some constant $-\infty < a < \infty$. Denoted $X \sim \text{Degenerate}(a)$, its mass function is:

$$p_X(x) = \begin{cases} 1 & \text{for } x = a \\ 0 & \text{otherwise} \end{cases}$$

and its distribution function is:

$$F_X(x) = \begin{cases} 0 & \text{for } x < a \\ 1 & \text{for } x \geq a. \end{cases}$$

3.7.2 Discrete uniform distribution

A **discrete uniform distribution** assigns equal (i.e. the same, or 'uniform') probabilities to each member of its support. If X is a discrete uniform random variable with support $\{x_1, x_2, \ldots, x_n\}$, denoted $X \sim \text{Uniform}\{x_1, x_n\}$, its mass function is:

$$p_X(x) = \begin{cases} \dfrac{1}{n} & \text{for } x \in \{x_1, x_2, \ldots, x_n\} \\ 0 & \text{otherwise.} \end{cases}$$

The corresponding distribution function is a step function with steps of equal magnitude $1/n$. This can be expressed as:

$$F_X(x) = \begin{cases} 0 & \text{for } x < x_1 \\ \dfrac{\lfloor x \rfloor - x_1 + 1}{n} & \text{for } x_1 \leq x \leq x_n \\ 1 & \text{for } x > x_n \end{cases}$$

where $\lfloor x \rfloor$ is the 'floor' of x, i.e. the largest integer which is less than or equal to x.

3.7.3 Bernoulli distribution

A **Bernoulli distribution** assigns probabilities π and $1 - \pi$ to the only two possible outcomes, often referred to as 'success' and 'failure', respectively, although these do not necessarily have to represent 'good' and 'bad' outcomes, respectively. Since the support must be a set of real numbers, these are assigned the values 1 and 0, respectively. If X is a Bernoulli random variable, denoted $X \sim \text{Bernoulli}(\pi)$, its mass function is:

$$p_X(x) = \begin{cases} \pi^x (1-\pi)^{1-x} & \text{for } x = 0, 1 \\ 0 & \text{otherwise.} \end{cases}$$

Sometimes the notation p is used to denote π. The corresponding distribution function is a step function, given by:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - \pi & \text{for } 0 \leq x < 1 \\ 1 & \text{for } x \geq 1. \end{cases}$$

3.7.4 Binomial distribution

A **binomial distribution** assigns probabilities to the number of successes in n independent trials each with only two possible outcomes with a constant probability of success. If X is a binomial random variable, denoted $X \sim \text{Bin}(n, \pi)$, its mass function is:

$$p_X(x) = \begin{cases} \binom{n}{x} \pi^x (1-\pi)^{n-x} & \text{for } x = 0, 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

where:

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

is the binomial coefficient. Sometimes the notation p and q are used to denote π and $1 - \pi$, respectively. This has the benefit of brevity, since q is more concise than $1 - \pi$. Proof of the validity of this mass function was shown in **ST104b Statistics 2**.

The corresponding distribution function is a step function, given by:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} \pi^i (1-\pi)^{n-i} & \text{for } 0 \leq x < n \\ 1 & \text{for } x \geq n. \end{cases}$$

Note the special case when $n = 1$ corresponds to the Bernoulli(π) distribution, and that the sum of n independent and identically distributed Bernoulli(π) random variables has a Bin(n, π) distribution.

3.7.5 Geometric distribution

Somewhat confusingly, there are two versions of the **geometric distribution** with subtle differences over the support and hence of what the random variable represents. Regardless of this nuance, both versions involve the repetition of independent and identically distributed Bernoulli trials *until the first success occurs*.

First version

In the first version of the geometric distribution, X represents the *trial number of the first success*. As such the support is $\{1, 2, \ldots\}$, with the mass function:

$$p_X(x) = \begin{cases} (1-\pi)^{x-1} \pi & \text{for } x = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

with its distribution function given by:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 1 \\ 1 - (1-\pi)^{\lfloor x \rfloor} & \text{for } x \geq 1. \end{cases}$$

Second version

In the second version of the geometric distribution, X represents the *number of failures before the first success*. As such the support is $\{0, 1, 2, \ldots\}$, with the mass function:

$$p_X(x) = \begin{cases} (1-\pi)^x \pi & \text{for } x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

with its distribution function given by:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - (1-\pi)^{\lfloor x \rfloor + 1} & \text{for } x \geq 0. \end{cases}$$

Either version could be denoted as $X \sim \text{Geo}(\pi)$, although be sure to be clear which version is being used in an application.

3.7.6 Negative binomial distribution

The **negative binomial distribution** extends the geometric distribution, and hence also has two versions.

First version

In the first version this distribution is used to represent the trial number of the rth success in independent Bernoulli(π) trials, where $r = 1, 2, \ldots$. When $r = 1$, this is a special case which is the (first) version of the geometric distribution above. If X is a negative binomial random variable, denoted $X \sim \text{Neg. Bin}(r, \pi)$, where π is the constant probability of success, its mass function is:

$$p_X(x) = \begin{cases} \binom{x-1}{r-1} \pi^r (1-\pi)^{x-r} & \text{for } x = r, r+1, r+2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

Note why the mass function has this form. In order to have r successes for the first time on the xth trial, we must have $r-1$ successes in the first $x-1$ trials. The number of ways in which these $r-1$ successes may occur is $\binom{x-1}{r-1}$ and the probability associated with each of these sequences is $\pi^{r-1}(1-\pi)^{x-r}$. Since the xth, i.e. final, trial must be a success, we then multiply $\binom{x-1}{r-1}\pi^{r-1}(1-\pi)^{x-r}$ by π. From the mass function it can be seen that the negative binomial distribution generalises the (first version of the) geometric distribution, such that if $X \sim \text{Geo}(\pi)$ then $X \sim \text{Neg. Bin}(1, \pi)$.

Its distribution function is given by:

$$F_X(x) = \begin{cases} 0 & \text{for } x < r \\ \displaystyle\sum_{i=r}^{\lfloor x \rfloor} p_X(i) & \text{for } x \geq r. \end{cases}$$

Second version

The second version of the negative binomial distribution is formulated as the number of failures before the rth success occurs. In this formulation the mass function is:

$$p_X(x) = \begin{cases} \binom{x+r-1}{r-1}\pi^r(1-\pi)^x & \text{for } x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

while its distribution function is:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \sum_{i=0}^{\lfloor x \rfloor} p_X(i) & \text{for } x \geq 0. \end{cases}$$

3.7.7 Polya distribution

The **Polya** distribution extends the second version of the negative binomial distribution to allow for non-integer values of r. If X is a Polya random variable, denoted by $X \sim \text{Polya}(r, \pi)$, its mass function is:

$$p_X(x) = \begin{cases} \dfrac{(r+x)}{x!\,\Gamma(r)}\,\pi^r(1-\pi)^x & \text{for } x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

where Γ is the **gamma function** defined as:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}\mathrm{e}^{-t}\,\mathrm{d}t.$$

Integration by parts yields a useful property of the gamma function:

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1) \quad \text{for } \alpha > 1$$

and we also have:

$$\Gamma(1) = 1. \tag{3.1}$$

One interpretation of the above property is that the gamma function extends the factorial function to non-integer values. It is clear from (3.1) that:

$$\Gamma(n) = (n-1)!$$

for any positive integer n.

3.7.8 Hypergeometric distribution

A **hypergeometric distribution** is used to represent the number of successes when n objects are selected *without* replacement from a population of N objects, where $K \leq N$ of the objects represent 'success' and the remaining $N - K$ objects represent 'failure'. If

X is a hypergeometric random variable, denoted $X \sim \text{Hyper}(n, N, K)$, with support $\{\max(0, n+K-N), \ldots, \min(n, K)\}$, its mass function is:

$$p_X(x) = \begin{cases} \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} & \text{for } x = \max(0, n+K-N), \ldots, \min(n, K) \\ 0 & \text{otherwise} \end{cases}$$

where $n \in \{0, 1, 2, \ldots, N\}$, $K \in \{0, 1, 2, \ldots, N\}$ and $N \in \{0, 1, 2, \ldots\}$ are parameters. We omit the distribution function of the hypergeometric distribution in this course.

Note this is similar to the binomial distribution except that sampling is *with replacement* in a binomial model, whereas sampling is *without replacement* in a hypergeometric model.

3.7.9 Poisson distribution

A **Poisson distribution** is used to model the number of occurrences of events over a fixed interval, typically in space or time. This distribution has a single parameter, $\lambda > 0$, and the support of the distribution is $\{0, 1, 2, \ldots\}$. If X is a Poisson random variable, denoted $X \sim \text{Pois}(\lambda)$, its mass function is:

$$p_X(x) = \begin{cases} \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} & \text{for } x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

while its distribution function is:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \mathrm{e}^{-\lambda} \sum_{i=0}^{\lfloor x \rfloor} \frac{\lambda^i}{i!} & \text{for } x \geq 0. \end{cases}$$

Example 3.9 Suppose a fair die is rolled 10 times. Let X be the random variable to represent the number of 6s which appear. Derive the distribution of X, i.e. F_X.

Solution

For a fair die, the probability of a 6 is 1/6, and hence the probability of a non-6 is 5/6. By independence of the outcome of each roll, we have:

$$F_X(x) = P(X \leq x) = \sum_{i=0}^{x} P(X = i) = \sum_{i=0}^{x} \binom{10}{i} \left(\frac{1}{6}\right)^i \left(\frac{5}{6}\right)^{10-i}.$$

In full:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \sum_{i=0}^{\lfloor x \rfloor} \binom{10}{i} \left(\frac{1}{6}\right)^i \left(\frac{5}{6}\right)^{10-i} & \text{for } 0 \leq x \leq 10 \\ 1 & \text{for } x > 10. \end{cases}$$

Example 3.10 Show that the ratio of any two *successive* hypergeometric probabilities, i.e. $P(X = x + 1)$ and $P(X = x)$ equals:

$$\frac{n-x}{x+1}\frac{K-x}{N-K-n+x+1}$$

for any valid x and $x + 1$.

Solution

If $X \sim \text{Hyper}(n, N, K)$, then from its mass function we have:

$$\begin{aligned}
&\frac{p_X(x+1)}{p_X(x)} \\
&= \frac{\binom{K}{x+1}\binom{N-K}{n-x-1}}{\binom{N}{n}} \Bigg/ \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \\
&= \frac{\binom{K}{x+1}\binom{N-K}{n-x-1}}{\binom{K}{x}\binom{N-K}{n-x}} \\
&= \frac{K!}{(x+1)!\,(K-x-1)!}\,\frac{(N-K)!}{(n-x-1)!\,(N-K-n+x+1)!} \\
&\quad \times \frac{x!\,(K-x)!}{K!}\,\frac{(n-x)!\,(N-K-n+x)!}{(N-K)!} \\
&= \frac{n-x}{x+1}\frac{K-x}{N-K-n+x+1}.
\end{aligned}$$

Example 3.11 Consider a generalisation of the hypergeometric distribution, such that in a population of N objects, N_1 are of type 1, N_2 are of type 2, ... and N_k are of type k, where:

$$\sum_{i=1}^{k} N_i = N.$$

Derive an expression for the probability of $n_1, n_2, \ldots, n_k$ of the first, second etc. up to the kth type of object, when a random sample of size n is selected without replacement.

Solution

In total there are $\binom{N}{n}$ ways of selecting a random sample of size n from the population of N objects. There are $\binom{N_i}{n_i}$ ways to arrange the n_i of the N_i objects for $i = 1, 2, \ldots, k$. By the rule of product, the required probability is:

$$\frac{\prod_{i=1}^{k}\binom{N_i}{n_i}}{\binom{N}{n}}.$$

Activity 3.4 Consider the first version of the geometric distribution.

(a) Show that its mass function:

$$p_X(x) = \begin{cases} (1-\pi)^{x-1}\pi & \text{for } x = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

is a valid mass function.

(b) Use the mass function in (a) to derive the associated distribution function.

Activity 3.5 Consider the first version of the negative binomial distribution. Show that its mass function:

$$p_X(x) = \begin{cases} \binom{x-1}{r-1}\pi^r(1-\pi)^{x-r} & \text{for } x = r, r+1, r+2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

is a valid mass function.

3.8 Continuous random variables

Continuous random variables have supports which are either the real numbers $\mathbb{R}$, or one or more intervals in $\mathbb{R}$. Equivalently, this means that the distribution function of a continuous random variable is continuous (unlike the step functions for discrete random variables). Instead of a (probability) *mass* function, we describe a continuous distribution with a (probability) *density* function.

Continuous random variable

A random variable X is **continuous** if its distribution function can be expressed as:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \quad \text{for } x \in \mathbb{R}$$

for some integrable function $f_X : \mathbb{R} \to [0, \infty)$, known as the (probability) density function. In reverse, the density function can be derived from the distribution function by differentiating:

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}t} F_X(t)\Big|_{t=x} = F_X'(x) \quad \text{for all } x \in \mathbb{R}.$$

If f_X is a valid density function, then:

i. $f_X(x) \geq 0$ for all $x \in \mathbb{R}$

ii. $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$.

Example 3.12 Show that:

$$f_X(x) = \begin{cases} (n+2)(n+1)x^n(1-x) & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

is a valid density function, where n is a positive integer.

Solution

Immediately, $f_X(x) \geq 0$ for all real x. So it remains to check that the function integrates to 1 over its support. We have:

$$\begin{aligned} f_X(x) = \int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x &= \int_0^1 (n+2)(n+1)x^n(1-x)\,\mathrm{d}x \\ &= \int_0^1 (n+2)(n+1)(x^n - x^{n+1})\,\mathrm{d}x \\ &= \left[(n+2)(n+1)\left(\frac{x^{n+1}}{n+1} - \frac{x^{n+2}}{n+2}\right)\right]_0^1 \\ &= \Big[(n+2)x^{n+1} - (n+1)x^{n+2}\Big]_0^1 \\ &= (n+2) - (n+1) \\ &= 1. \end{aligned}$$

Hence f_X is a valid density function.

Example 3.13 Let X be a random variable with density function:

$$f_X(x) = \begin{cases} x\mathrm{e}^{-x} & \text{for } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Derive the distribution function of X, i.e. F_X.

Solution

Applying integration by parts, we have:

$$\begin{aligned} F_X(x) = \int_0^x t\mathrm{e}^{-t}\,\mathrm{d}t &= \Big[-t\mathrm{e}^{-t}\Big]_0^x + \int_0^1 \mathrm{e}^{-t}\,\mathrm{d}t \\ &= -x\mathrm{e}^{-x} + \Big[-\mathrm{e}^{-t}\Big]_0^x \\ &= -x\mathrm{e}^{-x} - \mathrm{e}^{-x} + 1 \\ &= 1 - (1+x)\mathrm{e}^{-x}. \end{aligned}$$

In full:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - (1+x)\mathrm{e}^{-x} & \text{for } x \geq 0. \end{cases}$$

Example 3.14 A *logistic curve* has distribution function:

$$F_X(x) = \frac{1}{1+\mathrm{e}^{-x}} \quad \text{for } -\infty < x < \infty.$$

(a) Verify this is a valid distribution function.

(b) Derive the corresponding density function.

Solution

(a) Applying the chain rule, we have that:

$$F'_X(x) = \frac{\mathrm{e}^{-x}}{(1+\mathrm{e}^{-x})^2} > 0$$

and so F_X is strictly increasing. We also have that:

$$\lim_{x\to-\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} F_X(x) = 1$$

verifying this is a valid distribution function.

(b) The density function is simply $F'_X(x)$, i.e. we have:

$$f_X(x) = \frac{\mathrm{e}^{-x}}{(1+\mathrm{e}^{-x})^2} \quad \text{for } -\infty < x < \infty.$$

Note this is a member of *logistic distribution* family.

Example 3.15 Let X be a continuous random variable with density function $f_X(x)$ which is symmetric, i.e. $f_X(x) = f_X(-x)$ for all x. For any real constant k, show that:

$$P(-k < X < k) = 2F_X(k) - 1.$$

Solution

We have:

$$\begin{aligned} P(-k < X < k) &= P(-k < X \le 0) + P(0 < X < k) \\ &= \int_{-k}^{0} f_X(x)\,\mathrm{d}x + \int_{0}^{k} f_X(x)\,\mathrm{d}x \\ &= -\int_{k}^{0} f_X(-x)\,\mathrm{d}x + \int_{0}^{k} f_X(x)\,\mathrm{d}x \\ &= \int_{0}^{k} f_X(x)\,\mathrm{d}x + \int_{0}^{k} f_X(x)\,\mathrm{d}x \\ &= 2(F_X(k) - F_X(0)). \end{aligned}$$

Due to the symmetry of X, we have that $F_X(0) = 0.5$. Therefore:

$$2(F_X(k) - F_X(0)) = 2(F_X(k) - 0.5) = 2F_X(k) - 1.$$

It is important to remember that whereas mass functions return probabilities, since $p_X(x) = P(X = x)$, hence values of mass functions must be within $[0, 1]$, values of a density fucntion are *not* probabilities, rather probabilities are given by the area below the density function (and above the x-axis).

Probability of an event for a continuous random variable

If X is a continuous random variable with density function f_X, then for $a, b \in \mathbb{R}$ such that $a \leq b$, we have:

$$P(a < X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x = F_X(b) - F_X(a).$$

Setting $a = b = x$, we have that:

$$P(X = x) = 0 \quad \text{for all } x \in \mathbb{R}.$$

More generally, for any well-behaved subset A of $\mathbb{R}$, i.e. $A \subseteq \mathbb{R}$, then:

$$P(X \in A) = \int_A f_X(x)\,\mathrm{d}x$$

where $\{X \in A\}$ means, in full, $\{\omega \in \Omega : X(\omega) \in A\}$ such that A is an interval or a countable union of intervals.

Note the seemingly counterintuitive result that $P(X = x) = 0$. This seems strange because we can observe real-valued measurements of a continuous variable, such as height. If a person was measured to be 170 cm, what does it mean if we said $P(X = 170) = 0$, when clearly we have observed this event? Well, 170 cm has been expressed to the nearest centimetre, so this simply means the observed height (in centimetres) fell in the interval $[169.5, 170.5]$, and there is a strictly positive probability associated with this interval. Even if we used a measuring device with (far) greater accuracy, there will always be practical limitations to the precision with which we can measure. Therefore, any measurement on a continuous scale produces an *interval* rather than a *single value*.

We now consider some common continuous probability distributions, several of which were first mentioned in **ST104b Statistics 2**. As with discrete distributions, we refer to 'families' of probability distributions, with different members of each family distinguished by one or more *parameters*.

3.8.1 Continuous uniform distribution

A **continuous uniform distribution** assigns probability equally (*uniformly*, hence the name) over its support $\{[a, b]\}$, for $a < b$. If X is a continuous random variable, denoted

$X \sim \text{Uniform}[a, b]$, its density function is:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

Its distribution function is given by:

$$F_X(x) = \begin{cases} 0 & \text{for } x < a \\ \dfrac{x-a}{b-a} & \text{for } a \leq x \leq b \\ 1 & \text{for } x > b. \end{cases}$$

The special case of $X \sim \text{Uniform}[0, 1]$, i.e. when the support is the unit interval, is used in simulations of random samples from distributions, by treating a random drawing from Uniform$[0, 1]$ as a randomly drawn value of a distribution function, since:

$$\lim_{x \to -\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F_X(x) = 1.$$

Inverting the distribution function recovers the (simulated) random drawing from the desired distribution.

3.8.2 Exponential distribution

An **exponential distribution** arises in reliability theory and queuing theory. For example, in queuing theory we can model the distribution of *interarrival* times (if, as is often assumed, arrivals are treated as having a Poisson distribution with a rate parameter of $\lambda > 0$) with a positive-valued random variable following an exponential distribution, i.e. with support $\{x \geq 0\}$. If X is an exponential random variable, denoted $X \sim \text{Exp}(\lambda)$, its density function is:

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

Its distribution function is given by:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - \mathrm{e}^{-\lambda x} & \text{for } x \geq 0. \end{cases}$$

3.8.3 Normal distribution

A **normal distribution** is (one of) the most important distribution(s) in statistics. It has been covered extensively in **ST104a Statistics 1** and **ST104b Statistics 2**. Recall that a normal distribution is completely specified by its mean, μ, and its variance, σ^2 (such that $-\infty < \mu < \infty$ and $\sigma^2 > 0$) and has a support of $\mathbb{R}$. If X is a normal random variable, denoted $X \sim N(\mu, \sigma^2)$, its density function is:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for } -\infty < x < \infty.$$

The distribution function of a normal random variable does not have a closed form.

An important special case is the **standard normal distribution** with $\mu = 0$ and $\sigma^2 = 1$, denoted $Z \sim N(0, 1)$, with density function:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad \text{for } -\infty < z < \infty.$$

X and Z are related through the linear transformation:

$$Z = \frac{X - \mu}{\sigma} \quad \Leftrightarrow \quad X = \mu + \sigma Z.$$

The distribution function of Z is denoted by Φ, such that if $Z \sim N(0, 1)$, then:

$$\Phi(z) = F_Z(z) = P(Z \leq z).$$

3.8.4 Gamma distribution

A **gamma distribution** is a positively-skewed distribution with numerous practical applications, such as modelling the size of insurance claims and the size of defaults on loans. The distribution is characterised by two parameters – a shape parameter, $\alpha > 0$, and a scale parameter, $\beta > 0$. If X is a gamma-distributed random variable, denoted $X \sim \text{Gamma}(\alpha, \beta)$, its density function is:

$$f_X(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} \mathrm{e}^{-\beta x} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where recall Γ is the gamma function, defined in Section 3.7.7. We omit the distribution function of the gamma distribution in this course.

Note that when $\alpha = 1$ the density function reduces to an exponential distribution, i.e. if $X \sim \text{Gamma}(1, \beta)$, then $X \sim \text{Exp}(\beta)$.

3.8.5 Beta distribution

A **beta distribution** is a generalisation of the continuous uniform distribution, defined over the support $[0, 1]$. A beta distribution is characterised by two shape parameters, $\alpha > 0$ and $\beta > 0$. If X is a beta random variable, denoted $X \sim \text{Beta}(\alpha, \beta)$, its density function is:

$$f_X(x) = \begin{cases} \dfrac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

where $B(\alpha, \beta)$ is the **beta function** defined as:

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} \, \mathrm{d}t = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

We omit the distribution function of the beta distribution in this course.

Note that when $\alpha = 1$ and $\beta = 1$ the density function reduces to the continuous uniform distribution over $[0, 1]$, i.e. if $X \sim \text{Beta}(1, 1)$, then $X \sim \text{Uniform}[0, 1]$.

3.8.6 Triangular distribution

A **triangular distribution** is a popular choice of input distribution in Monte Carlo simulation studies. It is specified by easy-to-understand parameters: the minimum possible value, a, the maximum possible value, b (with $a < b$), and the modal (i.e. most likely) value, c, such that $a \leq c \leq b$. The support is $[a, b]$. If X is a triangular random variable, denoted $X \sim \text{Triangular}(a, b, c)$, its density function is:

$$f_X(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)} & \text{for } a \leq x < c \\ \dfrac{2}{b-a} & \text{for } x = c \\ \dfrac{2(b-x)}{(b-a)(b-c)} & \text{for } c < x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

Its distribution function is given by:

$$F_X(x) = \begin{cases} 0 & \text{for } x < a \\ \dfrac{(x-a)^2}{(b-a)(c-a)} & \text{for } a \leq x \leq c \\ 1 - \dfrac{(b-x)^2}{(b-a)(b-c)} & \text{for } c < x \leq b \\ 1 & \text{for } x > b. \end{cases}$$

Example 3.16 A random variable has 'no memory' if for all x and for $y > 0$ it holds that:

$$P(X > x + y \,|\, X > x) = P(X > y).$$

Show that if X has either the exponential distribution, or a geometric distribution with $P(X = x) = q^{x-1}p$, then X has no memory. Interpret this property.

Solution

We must check that $P(X > x + y \,|\, X > x) = P(X > y)$. This can be written in terms of the distribution function of X because for $y > 0$ we have:

$$\begin{aligned} 1 - F_X(y) = P(X > y) = P(X > x + y \,|\, X > x) &= \frac{P(\{X > x + y\} \cap \{X > x\})}{P(X > x)} \\ &= \frac{P(X > x + y)}{P(X > x)} \\ &= \frac{1 - F_X(x + y)}{1 - F_X(x)}. \end{aligned}$$

If $X \sim \text{Exp}(\lambda)$, then:

$$1 - F_X(x) = \mathrm{e}^{-\lambda x}.$$

The 'no memory' property is verified by noting that:

$$\mathrm{e}^{-\lambda y} = \frac{\mathrm{e}^{-\lambda(x+y)}}{\mathrm{e}^{-\lambda x}}.$$

If X has a geometric distribution, then:

$$1 - F_X(x) = q^x.$$

The 'no memory' property is verified by noting that:

$$q^y = \frac{q^{x+y}}{q^x}.$$

The 'no memory' property is saying that 'old is as good as new'. If we think in terms of lifetimes, it says that you are equally likely to survive for y more years whatever your current age x may be. This is unrealistic for humans for widely different ages x, but may work as a base model in other applications.

3.9 Expectation, variance and higher moments

3.9.1 Mean of a random variable

Measures of central tendency (or location) were introduced as descriptive statistics in **ST104a Statistics 1**. For simple datasets, the mean, median and mode were considered. Central tendency allows us to summarise a single feature of datasets as a 'typical' value. This is a simple example of data reduction, i.e. reducing a random sample of $n > 1$ observations into a single value, in this instance to reflect where a sample distribution is centred.

As seen in **ST104b Statistics 2**, central tendency measures can also be applied to probability distributions. For example, if X is a continuous random variable with density function f_X and distribution function F_X, then the mode is:

$$\text{mode}(X) = \max_x f_X(x)$$

i.e. the value of X where the density function reaches a maximum (which may or may not be unique), and the median is the value m satisfying:

$$\text{median}(X) = F_X(m) = 0.5.$$

Hereafter, we will focus our attention on the **mean** of X, often referred to as the **expected value** of X, or simply the **expectation** of X.

Mean of a random variable

If X is a random variable with mean μ, then:

$$\mu = \mathrm{E}(X) = \begin{cases} \sum_x x\, p_X(x) & \text{for discrete } X \\ \int_{-\infty}^{\infty} x\, f_X(x)\, \mathrm{d}x & \text{for continuous } X \end{cases} \tag{3.2}$$

where to ensure that $\mathrm{E}(X)$ is well-defined, we usually require that $\sum_x |x|\, p_X(x) < \infty$ for discrete X, and that $\int_{-\infty}^{\infty} |x|\, f_X(x)\, \mathrm{d}x < \infty$ for continuous X.

Example 3.17 Suppose that $X \sim \text{Bin}(n, \pi)$. Hence its mass function is:

$$p_X(x) = \begin{cases} \binom{n}{x}\pi^x(1-\pi)^{n-x} & \text{for } x = 0, 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

and so the mean, $\text{E}(X)$, is:

$$\begin{aligned}
\text{E}(X) &= \sum_x x\, p_X(x) && \text{(by definition)} \\
&= \sum_{x=0}^{n} x \binom{n}{x} \pi^x (1-\pi)^{n-x} && \text{(substituting } p_X(x)\text{)} \\
&= \sum_{x=1}^{n} x \binom{n}{x} \pi^x (1-\pi)^{n-x} && \text{(since } x\, p_X(x) \text{ when } x = 0\text{)} \\
&= \sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)!\,[(n-1)-(x-1)]!}\, \pi\pi^{x-1}(1-\pi)^{n-x} && \text{(as } x/x! = 1/(x-1)!\text{)} \\
&= n\pi \sum_{x=1}^{n} \binom{n-1}{x-1} \pi^{x-1}(1-\pi)^{n-x} && \text{(taking } n\pi \text{ outside)} \\
&= n\pi \sum_{y=0}^{n-1} \binom{n-1}{y} \pi^{y}(1-\pi)^{(n-1)-y} && \text{(setting } y = x-1\text{)} \\
&= n\pi \times 1 && \text{(since the sum is a Bin}(n-1,\, \pi)\text{ mass function)} \\
&= n\pi.
\end{aligned}$$

Example 3.18 Suppose that $X \sim \text{Exp}(\lambda)$. Hence its density function is:

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

and so the mean, $\text{E}(X)$, is:

$$\text{E}(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, \text{d}x = \int_0^{\infty} x\, \lambda \text{e}^{-\lambda x}\, \text{d}x.$$

Note that:

$$x\, \lambda \text{e}^{-\lambda x} = -\lambda \frac{\text{d}}{\text{d}\lambda} \text{e}^{-\lambda x}.$$

We may switch the order of differentation with respect to λ and integration with respect to x, hence:

$$\text{E}(X) = -\lambda \frac{\text{d}}{\text{d}\lambda} \int_0^{\infty} \text{e}^{-\lambda x}\, \text{d}x = -\lambda \frac{\text{d}}{\text{d}\lambda} \left[-\frac{1}{\lambda} \text{e}^{-\lambda x} \right]_0^{\infty} = -\lambda \frac{\text{d}}{\text{d}\lambda} \lambda^{-1} = \lambda\lambda^{-2} = \frac{1}{\lambda}.$$

An alternative approach makes use of integration by parts.

Example 3.19 If the random variable $X \sim \text{Hyper}(n, N, K)$, show that for a random sample of size n, the expected value of the number of successes is:

$$\text{E}(X) = \frac{nK}{N}.$$

Solution

We have:

$$\text{E}(X) = \sum_{x=0}^{K} x \, \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} = \sum_{x=1}^{K} x \, \frac{\frac{K!}{x!\,(K-x)!}\binom{N-K}{n-x}}{\frac{N!}{n!\,(N-n)!}}$$

noting that $0\,p_X(0) = 0$, so we may start the summation at $x = 1$. We proceed to factor out the result of nK/N, such that:

$$\text{E}(X) = \frac{nK}{N} \sum_{x=1}^{K} \frac{\frac{(K-1)!}{(x-1)!\,(K-x)!}\binom{N-K}{n-x}}{\frac{(N-1)!}{(n-1)!\,(N-n)!}} = \frac{nK}{N} \sum_{x=1}^{K} \frac{\binom{K-1}{x-1}\binom{N-K}{n-x}}{\binom{N-1}{n-1}}.$$

If we now change the summation index to start at 0, we have:

$$\text{E}(X) = \frac{nK}{N} \sum_{x=0}^{K-1} \frac{\binom{K-1}{x}\binom{N-K}{n-1-x}}{\binom{N-1}{n-1}} = \frac{nK}{N}$$

since the summation is of $X \sim \text{Hyper}(n-1, N-1, K-1)$ over its support, and hence is equal to 1.

3.9.2 Expectation operator

ST104b Statistics 2 introduced properties of the expectation operator, E, notably 'the expectation of the sum' equals the 'sum of the expectations', i.e. *linearity* – reviewed below. We have seen above how the expectation operator is applied to determine the mean of a random variable X. On many occasions we may be interested in a function of X, i.e. $g(X)$. Rather than determine the mass or density function of $g(X)$ and then work out $\text{E}(g(X))$ using (3.2), it is often easier to work directly with the original mass or density function of X.

Expectation of functions of a random variable

For any well-behaved function $g : \mathbb{R} \to \mathbb{R}$, the expectation of $g(X)$ is defined as:

$$\text{E}(g(X)) = \begin{cases} \sum\limits_{x} g(x)\, p_X(x) & \text{for discrete } X \\ \int_{-\infty}^{\infty} g(x)\, f_X(x)\, \mathrm{d}x & \text{for continuous } X \end{cases}$$

where we usually require that $\sum_x |g(x)|\, p_X(x) < \infty$ for discrete X, and that $\int_{-\infty}^{\infty} |g(x)|\, f_X(x)\, \mathrm{d}x < \infty$ for continuous X, to ensure that $\text{E}(g(X))$ is well-defined.

A key property of the expectation is linearity.

Linearity of the expectation operator

For a random variable X and real constants $a_0, a_1, a_2, \ldots, a_k$, then:

$$\mathrm{E}\left(\sum_{i=0}^{k} a_i X^i\right) = \sum_{i=0}^{k} a_i \mathrm{E}(X^i).$$

The proof of this result is trivial since the property of linearity is inherited directly from the definition of expectation in terms of a sum or an integral. Note that since:

$$\mathrm{E}(a_0 + a_1 X + a_2 X^2 + \cdots + a_k X^k) = a_0 + a_1 \mathrm{E}(X) + a_2 \mathrm{E}(X^2) + \cdots + a_k \mathrm{E}(X^k)$$

when $k = 1$, i.e. for any real constants a_0 and a_1, we have:

$$\mathrm{E}(a_0) = a_0 \quad \text{and} \quad \mathrm{E}(a_0 + a_1 X) = a_0 + a_1 \mathrm{E}(X).$$

Also note that if X is a positive random variable, then $\mathrm{E}(X) \geq 0$, as we would expect.

Example 3.20 Suppose X is a random variable with density function:

$$f_X(x) = \begin{cases} 2(1-x) & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Define $Y = X^2$, with corresponding density function:

$$f_Y(y) \begin{cases} 1/\sqrt{y} - 1 & \text{for } 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Determine $\mathrm{E}(Y)$.

Solution

We can derive $\mathrm{E}(Y)$ in one of two ways. Working directly with $f_Y(y)$, we have:

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} y\, f_Y(y)\, \mathrm{d}y = \int_0^1 y \left(\frac{1}{\sqrt{y}} - 1\right) \mathrm{d}y = \left[\frac{2y^{3/2}}{3} - \frac{y^2}{2}\right]_0^1 = \frac{1}{6}.$$

Working with $f_X(x)$, we have:

$$\mathrm{E}(Y) = \mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\, \mathrm{d}x = \int_0^1 x^2(1-x)\, \mathrm{d}x = \left[\frac{2x^3}{3} - \frac{x^4}{2}\right]_0^1 = \frac{1}{6}.$$

Example 3.21 A bowl has n balls, numbered 1 to n. One ball is selected at random. Let X be the random variable representing the number of this ball, hence the support of X is $\{1, 2, \ldots, n\}$. Suppose the probability that ball x is chosen is kx. Calculate $\mathrm{E}(1/X)$.

Solution

We must have that:

$$\sum_x p_X(x) = \sum_{x=1}^{n} kx = k\sum_{x=1}^{n} x = \frac{kn(n+1)}{2} = 1$$

noting that $\sum_{i=1}^{n} i = n(n+1)/2$. Therefore:

$$k = \frac{2}{n(n+1)}.$$

Hence:

$$\mathrm{E}\left(\frac{1}{X}\right) = \sum_x \frac{1}{x} p_X(x) = \sum_{x=1}^{n} \frac{1}{x} \frac{2x}{n(n+1)} = \sum_{x=1}^{n} \frac{2}{n(n+1)} = \frac{2}{n+1}.$$

3.9.3 Variance of a random variable

Measures of dispersion (or spread) were similarly introduced in **ST104a Statistics 1** in the context of descriptive statistics for simple univariate datasets. As with measures of central tendency, we may apply these to probability distributions. For example, if X is a random variable with distribution function F_X, then the interquartile range (IQR) is:

$$\mathrm{IQR}(X) = Q_3 - Q_1 = F_X^{-1}(0.75) - F_X^{-1}(0.25).$$

Hereafter, we will focus our attention on the **variance** of X – the average squared distance from the mean (the **standard deviation** of X is then simply the positive square root of the variance).

Variance and standard deviation of a random variable

If X is a random variable, the variance of X is defined as:

$$\sigma^2 = \mathrm{Var}(X) = \mathrm{E}((X - \mathrm{E}(X))^2) = \begin{cases} \sum_x (x - \mathrm{E}(X))^2 p_X(x) & \text{for discrete } X \\ \int_{-\infty}^{\infty} (x - \mathrm{E}(X))^2 f_X(x)\,\mathrm{d}x & \text{for continuous } X \end{cases}$$

whenever the sum or integral is finite. The standard deviation is defined as:

$$\sigma = \sqrt{\mathrm{Var}(X)}.$$

Recall the following properties of variance from **ST104b Statistics 2**.

i. $\mathrm{Var}(X) \geq 0$, i.e. variance is always non-negative.

ii. $\mathrm{Var}(a_0 + a_1 X) = a_1^2 \mathrm{Var}(X)$, i.e. variance is invariant to a change in location.

Proof:

i. Since $(X - \mathrm{E}(X))^2$ is a positive random variable, it follows that:

$$\mathrm{Var}(X) = \mathrm{E}((X - \mathrm{E}(X))^2) \geq 0.$$

ii. Define $Y = a_0 + a_1 X$, i.e. Y is a linear transformation of X, then by linearity $\mathrm{E}(Y) = a_0 + a_1 \mathrm{E}(X)$. Hence $Y - \mathrm{E}(Y) = a_1(X - \mathrm{E}(X))$ and so:

$$\begin{aligned}\mathrm{Var}(a_0 + a_1 X) = \mathrm{Var}(Y) &= \mathrm{E}((Y - \mathrm{E}(Y))^2) \\ &= \mathrm{E}(a_1^2 (X - \mathrm{E}(X))^2) \\ &= a_1^2 \mathrm{Var}(X).\end{aligned}$$

■

In practice it is often easier to derive the variance of a random variable X using one of the following alternative, but equivalent, results.

i. $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$.

ii. $\mathrm{Var}(X) = \mathrm{E}(X(X-1)) - \mathrm{E}(X)\,\mathrm{E}(X-1)$.

Example 3.22 Suppose that $X \sim \text{Bernoulli}(\pi)$. The mass function is:

$$p_X(x) = \begin{cases} \pi^x (1-\pi)^{1-x} & \text{for } x = 0, 1 \\ 0 & \text{otherwise.} \end{cases}$$

Hence the mean of X is:

$$\begin{aligned}\mathrm{E}(X) = \sum_x x\, p_X(x) &= \sum_{x=0}^{1} x\, \pi^x (1-\pi)^{1-x} \\ &= 0 \times (1-\pi) + 1 \times \pi \\ &= \pi.\end{aligned}$$

Also, we have:

$$\begin{aligned}\mathrm{E}(X^2) = \sum_x x^2 p_X(x) &= \sum_{x=0}^{1} x^2 \pi^x (1-\pi)^{1-x} \\ &= 0^2 \times (1-\pi) + 1^2 \times \pi \\ &= \pi.\end{aligned}$$

Therefore:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \pi - \pi^2 = \pi(1-\pi).$$

Example 3.23 Suppose that $X \sim \text{Bin}(n, \pi)$. We know that $\text{E}(X) = n\pi$. To find $\text{Var}(X)$ it is most convenient to calculate $\text{E}(X(X-1))$, from which we can recover $\text{E}(X^2)$. We have:

$$\begin{aligned}
\text{E}(X(X-1)) &= \sum_x x(x-1)\, p_X(x) && \text{(by definition)} \\
&= \sum_{x=0}^{n} x(x-1) \binom{n}{x} \pi^x (1-\pi)^{n-x} && \text{(substituting } p_X(x)\text{)} \\
&= \sum_{x=2}^{n} x \binom{n}{x} \pi^x (1-\pi)^{n-x} && \text{(since } x\,p_X(x) = 0 \text{ when } x = 0 \text{ and } 1\text{)} \\
&= \sum_{x=2}^{n} \frac{n(n-1)(n-2)!}{(x-2)!\,[(n-2)-(x-2)]!}\, \pi^2 \pi^{x-2} (1-\pi)^{n-x} \\
& && \text{(as } x(x-1)/x! = 1/(x-2)!\text{)} \\
&= n(n-1)\pi^2 \sum_{x=2}^{n} \binom{n-2}{x-2} \pi^{x-2} (1-\pi)^{n-x} \\
& && \text{(taking } n(n-1)\pi^2 \text{ outside)} \\
&= n(n-1)\pi^2 \sum_{y=0}^{n-2} \binom{n-2}{y} \pi^y (1-\pi)^{(n-2)-y} && \text{(setting } y = x-2\text{)} \\
&= n(n-1)\pi^2 \times 1 && \text{(since the sum is a Bin}(n-2, \pi) \text{ mass function)} \\
&= n(n-1)\pi^2.
\end{aligned}$$

Therefore:

$$\begin{aligned}
\text{Var}(X) &= \text{E}(X(X-1)) - \text{E}(X)\,\text{E}(X-1) \\
&= n(n-1)\pi^2 - n\pi(n\pi - 1) \\
&= n\pi((n-1)\pi - (n\pi - 1)) \\
&= n\pi(1-\pi).
\end{aligned}$$

In practice, this is often written as $\text{Var}(X) = npq$, where $p = \pi$ and $q = 1 - p$.

Activity 3.6 Consider a continuous random variable X with density function:

$$f_X(x) = \begin{cases} x - \dfrac{x^3}{4} & \text{for } 0 \le x \le 2 \\ 0 & \text{otherwise.} \end{cases}$$

Calculate $\text{E}(X)$ and $\text{Var}(X)$.

Activity 3.7 Prove that, if $\lambda > 0$, then $x\mathrm{e}^{-\lambda x} \to 0$ as $x \to \infty$. Hence use integration by parts to show that if $X \sim \mathrm{Exp}(\lambda)$, then $\mathrm{E}(X) = 1/\lambda$.

Activity 3.8 For a random variable X, prove that:

(a) $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$.

(b) $\mathrm{Var}(X) = \mathrm{E}(X(X-1)) - \mathrm{E}(X)\,\mathrm{E}(X-1)$.

Activity 3.9 Suppose $X \sim \mathrm{Exp}(\lambda)$. Calculate $\mathrm{Var}(X)$.

Activity 3.10 Suppose X is a random variable. Show that:

$$\mathrm{E}(I_{(-\infty,x]}(X)) = F_X(x).$$

3.9.4 Inequalities involving expectation

Here we consider bounds for probabilities and expectations which are beneficial due to their generality. These can be useful in proofs of convergence results. We begin with the **Markov inequality**.

Markov inequality

Let X be a positive random variable with $\mathrm{E}(X) < \infty$, then:

$$P(X \geq a) \leq \frac{\mathrm{E}(X)}{a}$$

for any constant $a > 0$.

Proof: Here we consider the continuous case. A similar argument holds for the discrete case.

$$P(X \geq a) = \int_a^\infty f_X(x)\,\mathrm{d}x \qquad \text{(by definition)}$$

$$\Rightarrow \quad P(X \geq a) \leq \int_a^\infty \frac{x}{a} f_X(x)\,\mathrm{d}x \qquad \text{(since } 1 \leq x/a \text{ for } x \in [a,\infty)\text{)}$$

$$\Rightarrow \quad P(X \geq a) \leq \frac{1}{a}\int_0^\infty x\, f_X(x)\,\mathrm{d}x \qquad \text{(since } \textstyle\int_a^\infty g(x)\,\mathrm{d}x \leq \int_0^\infty g(x)\,\mathrm{d}x \text{ for positive } g\text{)}$$

$$\Rightarrow \quad P(X \geq a) \leq \frac{1}{a}\,\mathrm{E}(X). \qquad \text{(by definition of } \mathrm{E}(X)\text{)}$$

■

So the Markov inequality provides an upper bound on the probability in the upper tail of a distribution. Its appeal lies in its generality, since no distributional assumptions are required. However, a consequence is that the bound may be very loose as the following example demonstrates.

Example 3.24 Suppose human life expectancy in a developed country is 80 years. If we let X denote the positive random variable of lifespan, then $\mathrm{E}(X) = 80$. Without imposing a distributional assumption on lifespan, we may find an upper bound on the probability that a human in the country lives to be over 160. Using the Markov inequality we have:

$$P(X \geq 160) \leq \frac{\mathrm{E}(X)}{160} = \frac{80}{160} = 0.5$$

which is unrealistic as we would expect this probability to be (very) close to zero!

We now extend the Markov inequality to consider random variables which are not constrained to be positive, using the **Chebyshev inequality**.

Chebyshev inequality

Let X be a random variable with $\mathrm{Var}(X) < \infty$, then:

$$P(|X - \mathrm{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$$

for any constant $a > 0$.

Proof: This follows from the Markov inequality by setting $Y = (X - \mathrm{E}(X))^2$. Hence Y is a positive random variable so the Chebyshev inequality holds. By definition $\mathrm{E}(Y) = \mathrm{Var}(X)$, so the Markov inequality gives:

$$P((X - \mathrm{E}(X))^2 \geq a^2) = P(|X - \mathrm{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$$

using a^2 in place of a. ■

Applying the Chebyshev inequality to a *standardised* distribution, i.e. if X is a random variable with $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, then for a real constant $k > 0$ we have:

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

Proof: This follows immediately from the Chebyshev inequality by setting $a = k\sigma$. ■

Example 3.25 Suppose we seek an upper bound on the probability of a random variable lying beyond two standard deviations from its mean. We have:

$$P(|X - \mu| \geq 2\sigma) \leq \frac{1}{4}.$$

If $X \sim N(\mu, \sigma^2)$, then it is known that there is (approximately) a 0.05 probability of being beyond two standard deviations from the mean, which is clearly much lower than 0.25.

The above example demonstrates that (for the normal distribution at least) the bound can be very inaccurate. However, it is the generalisability of the result to all distributions with finite variance which makes this a useful result.

We now consider a final inequality – the **Jensen inequality** – but first we begin with the definition of a convex function.

Convex function

A function $g : \mathbb{R} \to \mathbb{R}$ is a **convex function** if for any real constant a we can find k such that:

$$g(x) \geq g(a) + k(x - a) \quad \text{for all } x \in \mathbb{R}.$$

Example 3.26 The function $g(x) = x^2$ is a convex function.

Jensen inequality

If X is a random variable with $\mathrm{E}(X) < \infty$ and g is a convex function such that $\mathrm{E}(g(X)) < \infty$, then:

$$\mathrm{E}(g(X)) \geq g(\mathrm{E}(X)).$$

Example 3.27 We consider applications of the Jensen inequality such that we may derive various relationships involving expectation.

1. *Linear:* If $g(x) = a_0 + a_1 x$, then g is convex. Applying the Jensen inequality we have:
$$\mathrm{E}(a_0 + a_1 X) \geq a_0 + a_1 \mathrm{E}(X).$$
Indeed, by linearity of expectation, $\mathrm{E}(a_0 + a_1 X) = a_0 + a_1 \mathrm{E}(X)$.

2. *Quadratic:* If $g(x) = x^2$, then g is convex. Applying the Jensen inequality we have:
$$\mathrm{E}(X^2) \geq (\mathrm{E}(X))^2.$$
It then follows that $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 \geq 0$, ensuring the variance is non-negative.

3. *Reciprocal:* If $g(x) = 1/x$, then g is convex for $x \geq 0$. Applying the Jensen inequality we have:
$$\mathrm{E}\left(\frac{1}{X}\right) \geq \frac{1}{\mathrm{E}(X)}.$$

3.9.5 Moments

Characterising a probability distribution by key attributes is desirable. For a random variable X the mean, $\mathrm{E}(X)$, is our preferred measure of central tendency, while the variance, $\mathrm{Var}(X) = \mathrm{E}((X - \mathrm{E}(X))^2)$, is our preferred measure of dispersion (or its standard deviation). However, these are not exhaustive of distribution attributes which may interest us. **Skewness** (the departure from symmetry) and **kurtosis** (the fatness of tails) are also important, albeit less important than the mean and variance on a *relative* basis.

On a rank-order basis we will think of the mean as being the most important attribute of a distribution, followed by the variance, skewness and then kurtosis. Nonetheless, all of these attributes may be expressed in terms of **moments** and **central moments**, now defined. (Note that moments were introduced in **ST104b Statistics 2** in the context of method of moments estimation.)

Moments

If X is a random variable, and r is a positive integer, then the rth **moment** of X is:

$$\mu_r = \mathrm{E}(X^r)$$

whenever this is well-defined.

Example 3.28 Setting $r = 1$ produces the first moment, which is the mean of the distribution since:

$$\mu_1 = \mathrm{E}(X^1) = \mathrm{E}(X) = \mu$$

provided $\mathrm{E}(X) < \infty$.

Setting $r = 2$ produces the second moment, which combined with the mean can be used to determine the variance since:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \mu_2 - (\mu_1)^2 = \sigma^2$$

provided $\mathrm{E}(X^2) < \infty$.

Moments are determined by the horizontal location of the distribution. For the variance, our preferred measure of dispersion, we would wish this to be invariant to a shift up or down the horizontal axis. This leads to *central moments* which account for the value of the mean.

Central moments

If X is a random variable, and r is a positive integer, then the rth **central moment** of X is:

$$\mu'_r = \mathrm{E}((X - \mathrm{E}(X))^r)$$

whenever this is well-defined.

Example 3.29 Setting $r = 1$ produces the first central moment, which is always zero since:

$$\mu'_1 = \mathrm{E}((X - \mathrm{E}(X))^1) = \mathrm{E}(X - \mu_1) = \mathrm{E}(X) - \mu_1 = \mathrm{E}(X) - \mathrm{E}(X) = 0$$

provided $\mathrm{E}(X) < \infty$.

Setting $r = 2$ produces the second central moment, which is the variance of the distribution since:

$$\mu'_2 = \mathrm{E}((X - \mathrm{E}(X))^2) = \mathrm{Var}(X) = \sigma^2$$

provided $\mathrm{E}(X^2) < \infty$.

Central moments can be expressed in terms of (non-central) moments by:

$$\mu'_r = \sum_{i=0}^{r} \binom{r}{i} (-\mu_1)^i \mu_{r-i}. \tag{3.3}$$

Example 3.30 Using (3.3), the second central moment can be expressed as:

$$\mu'_2 = \sum_{i=0}^{2} \binom{2}{i} (-\mu_1)^i \mu_{2-i} = \mu_2 - 2(\mu_1)^2 + (\mu_1)^2 = \mu_2 - (\mu_1)^2$$

noting that $\mu_0 = \mathrm{E}(X^0) = \mathrm{E}(1) = 1$. Of course, this is just an alternative way of saying that $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$.

Example 3.31 Using (3.3), the third central moment can be expressed as:

$$\mu'_3 = \sum_{i=0}^{3} \binom{3}{i} (-\mu_1)^i \mu_{3-i} = \mu_3 - 3\mu_1\mu_2 + 3(\mu_1)^3 - (\mu_1)^3 = \mu_3 - 3\mu_1\mu_2 + 2(\mu_1)^3.$$

Example 3.32 Let X be a random variable which has a Bernoulli distribution with parameter π.

(a) Show that $\mathrm{E}(X^r) = \pi$ for $r = 1, 2, \ldots$.

(b) Find the third central moment of X.

(c) Show that the mean of the binomial distribution with parameters n and π is equal to $n\pi$.

Solution

We have that $X \sim \text{Bernoulli}(\pi)$.

(a) Since $X^r = X$ it follows that:

$$\mathrm{E}(X^r) = \mathrm{E}(X) = 0 \times (1 - \pi) + 1 \times \pi = \pi.$$

(b) We have:

$$\begin{aligned} \mathrm{E}((X - \mathrm{E}(X))^3) &= \pi(1-\pi)^3 - (1-\pi)\pi^3 \\ &= \pi(1-\pi)((1-\pi)^2 - \pi^2) \\ &= \pi(1-\pi)(1-2\pi). \end{aligned}$$

(c) Define $Y = \sum_{i=1}^{n} X_i$, where the X_is are i.i.d. Bernoulli(π) random variables. Hence:

$$\mathrm{E}(Y) = \mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i) = \sum_{i=1}^{n} \pi = n\pi.$$

Example 3.28 showed that the first moment is the mean (our preferred measure of central tendency), while Example 3.29 showed that the second central moment is the variance (our preferred measure of dispersion). We now express skewness and kurtosis in terms of moments.

Coefficient of skewness

If X is a random variable with $\mathrm{Var}(X) = \sigma^2 < \infty$, the **coefficient of skewness** is:

$$\mathrm{Skew}(X) = \gamma_1 = \mathrm{E}\left(\left(\frac{X-\mu}{\sigma}\right)^3\right) = \frac{\mathrm{E}((X-\mu)^3)}{\sigma^3} = \frac{\mu_3'}{(\mu_2')^{3/2}}.$$

So we see that skewness depends on the third central moment, although the reason for this may not be immediately clear. It can be explained by first noting that if $g(x) = x^3$, then g is an odd function meaning $g(-x) = -g(x)$. For a continuous random variable X with density function f_X, the third central moment is:

$$\begin{aligned}
\mathrm{E}((X-\mu)^3) &= \int_{-\infty}^{\infty} (x-\mu)^3 f_X(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} z^3 f_X(\mu+z)\,\mathrm{d}z && \text{(where } z = x-\mu\text{)} \\
&= \int_{0}^{\infty} z^3 (f_X(\mu+z) - f_X(\mu-z))\,\mathrm{d}z. && \text{(since } z^3 \text{ is odd)}
\end{aligned}$$

The term $(f_X(\mu+z) - f_X(\mu-z))$ compares the density at the points a distance z above and below μ, such that any difference signals asymmetry. If such sources of asymmetry are far from μ, then when multiplied by z^3 they result in a large coefficent of skewness.

Example 3.33 Suppose $X \sim \mathrm{Exp}(\lambda)$. To derive the coefficient of skewness we require the second and third central moments, i.e. μ_2' and μ_3', respectively. Here these will be calculated from the first three (non-central) moments, μ_1, μ_2 and μ_3. We will proceed to find a general expression for the rth moment for no additional effort. We have:

$$\begin{aligned}
\mu_r = \mathrm{E}(X^r) &= \int_{-\infty}^{\infty} x^r f_X(x)\,\mathrm{d}x \\
&= \int_{0}^{\infty} x^r \lambda \mathrm{e}^{-\lambda x}\,\mathrm{d}x && \text{(using the exponential density)} \\
&= \Big[-x^r \mathrm{e}^{-\lambda x}\Big]_0^{\infty} + r\int_0^{\infty} x^{r-1}\mathrm{e}^{-\lambda x}\,\mathrm{d}x && \text{(using integration by parts)} \\
&= \frac{r}{\lambda}\,\mu_{r-1}.
\end{aligned}$$

Noting that $\mu_0 = 1$, by recursion we have:

$$\mu_r = \frac{r}{\lambda}\,\frac{r-1}{\lambda}\cdots\frac{2}{\lambda}\,\frac{1}{\lambda} = \frac{r!}{\lambda^r}$$

from which we obtain:

$$\mu_1 = \frac{1}{\lambda}, \quad \mu_2 = \frac{2}{\lambda^2} \quad \text{and} \quad \mu_3 = \frac{6}{\lambda^3}.$$

Hence the second central moment is:

$$\mu_2' = \mu_2 - (\mu_1)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

and the third central moment is:

$$\mu_3' = \mu_3 - 3\mu_1\mu_2 + 2(\mu_1)^3 = \frac{6}{\lambda^3} - 3 \times \frac{1}{\lambda}\frac{2}{\lambda^2} + 2 \times \left(\frac{1}{\lambda}\right)^3 = \frac{2}{\lambda^3}.$$

Therefore, the coefficient of skewness is:

$$\gamma_1 = \frac{\mu_3'}{(\mu_2')^{3/2}} = \frac{2/\lambda^3}{(1/\lambda^2)^{3/2}} = \frac{2/\lambda^3}{1/\lambda^3} = 2$$

which is positive (recall that the exponential distribution is positively skewed).

Coefficient of kurtosis

If X is a random variable with $\text{Var}(X) = \sigma^2 < \infty$, the **coefficient of kurtosis** is:

$$\text{Kurt}(X) = \gamma_2 = \text{E}\left(\left(\frac{X-\mu}{\sigma}\right)^4\right) - 3 = \frac{\text{E}((X-\mu)^4)}{\sigma^4} - 3 = \frac{\mu_4'}{(\mu_2')^2} - 3. \qquad (3.4)$$

We see that the term '-3' appears in the definition of kurtosis. Convention means we measure kurtosis with respect to a normal distribution. Noting that the fourth central moment of a normal distribution is $3\sigma^4$, we have that the kurtosis for a normal distribution is:

$$\gamma_2 = \frac{\text{E}((X-\mu)^4)}{\sigma^4} - 3 = \frac{3\sigma^4}{\sigma^4} - 3 = 0.$$

The coefficient of kurtosis as defined in (3.4) with the '-3' term is often called **excess kurtosis**, i.e. kurtosis in excess of that of a normal distribution.

Example 3.34 Suppose $X \sim \text{Exp}(\lambda)$. To derive the coefficient of kurtosis we require the fourth central moment. Example 3.33 gave us the general expression for the rth moment:

$$\mu_r = \frac{r!}{\lambda^r}.$$

The fourth central moment is:

$$\mu_4' = \sum_{i=0}^{4} \binom{4}{i} (-\mu_1)^i \mu_{4-i} = \mu_4 - 4\mu_1\mu_3 + 6(\mu_1)^2\mu_2 - 4(\mu_1)^4 + (\mu_1)^4 = \frac{9}{\lambda^4}.$$

Since $\mu_2' = 1/\lambda^2$, the coefficient of kurtosis is:

$$\gamma_2 = \frac{\mu_4'}{(\mu_2')^2} - 3 = \frac{9/\lambda^4}{(1/\lambda^2)^2} - 3 = 9 - 3 = 6.$$

Example 3.35 Find the mean and variance of the gamma distribution:

$$f_X(x) = \frac{1}{\Gamma(\alpha)}\,\beta^\alpha x^{\alpha-1}\mathrm{e}^{-\beta x} = \frac{1}{(\alpha-1)!}\,\beta^\alpha x^{\alpha-1}\mathrm{e}^{-\beta x}$$

for $x \geq 0$, $\alpha > 0$ and $\beta > 0$, where $\Gamma(\alpha) = (\alpha-1)!$.

Hint: Note that since $f_X(x)$ is a density function, we can write:

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

Solution

We can find the rth moment, and use this result to get the mean and variance. We have:

$$\begin{aligned}
\mathrm{E}(X^r) = \mu_r = \int_{-\infty}^{\infty} x^r f_X(x)\,\mathrm{d}x &= \int_0^\infty x^r \frac{1}{(\alpha-1)!}\,\beta^\alpha x^{\alpha-1}\mathrm{e}^{-\beta x}\,\mathrm{d}x \\
&= \int_0^\infty \frac{1}{(\alpha-1)!}\,x^{r+\alpha-1}\beta^\alpha \mathrm{e}^{-\beta x}\,\mathrm{d}x \\
&= \frac{(r+\alpha-1)!}{(\alpha-1)!}\,\frac{1}{\beta^r}\int_0^\infty \frac{1}{(r+\alpha-1)!}x^{(r+\alpha)-1}\beta^{r+\alpha}\mathrm{e}^{-\beta x}\,\mathrm{d}x \\
&= \frac{(r+\alpha-1)!}{(\alpha-1)!}\,\frac{1}{\beta^r}
\end{aligned}$$

since the integrand is a Gamma$(r+\alpha, \beta)$ density function, which integrates to 1. So:

$$\mu_r = \frac{(r+\alpha-1)!}{(\alpha-1)!\,\beta^r}.$$

Using the result:

$$\mathrm{E}(X) = \mu_1 = \frac{\alpha!}{(\alpha-1)!\,\beta} = \frac{\alpha}{\beta}$$

and:

$$\mathrm{E}(X^2) = \mu_2 = \frac{(\alpha+1)!}{(\alpha-1)!\,\beta^2} = \frac{\alpha(\alpha+1)}{\beta^2}.$$

Therefore, the variance is:

$$\mu_2 - (\mu_1)^2 = \frac{\alpha(\alpha+1)}{\beta^2} - \frac{\alpha^2}{\beta^2} = \frac{\alpha}{\beta^2}.$$

Both the mean and variance increase with α increasing and decrease with β increasing.

We can also compute $\mathrm{E}(X)$ and $\mathrm{E}(X^2)$ by substituting $y = \beta x$. Note that this gives $\mathrm{d}x = (1/\beta)\,\mathrm{d}y$. For example:

$$\begin{aligned}
\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f_X(x)\,\mathrm{d}x &= \int_0^\infty x\,\frac{1}{(\alpha-1)!}\,\beta^\alpha x^{\alpha-1}\mathrm{e}^{-\beta x}\,\mathrm{d}x \\
&= \frac{1}{(\alpha-1)!\,\beta}\int_0^\infty y^\alpha \mathrm{e}^{-y}\,\mathrm{d}y
\end{aligned}$$

and recognise that the integral is the definition of $\Gamma(\alpha+1) = \alpha!$.

Example 3.36 Find the mean and variance of the Poisson distribution:

$$p_X(x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}$$

for $x = 0, 1, 2, \ldots$, and $\lambda > 0$.

Hint: Note that since $p_X(x)$ is a mass function, we can write:

$$\sum_{x=0}^{\infty} \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} = 1.$$

Solution

By direct calculation we have:

$$\mathrm{E}(X) = \sum_{x=0}^{\infty} x\,\frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} = \mathrm{e}^{-\lambda}\lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \mathrm{e}^{-\lambda}\lambda \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = \mathrm{e}^{-\lambda}\lambda\mathrm{e}^{\lambda} = \lambda$$

setting $y = x - 1$, and:

$$\begin{aligned}
\mathrm{E}(X^2) = \sum_{x=0}^{\infty} x^2 \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} x^2 \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} &= \mathrm{e}^{-\lambda}\lambda \sum_{x=1}^{\infty} x \frac{\lambda^{x-1}}{(x-1)!} \\
&= \mathrm{e}^{-\lambda}\lambda \sum_{y=0}^{\infty} (y+1)\,\frac{\lambda^y}{y!} \\
&= \mathrm{e}^{-\lambda}\lambda \underbrace{\sum_{y=0}^{\infty} \frac{y\lambda^y}{y!}}_{=\mathrm{e}^{\lambda}\mathrm{E}(Y)} + \mathrm{e}^{-\lambda}\lambda \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \\
&= \mathrm{e}^{-\lambda}\lambda\mathrm{e}^{\lambda}\lambda + \mathrm{e}^{-\lambda}\lambda\mathrm{e}^{\lambda} \\
&= \lambda^2 + \lambda
\end{aligned}$$

again setting $y = x - 1$, and noting that for $Y \sim \mathrm{Pois}(\lambda)$ then $\mathrm{E}(Y) = \lambda$.

Another way here is to find the rth *factorial* moment:

$$\mu_{(r)} = \mathrm{E}(X^{(r)}) = \mathrm{E}(X(X-1)\cdots(X-r+1)).$$

This works out very simply. We can then convert to the mean and variance. The critical property that makes $\mu_{(r)}$ work out easily is that for an integer x from 1 to $r-1$ we have:

$$\frac{x^{(r)}}{x!} = \frac{x^{(r)}}{x^{(x)}} = \frac{1}{(x-r)!}.$$

We have:

$$\begin{aligned}
\mathrm{E}(X^{(r)}) = \sum_{x=0}^{\infty} x^{(r)} \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} &= \sum_{x=r}^{\infty} \frac{x^{(r)}}{x!} \mathrm{e}^{-\lambda}\lambda^x \\
&= \sum_{x=r}^{\infty} \frac{1}{(x-r)!}\, \mathrm{e}^{-\lambda}\lambda^x \\
&= \lambda^r \sum_{x=r}^{\infty} \frac{1}{(x-r)!}\, \mathrm{e}^{-\lambda}\lambda^{x-r} \\
&= \lambda^r \sum_{y=0}^{\infty} \frac{\mathrm{e}^{-\lambda}\lambda^y}{y!} \\
&= \lambda^r.
\end{aligned}$$

The last step follows because we are adding together all the probabilities for a Poisson distribution with parameter λ.

Now it is straightforward to get the mean and variance. For the mean we have:

$$\mathrm{E}(X) = \mathrm{E}(X^{(1)}) = \lambda.$$

Since $\mathrm{E}(X(X-1)) + \mathrm{E}(X) = \mathrm{E}(X^2)$, then:

$$\mu_2 = \mathrm{E}(X^{(2)}) + \mathrm{E}(X) = \mu_{(2)} + \mu_{(1)} = \lambda^2 + \lambda$$

and so:

$$\mathrm{Var}(X) = \mu_2 - (\mu_1)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Example 3.37 Find the mean and variance of the Pareto distribution:

$$f_X(x) = \frac{a-1}{(1+x)^a}$$

for $x > 0$ and $a > 1$.

Hint: It is easier to find $\mathrm{E}(X+1)$ and $\mathrm{E}((X+1)^2)$. We then have that $\mathrm{E}(X) = \mathrm{E}(X+1) - 1$ and $\mathrm{E}(X^2)$ comes from $\mathrm{E}((X+1)^2)$ in a similar manner.

Solution

We can directly integrate for μ_r, writing the integral as a Pareto distribution with parameter $a-r$ after a transformation. It is easier to work with $Y = X+1$, noting that $\mathrm{E}(X) = \mathrm{E}(Y) - 1$ and $\mathrm{Var}(Y) = \mathrm{Var}(X)$. We have:

$$\mathrm{E}(Y^r) = \int_0^{\infty} (x+1)^r \frac{a-1}{(1+x)^a}\,\mathrm{d}x = \frac{a-1}{a-r-1} \int_0^{\infty} \frac{a-r-1}{(1+x)^{a-r}}\,\mathrm{d}x = \frac{a-1}{a-r-1}$$

provided that $a - r > 1$ (otherwise the integral is not defined). So:

$$\mathrm{E}(Y) = \frac{a-1}{a-2} \quad \Rightarrow \quad \mathrm{E}(X) = \frac{1}{a-2}$$

for $a > 2$. Provided $a > 3$, then:

$$\begin{aligned}\mathrm{Var}(X) = \mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 &= \frac{a-1}{a-3} - \left(\frac{a-1}{a-2}\right)^2 \\ &= \frac{(a-1)((a-2)^2 - (a-1)(a-3))}{(a-2)^2(a-3)} \\ &= \frac{a-1}{(a-2)^2(a-3)}.\end{aligned}$$

3.10 Generating functions

3.10.1 Moment generating functions

In the previous section we saw how useful moments (and central moments) are for expressing important attributes of a probability distribution such as the mean, variance, skewness and kurtosis. Many distributions can summarise all of their moments, $\mathrm{E}(X), \mathrm{E}(X^2), \ldots$, into a single function known as the **moment generating function**. This is, literally, a function (when it exists) which can be used to generate the moments of a distribution.

Moment generating function

The **moment generating function** (mgf) of a random variable X is a function $M_X : \mathbb{R} \to [0, \infty)$ defined as:

$$M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \begin{cases} \sum\limits_x \mathrm{e}^{tx} p_X(x) & \text{for discrete } X \\ \int_{-\infty}^{\infty} \mathrm{e}^{tx} f_X(x)\,\mathrm{d}x & \text{for continuous } X \end{cases}$$

where to be well-defined we require $M_X(t) < \infty$ for all $t \in [-h, h]$ for some $h > 0$. So the mgf must be defined in an interval around the origin, which will be necessary when derivatives of the mgf with respect to t are taken and evaluated when $t = 0$, i.e. such as $M'_X(0)$, $M''_X(0)$ etc.

If the expected value $\mathrm{E}(\mathrm{e}^{tX})$ is infinite, the random variable X does not have an mgf.

$M_X(t)$ is a function of real numbers t. It is not a random variable itself.

The form of the mgf is *not* interesting or informative in itself. Instead, the reason we define the mgf is that it is a convenient tool for deriving means and variances of distributions, using the following results:

$$M'_X(0) = \mathrm{E}(X) \quad \text{and} \quad M''_X(0) = \mathrm{E}(X^2)$$

which also gives:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = M''_X(0) - (M'_X(0))^2.$$

This is useful if the mgf is easier to derive than $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ directly.

Other moments are obtained from the mgf similarly:

$$M_X^{(r)}(0) = \mathrm{E}(X^r) \quad \text{for } r = 1, 2, \ldots.$$

To see why, note that $M_X(t)$ is the expected value of an exponential function of X. Recall the Taylor expansion of e^x is:

$$\mathrm{e}^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^r}{r!} + \cdots = \sum_{i=0}^{\infty} \frac{x^i}{i!}.$$

All the derivatives of e^x are also e^x, i.e. the rth derivative is:

$$\frac{\mathrm{d}^r}{\mathrm{d}x^r}\,\mathrm{e}^x = \mathrm{e}^x \quad \text{for } r = 1, 2, \ldots.$$

Therefore, we may express the moment generating function as a polynomial in t, i.e. we have:

$$M_X(t) = 1 + t\,\mathrm{E}(X) + \frac{t^2}{2!}\,\mathrm{E}(X^2) + \cdots + \frac{t^r}{r!}\,\mathrm{E}(X^r) + \cdots = \sum_{i=0}^{\infty} \frac{t^i}{i!}\,\mathrm{E}(X^i).$$

Proof: This follows immediately from the series expansion of e^x and the linearity of expectation:

$$M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \mathrm{E}\left(\sum_{i=0}^{\infty} \frac{(tX)^i}{i!}\right) = \sum_{i=0}^{\infty} \frac{t^i}{i!}\,\mathrm{E}(X^i).$$

■

We are now in a position to understand how moments can be generated from a moment generating function. There are two approaches.

1. Use the coefficients of $\mathrm{E}(X^r)$, for $r = 1, 2, \ldots$, in the series expansion of $M_X(t)$.

2. Use derivatives of $M_X(t)$.

Determining moments by comparing coefficients

The coefficient of t^r in the series expansion of $M_X(t)$ is the rth moment divided by $r!$. Hence the rth moment can be determined by comparing coefficients. We have:

$$M_X(t) = \sum_{i=0}^{\infty} \frac{t^i}{i!}\,\mathrm{E}(X^i) = \sum_{i=0}^{\infty} a_i t^i \quad \Rightarrow \quad \mathrm{E}(X^r) = r!\,a_r$$

where $a_i = \mathrm{E}(X^i)/i!$.

This method is quick provided it is easy to derive the polynomial expansion in t.

Determining moments using derivatives

The rth derivative of a moment generating function evaluated at zero is the rth moment, that is:

$$M_X^{(r)}(0) = \frac{\mathrm{d}^r}{\mathrm{d}t^r} M_X(t)\Big|_{t=0} = \mathrm{E}(X^r) = \mu_r.$$

When the polynomial expansion in t is not easy to derive, calculating derivatives of $M_X(t)$ is more convenient.

Proof: Since:

$$M_X(t) = 1 + t\,\mathrm{E}(X) + \frac{t^2}{2!}\mathrm{E}(X^2) + \cdots + \frac{t^r}{r!}\mathrm{E}(X^r) + \cdots$$

then:

$$M_X^{(r)}(t) = \mathrm{E}(X^r) + t\,\mathrm{E}(X^{r+1}) + \frac{t^2}{2!}\mathrm{E}(X^{r+2}) + \cdots = \sum_{i=r}^{\infty} \frac{t^{i-r}}{(i-r)!}\mathrm{E}(X^i).$$

When evaluated at $t = 0$ only the first term, $\mathrm{E}(X^r)$, is non-zero, proving the result. ■

The moment generating function uniquely determines a probability distribution. In other words, if for two random variables X and Y we have $M_X(t) = M_Y(t)$ (for points around $t = 0$), then X and Y have the same distribution.

Uniqueness of the moment generating function

If X and Y are random variables and we can find $h > 0$ such that $M_X(t) = M_Y(t)$ for all $t \in [-h, h]$, then $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$.

We now show examples of deriving the moment generating function and subsequently using them to obtain moments.

Example 3.38 Suppose $X \sim \mathrm{Pois}(\lambda)$, i.e. we have:

$$p_X(x) = \begin{cases} \dfrac{\mathrm{e}^{-\lambda}\lambda^x}{x!} & \text{for } x = 0, 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

The moment generating function for this distribution is:

$$M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \sum_x \mathrm{e}^{tx} p_X(x) = \sum_{x=0}^{\infty} \mathrm{e}^{tx}\frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} = \mathrm{e}^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda\mathrm{e}^t)^x}{x!} = \exp(\lambda(\mathrm{e}^t - 1)).$$

From $M_X(t) = \exp(\lambda(\mathrm{e}^t - 1))$ we obtain:

$$M_X'(t) = \lambda\mathrm{e}^t\mathrm{e}^{\lambda(\mathrm{e}^t-1)}$$

and:

$$M_X''(t) = \lambda\mathrm{e}^t(1 + \lambda\mathrm{e}^t)\mathrm{e}^{\lambda(\mathrm{e}^t-1)}$$

and hence (since $e^0 = 1$):

$$M'_X(0) = \lambda = \mathrm{E}(X) \quad \text{and} \quad M''_X(0) = \lambda(1+\lambda) = \mathrm{E}(X^2).$$

Therefore:

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

Example 3.39 Suppose $X \sim \mathrm{Geo}(\pi)$, for the second version of the geometric distribution, i.e. we have:

$$p_X(x) = \begin{cases} (1-\pi)^x \pi & \text{for } x = 0, 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

The moment generating function for this distribution is:

$$M_X(t) = \mathrm{E}(e^{tX}) = \sum_x e^{tx} p_X(x) = \sum_{x=0}^{\infty} e^{tx}(1-\pi)^x \pi = \pi \sum_{x=0}^{\infty} (e^t(1-\pi))^x = \frac{\pi}{1 - e^t(1-\pi)}$$

using the sum to infinity of a geometric series, for $t < -\ln(1-\pi)$ to ensure convergence of the sum.

From $M_X(t) = \pi/(1 - e^t(1-\pi))$ we obtain, using the chain rule:

$$M'_X(t) = \frac{\pi(1-\pi)e^t}{(1 - e^t(1-\pi))^2}$$

also, using the quotient rule:

$$M''_X(t) = \frac{\pi(1-\pi)e^t(1-(1-\pi)e^t)(1+(1-\pi)e^t)}{(1 - e^t(1-\pi))^4}$$

and hence (since $e^0 = 1$):

$$M'_X(0) = \frac{1-\pi}{\pi} = \mathrm{E}(X).$$

For the variance:

$$M''_X(0) = \frac{(1-\pi)(2-\pi)}{\pi^2} = \mathrm{E}(X^2)$$

and so:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{(1-\pi)(2-\pi)}{\pi^2} - \frac{(1-\pi)^2}{\pi^2} = \frac{1-\pi}{\pi^2}.$$

Example 3.40 Suppose $X \sim \mathrm{Exp}(\lambda)$, i.e. with density function:

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

The moment generating function for this distribution is:

$$\begin{aligned} M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) &= \int_{-\infty}^{\infty} \mathrm{e}^{tx} f_X(x)\,\mathrm{d}x \\ &= \int_0^{\infty} \mathrm{e}^{tx}\lambda\mathrm{e}^{-\lambda x}\,\mathrm{d}x \\ &= \int_0^{\infty} \lambda\mathrm{e}^{-(\lambda-t)x}\,\mathrm{d}x \\ &= \frac{\lambda}{\lambda-t}\underbrace{\int_0^{\infty}(\lambda-t)\mathrm{e}^{-(\lambda-t)x}\,\mathrm{d}x}_{=1} \\ &= \frac{\lambda}{\lambda-t} \quad \text{for } t<\lambda \end{aligned}$$

where note the integral is that of an $\mathrm{Exp}(\lambda - t)$ distribution over its support, hence is equal to 1.

From $M_X(t) = \lambda/(\lambda - t)$ we obtain:

$$M_X'(t) = \frac{\lambda}{(\lambda-t)^2} \quad \text{and} \quad M_X''(t) = \frac{2\lambda}{(\lambda-t)^3}$$

so:

$$\mathrm{E}(X) = M_X'(0) = \frac{1}{\lambda} \quad \text{and} \quad \mathrm{E}(X^2) = M_X''(0) = \frac{2}{\lambda^2}$$

and:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Example 3.41 Suppose $X \sim \mathrm{Gamma}(\alpha, \beta)$, i.e. with density function:

$$f_X(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)}\,\beta^{\alpha}x^{\alpha-1}\mathrm{e}^{-\beta x} & \text{for } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The moment generating function for this distribution is:

$$\begin{aligned} M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) &= \int_{-\infty}^{\infty} \mathrm{e}^{tx} f_X(x)\,\mathrm{d}x \\ &= \int_0^{\infty} \mathrm{e}^{tx}\,\frac{1}{\Gamma(\alpha)}\,\beta^{\alpha}x^{\alpha-1}\mathrm{e}^{-\beta x}\,\mathrm{d}x \\ &= \frac{\beta^{\alpha}}{(\beta-t)^{\alpha}}\underbrace{\int_0^{\infty}\frac{1}{\Gamma(\alpha)}(\beta-t)^{\alpha}x^{\alpha-1}\mathrm{e}^{-(\beta-t)x}\,\mathrm{d}x}_{=1} \\ &= \left(\frac{\beta}{\beta-t}\right)^{\alpha} \quad \text{for } t<\beta \end{aligned}$$

where note the integral is that of a Gamma$(\alpha, \beta - t)$ distribution over its support, hence is equal to 1, and where we multiplied by $(\beta - t)^\alpha/(\beta - t)^\alpha = 1$ (hence not affecting the integral) to 'create' a Gamma$(\alpha, \beta - t)$ density function. Since $(\beta - t)^\alpha$ does not depend on x we can place the numerator term inside the integral and the denominator term outside the integral.

We can divide through by β to obtain:

$$M_X(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha} \quad \text{for } t < \beta.$$

Noting the *negative binomial expansion* given by:

$$(1-a)^{-n} = \sum_{i=0}^{\infty} \binom{i+n-1}{n-1} a^i$$

we have:

$$M_X(t) = \sum_{i=0}^{\infty} \binom{i+\alpha-1}{\alpha-1} \left(\frac{t}{\beta}\right)^i = \sum_{i=0}^{\infty} \frac{(i+\alpha-1)!}{(\alpha-1)!\,\beta^i} \frac{t^i}{i!}.$$

Since the rth moment is the coefficient of $t^r/r!$ in the polynomial expansion of $M_X(t)$, we deduce that if $X \sim \text{Gamma}(\alpha, \beta)$, the rth moment is:

$$\text{E}(X^r) = \mu_r = \frac{(r+\alpha-1)!}{(\alpha-1)!\,\beta^r}.$$

We have previously seen that if $X \sim \text{Gamma}(1, \beta)$, then $X \sim \text{Exp}(\beta)$, and so:

$$M_X(t) = \left(\frac{\beta}{\beta - t}\right)^\alpha = \frac{\beta}{\beta - t} = \frac{\lambda}{\lambda - t} \quad \text{for } t < \lambda$$

as derived in Example 3.40. Note the choice of parameter symbol is arbitrary, such that $\beta = \lambda$.

Example 3.42 Find the moment generating function of X, where $X \sim N(\mu, \sigma^2)$. Also, find the mean and the variance of $Y = \text{e}^X$. (Y has a log-normal distribution, popular as a skewed distribution for positive random variables.)

Hint: Compute the mgf for a standard normal random variable $Z = (X - \mu)/\sigma$, where the density function of Z is:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \text{e}^{-z^2/2}$$

then once you have the mgf of Z you can easily find the mgfs of X and Y without integrals.

Solution

We are asked to find:

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

for $X \sim N(\mu, \sigma^2)$. We may write $X = \mu + \sigma Z$, where $Z \sim N(0,1)$, such that:

$$M_X(t) = \mathrm{E}(\mathrm{e}^{t(\mu+\sigma Z)}) = \mathrm{e}^{\mu t}\mathrm{E}(\mathrm{e}^{\sigma t Z}) = \mathrm{e}^{\mu t} M_Z(\sigma t).$$

So we only need to derive the mgf for a standard normal random variable. We have:

$$\begin{aligned} M_Z(t) = \int_{-\infty}^{\infty} \mathrm{e}^{zt} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}\,\mathrm{d}z &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-((z-t)^2 - t^2)/2}\,\mathrm{d}z \\ &= \mathrm{e}^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(z-t)^2/2}\,\mathrm{d}z \\ &= \mathrm{e}^{t^2/2}. \end{aligned}$$

The last integral is that of a $N(t,1)$ density function, and so is equal to 1. The step from line 1 to line 2 follows from the simple algebraic identity:

$$-\frac{1}{2}((z-t)^2 - t^2) = -\frac{1}{2}z^2 + zt.$$

The mgf for the general normal distribution is:

$$M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

The mean of Y is:

$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{e}^X) = M_X(1) = \exp\left(\mu + \frac{\sigma^2}{2}\right).$$

Also, noting $Y^2 = \mathrm{e}^{2X}$, we have:

$$\mathrm{E}(Y^2) = \mathrm{E}(\mathrm{e}^{2X}) = M_X(2) = \exp\left(2\mu + \frac{4\sigma^2}{2}\right)$$

hence the variance of Y is:

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 \\ &= \exp\left(2\mu + \frac{4\sigma^2}{2}\right) - \exp\left(2\mu + \frac{2\sigma^2}{2}\right) \\ &= \mathrm{e}^{2\mu+\sigma^2}(\mathrm{e}^{\sigma^2} - 1). \end{aligned}$$

Example 3.43 Find the moment generating function of the double exponential, or Laplace, distribution with density function:

$$f_X(x) = \frac{1}{2}\mathrm{e}^{-|x|} \quad \text{for } -\infty < x < \infty.$$

Solution

We have:

$$\begin{aligned} M_X(t) = \int_{-\infty}^{\infty} \mathrm{e}^{xt}\,\frac{\mathrm{e}^{-|x|}}{2}\,\mathrm{d}x &= \int_{-\infty}^{0} \mathrm{e}^{xt}\,\frac{\mathrm{e}^{x}}{2}\,\mathrm{d}x + \int_{0}^{\infty} \mathrm{e}^{xt}\,\frac{\mathrm{e}^{-x}}{2}\,\mathrm{d}x \\ &= \int_{-\infty}^{0} \frac{\mathrm{e}^{x(1+t)}}{2}\,\mathrm{d}x + \int_{0}^{\infty} \frac{\mathrm{e}^{-x(1-t)}}{2}\,\mathrm{d}x \\ &= \left[\frac{\mathrm{e}^{x(1+t)}}{2(1+t)}\right]_{-\infty}^{0} + \left[-\frac{\mathrm{e}^{-x(1-t)}}{2(1-t)}\right]_{0}^{\infty} \\ &= \frac{1}{2(1+t)} + \frac{1}{2(1-t)} \\ &= \frac{1}{1-t^2} \end{aligned}$$

where we require $|t| < 1$.

Example 3.44 A random variable X follows the Laplace distribution with parameter $\lambda > 0$ if its density function has the form:

$$f_X(x) = k\mathrm{e}^{-\lambda|x|} \quad \text{for } -\infty < x < \infty$$

where k is a normalising constant.

(a) Find k in terms of λ.

(b) Compute $\mathrm{E}(X^3)$.

(c) Derive the moment generating function of X and provide the interval on which this function is well-defined.

(d) Find the variance of X using the moment generating function derived in (c).

Solution

(a) Since:

$$\begin{aligned} 1 = \int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = k\int_{-\infty}^{\infty} \mathrm{e}^{-\lambda|x|}\,\mathrm{d}x &= k\left(\int_{-\infty}^{0} \mathrm{e}^{\lambda x}\,\mathrm{d}x + \int_{0}^{\infty} \mathrm{e}^{-\lambda x}\,\mathrm{d}x\right) \\ &= k\left(\left[\frac{\mathrm{e}^{\lambda x}}{\lambda}\right]_{-\infty}^{0} + \left[-\frac{\mathrm{e}^{-\lambda x}}{\lambda}\right]_{0}^{\infty}\right) \\ &= k\left(\frac{1}{\lambda} + \frac{1}{\lambda}\right) \\ &= \frac{2k}{\lambda} \end{aligned}$$

it follows that $k = \lambda/2$.

(b) As X has a symmetric distribution, i.e. $f(-x) = f(x)$, and $g(x) = x^3$ is an odd function then $g(-x) = -g(x)$, it follows that:

$$\mathrm{E}(X^3) = 0.$$

(c) We have:

$$\begin{aligned} M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) &= \frac{\lambda}{2}\int_{-\infty}^{\infty} \mathrm{e}^{tx}\mathrm{e}^{-\lambda|x|}\,\mathrm{d}x \\ &= \frac{\lambda}{2}\left(\int_{-\infty}^{0} \mathrm{e}^{(t+\lambda)x}\,\mathrm{d}x + \int_{0}^{\infty} \mathrm{e}^{(t-\lambda)x}\,\mathrm{d}x\right) \\ &= \frac{\lambda}{2}\left(\frac{1}{t+\lambda} - \frac{1}{t-\lambda}\right) \\ &= \frac{\lambda^2}{\lambda^2 - t^2} \end{aligned}$$

for $|t| < \lambda$.

(d) Since $\mathrm{E}(X) = 0$ it holds that the variance of X is equal to $\mathrm{E}(X^2)$. For $|t| < \lambda$, we have:

$$M_X''(t) = \frac{\mathrm{d}}{\mathrm{d}t}\frac{2\lambda^2 t}{(\lambda^2 - t^2)^2} = \frac{2\lambda^2(\lambda^2 - t^2)^2 + 2\lambda^2 t(4t(\lambda^2 - t^2))}{(\lambda^2 - t^2)^4}.$$

Setting $t = 0$ we have:

$$\mathrm{E}(X^2) = M_X''(0) = \frac{2\lambda^6}{\lambda^8} = \frac{2}{\lambda^2}.$$

Note that this should not come as a surprise since X can be written as the difference of two independent and identically distributed exponential random variables, each with parameter λ, say $X = T_1 - T_2$. Hence, due to independence:

$$\begin{aligned} \mathrm{Var}(X) = \mathrm{Var}(T_1 - T_2) &= \mathrm{Var}(T_1) + \mathrm{Var}(T_2) \\ &= \frac{1}{\lambda^2} + \frac{1}{\lambda^2} \\ &= \frac{2}{\lambda^2}. \end{aligned}$$

Activity 3.11 Consider the following game. You pay £5 to play the game. A fair coin is tossed three times. If the first and last tosses are heads, you receive £10 for each head.

(a) What is the expected return from playing this game?

(b) Derive the moment generating function of the return.

3.10.2 Cumulant generating functions and cumulants

Often we may choose to work with the logarithm of the moment generating function as the coefficients of the polynomial expansion of this log-transformation have convenient moment and central moment interpretations.

Cumulant generating function and cumulants

A random variable X with moment generating function $M_X(t)$ has a **cumulant generating function** defined as:

$$K_X(t) = \log M_X(t)$$

where 'log' is the natural logarithm, i.e. to the base e.

The rth **cumulant**, κ_r, is the coefficient of $t^r/r!$ in the expansion of $K_X(t)$, so:

$$K_X(t) = \kappa_1 t + \kappa_2 \frac{t^2}{2!} + \cdots + \kappa_r \frac{t^r}{r!} + \cdots = \sum_{i=1}^{\infty} \frac{\kappa_i t^i}{i!}.$$

As with the relationship between a moment generating function and moments, the same relationship holds for a cumulant generating function and cumulants. There are two approaches.

1. Use the coefficients of κ_k, for $k = 1, 2, \ldots$, in the series expansion of $K_X(t)$.
2. Use derivatives of $K_X(t)$.

Determining cumulants by comparing coefficients

The coefficient of t^r in the series expansion of $K_X(t)$ is the rth cumulant divided by $r!$. Hence the rth cumulant can be determined by comparing coefficients. We have:

$$K_X(t) = \sum_{i=0}^{\infty} \frac{t^i}{i!}\kappa_i = \sum_{i=0}^{\infty} a_i t^i \quad \Rightarrow \quad \kappa_r = r!\, a_r$$

where $a_i = \kappa_i / i!$.

Determining cumulants using derivatives

The rth derivative of a cumulant generating function evaluated at zero is the rth cumulant, that is:

$$K_X^{(r)}(0) = \frac{\mathrm{d}^r}{\mathrm{d}t^r} K_X(t)\Big|_{t=0} = \kappa_r.$$

Cumulants may be expressed in terms of moments and central moments. In particular, the first cumulant is the mean and the second cumulant is the variance.

Relationship between cumulants and moments

If X is a random variable with moments $\{\mu_r\}$, central moments $\{\mu'_r\}$ and cumulants $\{\kappa_r\}$, for $r = 1, 2, \ldots$, then:

i. the first cumulant is the mean:

$$\kappa_1 = \mathrm{E}(X) = \mu_1 = \mu$$

ii. the second cumulant is the variance:

$$\kappa_2 = \mathrm{Var}(X) = \mu'_2 = \sigma^2$$

iii. the third cumulant is the third central moment:

$$\kappa_3 = \mu'_3$$

iv. a function of the fourth and second cumulants yields the fourth central moment:

$$\kappa_4 + 3\kappa_2^2 = \mu'_4.$$

In this course we only prove the first two of these results.

Proof:

i. Applying the chain rule, noting that $M_X(0) = \mathrm{E}(\mathrm{e}^0) = 1$, we have:

$$K'_X(t) = \frac{M'_X(t)}{M_X(t)} \quad \Rightarrow \quad K'_X(0) = M'_X(0) = \mu_1 = \mu \quad \Rightarrow \quad \kappa_1 = \mu.$$

ii. Applying the product rule, writing $K'_X(t) = M'_X(t)(M_X(t))^{-1}$, we have:

$$K''_X(t) = \frac{M''_X(t)}{M_X(t)} - \frac{(M'_X(t))^2}{(M_X(t))^2} \quad \Rightarrow \quad K''_X(0) = \mu_2 - (\mu_1)^2 \quad \Rightarrow \quad \kappa_2 = \sigma^2.$$

■

Example 3.45 Suppose that X is a degenerate random variable with probability mass function:

$$p_X(x) = \begin{cases} 1 & \text{for } x = \mu \\ 0 & \text{otherwise.} \end{cases}$$

Show that the cumulant generating function of X is:

$$K_X(t) = \mu t.$$

Solution

We have:

$$M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \sum_x \mathrm{e}^{tx} p_X(x) = \mathrm{e}^{t\mu} p_X(\mu) = \mathrm{e}^{t\mu} \times 1 = \mathrm{e}^{t\mu}.$$

Hence:

$$K_X(t) = \log M_X(t) = \mu t.$$

Example 3.46 For the Poisson distribution, its cumulant generating function has a simpler functional form than its moment generating function. If $X \sim \text{Pois}(\lambda)$, its cumulant generating function is:

$$K_X(t) = \log M_X(t) = \log\left(\exp(\lambda(\mathrm{e}^t - 1))\right) = \lambda(\mathrm{e}^t - 1).$$

Applying the series expansion of e^t, we have:

$$K_X(t) = \lambda\left(t + \frac{t^2}{2!} + \cdots\right) = \sum_{i=1}^{\infty} \frac{\lambda t^i}{i!}.$$

Comparing coefficients, the rth cumulant is λ, i.e. $\kappa_r = \lambda$ for all $r = 1, 2, \ldots$.

Example 3.47 Suppose $Z \sim N(0, 1)$. From Example 3.42 we have $M_Z(t) = \mathrm{e}^{t^2/2}$, hence taking the logarithm yields the cumulant generating function:

$$K_Z(t) = \log M_Z(t) = \log \mathrm{e}^{t^2/2} = \frac{t^2}{2}.$$

Hence for a standard normal random variable, $\kappa_2 = 1$ and all other cumulants are zero.

Activity 3.12 Suppose $X \sim \text{Exp}(\lambda)$. Use the moment generating function of X to show that $\mathrm{E}(X^r) = r!/\lambda^r$.

Activity 3.13 For each of the following distributions derive the moment generating function and the cumulant generating function:

(a) Bernoulli(π)

(b) Bin(n, π)

(c) Geometric(π), first version

(d) Neg. Bin(r, π), first version.

Comment on any relationships found.

Activity 3.14 Use cumulants to calculate the coefficient of skewness for a Poisson distribution.

3.11 Functions of random variables

A well-behaved function, $g : \mathbb{R} \to \mathbb{R}$ of a random variable X is also a random variable. So, if $Y = g(X)$, then Y is a random variable. While we have seen how to determine the

expectation of functions of a random variable (Section 3.9.2), in matters of statistical inference we often need to know the actual distribution of $g(X)$, not just its expectation. We now proceed to show how to determine the distribution.

3.11.1 Distribution, mass and density functions of $Y = g(X)$

Let X be a random variable defined on $(\Omega, \mathcal{F}, P)$ and let $g : \mathbb{R} \to \mathbb{R}$ be a well-behaved function. Suppose $Y = g(X)$. While g is a function, and hence each input has a single output, g^{-1} is not guaranteed to be a function as a single input could produce multiple outputs. An obvious example is $g(x) = x^2$ such that $g^{-1}(x^2) = \pm\sqrt{x}$ with positive and negative roots, hence there is not a single output for $x^2 \neq 0$. In general:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) \neq P(X \leq g^{-1}(y)).$$

How to proceed? Well, we begin with the concept of the inverse image of a set.

Inverse image of a set

If $g : \mathbb{R} \to \mathbb{R}$ is a function and B is a subset of $\mathbb{R}$, then the **inverse image** of B under g is the set of real numbers whose images under g lie in B, i.e. for all $B \subseteq \mathbb{R}$ the inverse image of B under g is:

$$g^{-1}(B) = \{x \in \mathbb{R} : g(x) \in B\}.$$

So the inverse image of B under g is the image of B under g^{-1}. Hence for any well-behaved $B \subseteq \mathbb{R}$, we have:

$$P(Y \in B) = P(g(X) \in B) = P(X \in g^{-1}(B))$$

that is, the probability that $g(X)$ is in B equals the probability that X is in the inverse image of B.

We can now derive the distribution function of $Y = g(X)$.

Distribution function of $Y = g(X)$

Suppose $Y = g(X)$. The distribution function of Y is:

$$F_Y(y) = P(Y \leq y) = P(Y \in (-\infty, y]) = P(g(X) \in (-\infty, y]) = P(X \in g^{-1}((-\infty, y])).$$

Hence:

$$F_Y(y) = \begin{cases} \sum\limits_{\{x:g(x)\leq y\}} p_X(x) & \text{for discrete } X \\ \int_{\{x:g(x)\leq y\}} f_X(x)\,\mathrm{d}x & \text{for continuous } X. \end{cases}$$

Probability mass and density functions of $Y = g(X)$

For the discrete case, the probability mass function of Y is simply:

$$p_Y(y) = P(Y = y) = P(g(X) = y) = P(X \in g^{-1}(y)) = \sum_{\{x:g(x)=y\}} p_X(x).$$

For the continuous case, the probability density function of Y is simply:

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y).$$

Example 3.48 Let X be a random variable with continuous cdf F_X. Find expressions for the cdf of the following random variables.

(a) X^2

(b) $\sqrt{X}$

(c) $G^{-1}(X)$

(d) $G^{-1}(F_X(X))$

where G is continuous and strictly increasing.

Solution

(a) If $y \geq 0$, then:

$$P(X^2 \leq y) = P(X \leq \sqrt{y}) - P(X < -\sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

(b) We must assume $X \geq 0$. If $y \geq 0$, then:

$$P(\sqrt{X} \leq y) = P(0 \leq X \leq y^2) = F_X(y^2).$$

(c) We have:

$$P(G^{-1}(X) \leq y) = P(X \leq G(y)) = F_X(G(y)).$$

(d) We have:

$$\begin{aligned} P(G^{-1}(F_X(X)) \leq y) = P(F_X(X) \leq G(y)) &= P(X \leq F_X^{-1}(G(y))) \\ &= F_X(F_X^{-1}(G(y))) \\ &= G(y). \end{aligned}$$

Example 3.49 Suppose $X \sim \mathrm{Bin}(n, \pi)$ and hence X is the total number of *successes* in n independent Bernoulli(π) trials. Hence $Y = n - X$ is the total number of *failures*. We seek the distribution of Y.

We have $Y = g(X)$ such that $g(x) = n - x$. Hence:

$$\begin{aligned} F_Y(y) &= \sum_{\{x:g(x)\leq y\}} p_X(x) \\ &= \sum_{x=n-y}^{n} \binom{n}{x} \pi^x (1-\pi)^{n-x} && \text{(note the limits of } x\text{)} \\ &= \sum_{i=0}^{y} \binom{n}{n-i} \pi^{n-i} (1-\pi)^i && \text{(setting } i = n-x\text{)} \\ &= \sum_{i=0}^{y} \binom{n}{i} (1-\pi)^i \pi^{n-i}. && \text{(since } \tbinom{n}{n-i} = \tbinom{n}{i}\text{)} \end{aligned}$$

Note that this is the mass function of a $\text{Bin}(n, 1-\pi)$ distribution, hence by symmetry $Y \sim \text{Bin}(n, 1-\pi)$, as we would expect.

Example 3.50 Let X be a continuous random variable and suppose $Y = X^2$. We seek the distribution function and density function of Y. We have:

$$\begin{aligned} F_Y(y) = P(Y \leq y) = P(X^2 \leq y) &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

Noting that the support must be for $\{y : y \geq 0\}$, in full the distribution function of Y is:

$$F_Y(y) = \begin{cases} F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \text{for } y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Differentiating, we obtain the density function of Y, noting the application of the chain rule:

$$f_Y(y) = \begin{cases} \dfrac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}} & \text{for } y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Example 3.51 Let X be a continuous random variable with cdf $F_X(x)$. Determine the distribution of $Y = F_X(X)$. What do you observe?

Solution

For $0 \leq y \leq 1$ we have:

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= P(F_X(X) \leq y) \\ &= P(X \leq F_X^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) \\ &= y. \end{aligned}$$

Hence the density function of Y is:

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = 1$$

for $0 \leq y \leq 1$, and 0 otherwise. Therefore, $Y \sim \text{Uniform}[0, 1]$.

Example 3.52 We apply the density function result in Example 3.50 to the case where $X \sim N(0, 1)$, i.e. the standard normal distribution. Since the support of X is $\mathbb{R}$, the support of $Y = X^2$ is the positive real line. The density function of X is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \quad \text{for} -\infty < x < \infty.$$

Therefore, setting $y = x^2$, we have:

$$f_Y(y) = \frac{1}{2\sqrt{y}} \left(\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y/2} + \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y/2} \right) = \frac{1}{\sqrt{\pi}} \left(\frac{1}{2} \right)^{1/2} y^{-1/2} \mathrm{e}^{-y/2}.$$

Noting that $\Gamma(1/2) = \sqrt{\pi}$, the density function is:

$$f_Y(y) = \begin{cases} \dfrac{1}{\Gamma(1/2)} \left(\dfrac{1}{2} \right)^{1/2} y^{1/2-1} \mathrm{e}^{-y/2} & \text{for } y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note that this is the density function of a $\text{Gamma}(1/2, 1/2)$ distribution, hence if $X \sim N(0, 1)$ and $Y = X^2$, then $Y \sim \text{Gamma}(1/2, 1/2)$.

In passing we also note, from **ST104b Statistics 2**, that the square of a standard normal random variable has a chi-squared distribution with 1 degree of freedom, i.e. χ^2_1, hence it is also true that $Y \sim \chi^2_1$ and so we can see that the chi-squared distribution is a special case of the gamma distribution – there are many relationships between the various families of distributions!

Example 3.53 Suppose that X is a continuous random variable taking values between $-\infty$ and ∞ with distribution function $F_X(x)$. Sometimes we want to *fold* the distribution of X about the value $x = a$, that is we want the distribution function $F_Y(y)$ of the random variable Y obtained from X by taking $Y = X - a$, for $X > a$, and $Y = a - X$, for $X < a$ (in other words $Y = |X - a|$). Find $F_Y(y)$ by working out directly $P(Y \leq y)$. What is the density function of Y?

A particularly important application is the case when X has a $N(\mu, \sigma^2)$ distribution, and $a = \mu$, which has pdf:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-a)^2}{2\sigma^2} \right).$$

Apply the result to this case.

Solution

The description of Y says that $Y = |X - a|$, hence:

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= P(a - y \leq X \leq a + y) \\ &= P(a - y < X \leq a + y) \\ &= \begin{cases} F_X(a + y) - F_X(a - y) & \text{for } y \geq 0 \\ 0 & \text{for } y < 0. \end{cases} \end{aligned}$$

The density function $f_Y(y)$ is obtained by differentiating with respect to y, hence:

$$f_Y(y) = \begin{cases} f_X(a + y) + f_X(a - y) & \text{for } y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

In the case where $X \sim N(\mu, \sigma^2)$ and $a = \mu$, the density function of Y is:

$$f_Y(y) = \begin{cases} \dfrac{\sqrt{2}}{\sqrt{\pi\sigma^2}} \exp\left(-\dfrac{y^2}{2\sigma^2}\right) & \text{for } y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

This is sometimes called a *half-normal distribution.*

Activity 3.15 Let X be a continuous random variable with a support of $\mathbb{R}$. Determine the density function of $|X|$ in terms of f_X.

Activity 3.16 Let $X \sim N(\mu, \sigma^2)$. Determine the density function of $|X - \mu|$.

3.11.2 Monotone functions of random variables

We now apply the material from Section 3.11.1 to the case when the function g is monotone.

Monotone function

A function $g : \mathbb{R} \to \mathbb{R}$ is **monotone** in both of the following cases.

i. g is **monotone increasing** if:

$$g(x_1) \leq g(x_2) \quad \text{for all } x_1 < x_2$$

ii. g **monotone decreasing** if:

$$g(x_1) \geq g(x_2) \quad \text{for all } x_1 < x_2.$$

Strict monotonicity replaces the above inequalities with strict inequalities. For a strictly monotone function, the inverse image of an interval is also an interval.

Distribution function of $Y = g(X)$ when g is strictly monotone

If X is a random variable, $g : \mathbb{R} \to \mathbb{R}$ is strictly monotone and $Y = g(X)$, then the distribution function of Y is:

$$F_Y(y) = P(X \in g^{-1}((-\infty, y])) = \begin{cases} P(X \leq g^{-1}(y)) & \text{for } g \text{ increasing} \\ P(X \geq g^{-1}(y)) & \text{for } g \text{ decreasing} \end{cases}$$

$$= \begin{cases} F_X(g^{-1}(y)) & \text{for } g \text{ increasing} \\ 1 - F_X(g^{-1}(y)-) & \text{for } g \text{ decreasing} \end{cases} \tag{3.5}$$

where y is in the range of g and:

$$F_X(g^{-1}(y)-) = \begin{cases} F_X(g^{-1}(y)) - P(X = g^{-1}(y)) & \text{for discrete } X \\ F_X(g^{-1}(y)) & \text{for continuous } X. \end{cases}$$

Density function of $Y = g(X)$ when g is monotone

If X is a continuous random variable, $g : \mathbb{R} \to \mathbb{R}$ is monotone and $Y = g(X)$, then the density function of Y is:

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \dfrac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right| & \text{for } y \text{ in the range of } g \\ 0 & \text{otherwise.} \end{cases}$$

Proof: Let X be a random variable with density function $f_X(x)$ and let $Y = g(X)$, i.e. $X = g^{-1}(Y)$.

If $g^{-1}(\cdot)$ is increasing, then:

$$F_Y(y) = P(Y \leq y) = P(g^{-1}(Y) \leq g^{-1}(y)) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

hence:

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y).$$

If $g^{-1}(\cdot)$ is decreasing, then:

$$F_Y(y) = P(Y \leq y) = P(g^{-1}(Y) \geq g^{-1}(y)) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$

hence:

$$f_Y(y) = -\frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \left(-\frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \right).$$

Recall that the derivative of a decreasing function is negative.

Combining both cases, if $g^{-1}(\cdot)$ is monotone, then:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right|.$$

■

Example 3.54 For any continuous random variable X we consider, its distribution function, F_X, is strictly increasing over its support, say $S \subseteq \mathbb{R}$. The inverse function, F_X^{-1}, known as the *quantile function*, is strictly increasing on $[0,1]$, such that $F_X^{-1} : [0,1] \to S$.

Let $U \sim \text{Uniform}[0,1]$, hence its distribution function is:

$$F_U(u) = \begin{cases} 0 & \text{for } u < 0 \\ u & \text{for } 0 \leq u \leq 1 \\ 1 & \text{for } u > 1. \end{cases}$$

Let $X = F_X^{-1}(U)$, hence its distribution function is:

$$F_X(x) = F_U((F_X^{-1})^{-1}(x)) \quad \text{for } x \in S$$

which can be used to simulate random samples from a required distribution by simulating values of u from Uniform$[0,1]$, and then view $F_X^{-1}(u)$ as a random drawing from F_X.

Suppose $X \sim \text{Exp}(\lambda)$, hence its distribution function is:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - \mathrm{e}^{-\lambda x} & \text{for } x \geq 0. \end{cases}$$

The quantile function (i.e. the inverse distribution function) is:

$$F_X^{-1}(u) = \frac{1}{\lambda} \log\left(\frac{1}{1-u}\right) \quad \text{for } 0 \leq u \leq 1.$$

Therefore, if $U \sim \text{Uniform}[0,1]$, then:

$$\frac{1}{\lambda} \log\left(\frac{1}{1-U}\right) \sim \text{Exp}(\lambda).$$

Example 3.55 Let $X \sim \text{Gamma}(\alpha, \beta)$, hence:

$$f_X(x) = \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} \mathrm{e}^{-\beta x}$$

for $x \geq 0$, and 0 otherwise.

We seek $Y = 1/X$. We have that:

$$g^{-1}(y) = \frac{1}{y} \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) = -\frac{1}{y^2} < 0.$$

Therefore:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right| = \frac{1}{\Gamma(\alpha)} \beta^\alpha (1/y)^{\alpha-1} \mathrm{e}^{-\beta/y} \frac{1}{y^2}$$

$$= \frac{1}{\Gamma(\alpha)} \beta^\alpha y^{-\alpha-1} \mathrm{e}^{-\beta/y}$$

for $y > 0$, and 0 otherwise. This is the inverse gamma distribution.

Example 3.56 Suppose X has the Weibull distribution:

$$f_X(x) = c\tau x^{\tau-1}\mathrm{e}^{-cx^\tau}$$

for $x \geq 0$, where $c, \tau > 0$ are constants. What is the density function of $Y = cX^\tau$?

Solution

The transformation $y = cx^\tau$ is strictly increasing on $[0, \infty)$. The inverse transformation is $x = (y/c)^{1/\tau}$. It follows that for $y \geq 0$ we have:

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= P\left(\left(\frac{Y}{c}\right)^{1/\tau} \leq \left(\frac{y}{c}\right)^{1/\tau}\right) \\ &= P\left(X \leq \left(\frac{y}{c}\right)^{1/\tau}\right) \\ &= F_X\left(\left(\frac{y}{c}\right)^{1/\tau}\right). \end{aligned}$$

Differentiating with respect to y, for $y \geq 0$ we have:

$$\begin{aligned} f_Y(y) &= f_X\left(\left(\frac{y}{c}\right)^{1/\tau}\right)\frac{1}{\tau}\left(\frac{y}{c}\right)^{1/\tau-1}\frac{1}{c} \\ &= c\tau\left(\left(\frac{y}{c}\right)^{1/\tau}\right)^{\tau-1}\mathrm{e}^{-y}\frac{1}{\tau}\left(\frac{y}{c}\right)^{1/\tau-1}\frac{1}{c} \\ &= \mathrm{e}^{-y}. \end{aligned}$$

This is an exponential distribution, specifically Exp(1).

Example 3.57 Standardisation and 'reverse' standardisation are examples of a *scale* or *location* transformation. A classic example is:

$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

or, in reverse:

$$Z \sim N(0,1) \quad \Rightarrow \quad X = \mu + \sigma Z \sim N(\mu, \sigma^2).$$

In fact, the distributional assumption of normality is *not* essential for standardisation (or its reverse). Let Z be any random variable, and g be the linear function:

$$g(z) = \mu + \sigma z$$

for $\mu \in \mathbb{R}$ and $\sigma > 0$. If $X = g(Z)$, i.e. $X = \mu + \sigma Z$, then X is a linear transformation of Z, being related through a scale (or location) transformation. This means properties of X can be derived from properties of Z.

The distribution function of X is:

$$F_X(x) = P(X \leq x) = P(\mu + \sigma Z \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = F_Z\left(\frac{x-\mu}{\sigma}\right).$$

If Z is a continuous random variable, then the density function of X is (applying the chain rule):

$$f_X(x) = \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right).$$

The moment generating functions of X and Z are also related, since:

$$M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \mathrm{E}(\mathrm{e}^{t(\mu+\sigma Z)}) = \mathrm{E}(\mathrm{e}^{\mu t}\mathrm{e}^{\sigma t Z}) = \mathrm{e}^{\mu t}\mathrm{E}(\mathrm{e}^{\sigma t Z}) = \mathrm{e}^{\mu t} M_Z(\sigma t).$$

The cumulant generating functions are also related, as seen by taking logarithms:

$$K_X(t) = \mu t + K_Z(\sigma t).$$

Cumulants of X and Z are hence related, as seen by comparing coefficients:

$$\kappa_{X,r} = \begin{cases} \mu + \sigma\kappa_{Z,1} & \text{for } r = 1 \\ \sigma^r \kappa_{Z,r} & \text{for } r = 2, 3, \ldots. \end{cases}$$

Therefore:

$$\mathrm{E}(X) = \mu + \sigma\,\mathrm{E}(Z) \quad \text{and} \quad \mathrm{Var}(X) = \sigma^2 \mathrm{Var}(Z).$$

If we impose the distributional assumption of normality, such that $Z \sim N(0,1)$, and continue to let $X = \mu + \sigma Z$, the density function of X is:

$$\begin{aligned} f_X(x) = \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right) &= \frac{1}{\sigma}\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

and hence $X \sim N(\mu, \sigma^2)$.

Recall that the cumulant generating function of the standard normal distribution (Example 3.47) is $K_Z(t) = t^2/2$, hence the cumulant generating function of X is:

$$K_X(t) = \mu t + K_Z(\sigma t) = \mu t + \frac{\sigma^2 t^2}{2}.$$

So if $X \sim N(\mu, \sigma^2)$, then:

$$K_X'(t) = \mu + \sigma^2 t \quad \text{and} \quad K_X''(t) = \sigma^2$$

hence $\kappa_1 = K_X'(0) = \mu$ and $\kappa_2 = K_X''(0) = \sigma^2$ (as always for the first two cumulants). However, we see that:

$$\kappa_r = 0 \quad \text{for } r > 2$$

and so by the one-to-one mapping (i.e. uniqueness) of a cumulant generating function to a probability distribution, *any* distribution for which $\kappa_r = 0$ for $r > 2$ is a normal distribution.

Example 3.58 Let $X \sim N(\mu, \sigma^2)$ and $Y = \mathrm{e}^X$. Find the density function of Y.

Solution

Let $Y = g(X) = \exp(X)$. Hence $X = g^{-1}(Y) = \ln(Y)$, and:

$$\frac{\mathrm{d}}{\mathrm{d}y}\, g^{-1}(y) = \frac{1}{y}.$$

Therefore:

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y}\, g^{-1}(y) \right| \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2} \right) \frac{1}{y}. \end{aligned}$$

Example 3.59 Suppose X has the density function (sometimes called a Type II Beta distribution):

$$f_X(x) = \frac{1}{B(\alpha, \beta)} \frac{x^{\beta-1}}{(1+x)^{\alpha+\beta}}$$

for $0 < x < 1$, where $\alpha, \beta > 0$ are constants. What is the density function of $Y = X/(1+X)$?

Solution

The transformation $y = x/(1+x)$ is strictly increasing on $(0, \infty)$, and has the unique inverse function $x = y/(1-y)$. Hence:

$$f_Y(y) = f_X(x) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right|.$$

Since $\mathrm{d}x/\mathrm{d}y = 1/(1-y)^2 > 0$, we have:

$$\begin{aligned} f_Y(y) &= f_X(x) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| \\ &= f_X\left(\frac{y}{1-y} \right) \frac{1}{(1-y)^2} \\ &= \frac{1}{B(\alpha, \beta)} \frac{(y/(1-y))^{\beta-1}}{(1+y/(1-y))^{\alpha+\beta}} \frac{1}{(1-y)^2} \\ &= \frac{1}{B(\alpha, \beta)} y^{\beta-1}(1-y)^{\alpha+1} \frac{1}{(1-y)^2} \\ &= \frac{1}{B(\alpha, \beta)} y^{\beta-1}(1-y)^{\alpha-1}. \end{aligned}$$

This is a beta distribution, i.e. $Y \sim \mathrm{Beta}(\alpha, \beta)$.

Example 3.60 Let $X \sim \text{Uniform}[0, 1]$. Suppose $Y = a + (b - a)X$.

(a) Determine the distribution of Y.

(b) Determine $\text{Var}(Y)$.

Solution

(a) Rearranging, we have $X = (Y - a)/(b - a)$, hence $\mathrm{d}x/\mathrm{d}y = 1/(b - a)$ so:

$$f_Y(y) = f_X\left(\frac{y-a}{b-a}\right)\left|\frac{1}{b-a}\right| = 1 \times \frac{1}{b-a} = \frac{1}{b-a}.$$

In full:

$$f_Y(y) = \begin{cases} 1/(b-a) & \text{for } a \leq y \leq b \\ 0 & \text{otherwise} \end{cases}$$

that is, $Y \sim \text{Uniform}[a, b]$.

(b) We have that $\text{Var}(X) = \text{E}(X^2) - (\text{E}(X))^2$, where:

$$\text{E}(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, \mathrm{d}x = \int_0^1 x\, \mathrm{d}x = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}$$

and:

$$\text{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\, \mathrm{d}x = \int_0^1 x^2\, \mathrm{d}x = \left[\frac{x^3}{3}\right]_0^1 = \frac{1}{3}$$

giving:

$$\text{Var}(X) = \text{E}(X^2) - (\text{E}(X))^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}.$$

Hence:

$$\text{Var}(Y) = \text{Var}(a + (b-a)X) = \text{Var}((b-a)X) = (b-a)^2\text{Var}(X) = \frac{(b-a)^2}{12}.$$

Activity 3.17 Let X be a positive, continuous random variable. Determine the density function of $1/X$ in terms of f_X.

Activity 3.18 Let $X \sim \text{Exp}(\lambda)$. Determine the density function of $1/X$.

3.12 Convergence of sequences of random variables

In this section we consider aspects related to convergence of sequences of random variables. However, we begin with the definition of convergence for a sequence of real numbers (that is, constants) $x_1, x_2, \ldots$ which we denote by $\{x_n\}$.

Convergence of a real sequence

If $\{x_n\}$ is a sequence of real numbers, then x_n **converges** to a real number x if and only if, for all $\epsilon > 0$, there exists an integer N such that:

$$|x_n - x| < \epsilon \quad \text{for all } n > N.$$

The convergence of a real sequence can be written as $x_n \to x$ as $n \to \infty$.

If we now consider a sequence of *random variables* rather than *constants*, i.e. $\{X_n\}$, in matters of convergence it does not make sense to compare $|X_n - X|$ (a random variable) to a constant $\epsilon > 0$. Below, we introduce four different types of convergence.

Convergence in distribution

A sequence of random variables $\{X_n\}$ **converges in distribution** if:

$$P(X_n \leq x) \to P(X \leq x) \quad \text{as } n \to \infty$$

equivalently:

$$F_{X_n}(x) \to F_X(x) \quad \text{as } n \to \infty$$

for all x at which the distribution function is continuous. This is denoted as:

$$X_n \xrightarrow{d} X.$$

Convergence in probability

A sequence of random variables $\{X_n\}$ **converges in probability** if, for any $\epsilon > 0$, we have:

$$P(|X_n - X| < \epsilon) \to 1 \quad \text{as } n \to \infty.$$

This is denoted as:

$$X_n \xrightarrow{p} X.$$

Convergence almost surely

A sequence of random variables $\{X_n\}$ **converges almost surely** to X if, for any $\epsilon > 0$, we have:

$$P\left(\lim_{n\to\infty} |X_n - X| < \epsilon\right) = 1.$$

This is denoted as:

$$X_n \xrightarrow{a.s.} X.$$

Convergence in mean square

A sequence of random variables $\{X_n\}$ **converges in mean square** to X if:

$$\mathrm{E}((X_n - X)^2) \to 0 \quad \text{as } n \to \infty.$$

This is denoted as:

$$X_n \xrightarrow{m.s.} X.$$

The above types of convergence differ in terms of their strength. If $\{X_n\}$ converges almost surely, then $\{X_n\}$ converges in probability:

$$X_n \xrightarrow{a.s.} X \quad \Rightarrow \quad X_n \xrightarrow{p} X.$$

If $\{X_n\}$ converges in mean square, then $\{X_n\}$ converges in probability:

$$X_n \xrightarrow{m.s.} X \quad \Rightarrow \quad X_n \xrightarrow{p} X.$$

If $\{X_n\}$ converges in probability, then $\{X_n\}$ converges in distribution:

$$X_n \xrightarrow{p} X \quad \Rightarrow \quad X_n \xrightarrow{d} X.$$

Combining these results, we can say that the set of all sequences which converge in distribution contains the set of all sequences which converge in probability, which in turn contains the set of all sequences which converge almost surely and in mean square. We may write this as:

$$\left.\begin{array}{l} X_n \xrightarrow{a.s.} X \\ X_n \xrightarrow{m.s.} X \end{array}\right\} \quad \Rightarrow \quad X_n \xrightarrow{p} X \quad \Rightarrow \quad X_n \xrightarrow{d} X.$$

Example 3.61 For a sequence of random variables $\{X_n\}$, we prove that convergence in mean square implies convergence in probability.

Consider $P(|X_n - X| > \epsilon)$ for $\epsilon > 0$. Applying the Chebyshev inequality, we have:

$$P(|X_n - X| > \epsilon) \leq \frac{\mathrm{E}((X_n - X)^2)}{\epsilon^2}.$$

If $X_n \xrightarrow{m.s.} X$, then $\mathrm{E}((X_n - X)^2) \to 0$. Therefore, $\{P(|X_n - X| > \epsilon)\}$ is a sequence of positive real numbers bounded above by a sequence which converges to zero. Hence we conclude $P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$, and so $X_n \xrightarrow{p} X$.

Example 3.62 For a sequence of random variables $\{X_n\}$, we prove that if $X_n \xrightarrow{d} a$, where a is a constant, then this implies $X_n \xrightarrow{p} a$.

Making use of the distribution function of the degenerate distribution (Section 3.7.1), we have:

$$\begin{aligned} P(|X_n - a| > \epsilon) &= P(X_n - a > \epsilon) + P(X_n - a < -\epsilon) \\ &= P(X_n > a + \epsilon) + P(X_n < a - \epsilon) \\ &\leq (1 - F_{X_n}(a + \epsilon)) + F_{X_n}(a - \epsilon). \end{aligned}$$

If $X_n \xrightarrow{d} a$, then $F_{X_n} \to F_a$ and hence $F_{X_n}(a+\epsilon) \to 1$ and $F_{X_n}(a-\epsilon) \to 0$. Therefore, $\{P(|X_n - a| > \epsilon)\}$ is a sequence of positive real numbers bounded above by a sequence which converges to zero. Hence we conclude $P(|X_n - a| > \epsilon) \to 0$ as $n \to \infty$ and hence $X_n \xrightarrow{p} a$.

Example 3.63 In the following two cases explain in which, if any, of the three modes (in mean square, in probability, in distribution) X_n converges to 0.

(a) Let $X_n = 1$ with probability 2^{-n} and 0 otherwise.

(b) Let $X_n = n$ with probability n^{-1} and 0 otherwise, and assume that the X_ns are independent.

Solution

(a) We have $P(X_n = 1) = 2^{-n}$ and $P(X_n = 0) = 1 - 2^{-n}$. Since:

$$\mathrm{E}(|X_n - 0|^2) = \mathrm{E}(X_n^2) = 0^2 \times P(X_n = 0) + 1^2 \times P(X_n = 1) = 2^{-n} \to 0$$

as $n \to \infty$, then $X_n \xrightarrow{m.s.} 0$. Hence also $X_n \xrightarrow{p} 0$ and $X_n \xrightarrow{d} 0$.

(b) We have $P(X_n = n) = n^{-1}$ and $P(X_n = 0) = 1 - n^{-1}$. Since:

$$\mathrm{E}(|X_n - 0|^2) = \mathrm{E}(X_n^2) = 0^2 \times P(X_n = 0) + n^2 \times P(X_n = n) = n \to \infty$$

as $n \to \infty$, then X_n is not mean square convergent. For all $\epsilon > 0$ we have:

$$P(|X_n - 0| > \epsilon) = 0 = P(X_n > \epsilon) = P(X_n = n) = n^{-1} \to 0$$

as $n \to \infty$. Therefore, $X_n \xrightarrow{p} 0$ and $X_n \xrightarrow{d} 0$.

Activity 3.19 Consider a sequence of random variables $\{X_n\}$ with cumulant generating functions $\{K_{X_n}\}$ and a random variable X with cumulant generating function $K_X(t)$. Suppose, in addition, that all these cumulant generating functions are well-defined for $|t| < a$. If $K_{X_n} \to K_X(t)$ as $n \to \infty$ for all t such that $|t| < a$, what can we conclude?

Activity 3.20 Consider a sequence of random variables $\{X_n\}$ and constant a. Prove that $X_n \xrightarrow{p} c$ implies $X_n \xrightarrow{d} a$.

3.13 A reminder of your learning outcomes

On completion of this chapter, you should be able to:

- provide both formal and informal definitions of a random variable

- formulate problems in terms of random variables
- explain the characteristics of distribution functions
- explain the distinction between discrete and continuous random variables
- provide the probability mass function (pmf) and support for some common discrete distributions
- provide the probability density function (pdf) and support for some common continuous distributions
- explain whether a function defines a valid mass or density
- calculate moments for discrete and continuous distributions
- prove and manipulate inequalities involving the expectation operator
- derive moment generating functions for discrete and continuous distributions
- calculate moments from a moment generating function
- calculate cumulants from a cumulant generating function
- determine the distribution of a function of a random variable
- summarise scale/location and probability integral transformations.

3.14 Sample examination questions

Solutions can be found in Appendix C.

1. Let X be a discrete random variable with mass function defined by:

$$p_X(x) = \begin{cases} k^x & \text{for } x = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

where k is a constant with $0 < k < 1$.

(a) Show that $k = 1/2$.

(b) Show that the distribution function $F_X(x)$ of X has values at $x = 1, 2, \ldots$ given by:

$$F_X(x) = 1 - \left(\frac{1}{2}\right)^x \quad \text{for } x = 1, 2, \ldots.$$

(c) Show that the moment generating function of X is given by:

$$M_X(t) = \frac{\mathrm{e}^t}{2 - \mathrm{e}^t} \quad \text{for } t < \log 2.$$

Hence find $\mathrm{E}(X)$.

2. (a) Let X be a positive random variable with $\mathrm{E}(X) < \infty$. Prove the Markov inequality:
$$P(X \geq a) \leq \frac{\mathrm{E}(X)}{a}$$
for any constant $a > 0$.

   (b) For a random variable X with $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, state Chebyshev's inequality.

   (c) By considering an Exp(1) random variable show, for $0 < a < 1$, that:
$$a\mathrm{e}^{-a} \leq 1 \quad \text{and} \quad a^2(1 + \mathrm{e}^{-(1+a)} - \mathrm{e}^{-(1-a)}) \leq 1.$$
You can use the mean and variance of an exponential random variable without proof, as long as they are stated clearly.

3. Let X_1 and X_2 be independent continuous uniform random variables, with X_1 defined over $[-1, 1]$ and X_2 defined over $[-2, 1]$.

   (a) Show, by considering $P(W_1 < w)$, that the density function of $W_1 = X_1^2$ is:
$$f_{W_1}(w) = \begin{cases} \dfrac{1}{2\sqrt{w}} & \text{for } 0 \leq w \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

   (b) Show that the density function of $W_2 = X_2^2$ is:
$$f_{W_2}(w) = \begin{cases} \dfrac{1}{3\sqrt{w}} & \text{for } 0 \leq w \leq 1 \\ \dfrac{1}{6\sqrt{w}} & \text{for } 1 \leq w < 4 \\ 0 & \text{otherwise.} \end{cases}$$
(Hint: Consider $P(W_2 < w)$ and $P(1 \leq W_2 < w)$ for $0 < w < 1$ and $1 \leq w < 4$, respectively.)

   (c) Show that the density function of $Y = \sqrt{W_2}$ is:
$$f_Y(y) = \begin{cases} 2/3 & \text{for } 0 \leq y < 1 \\ 1/3 & \text{for } 1 \leq y < 2 \\ 0 & \text{otherwise.} \end{cases}$$

Chapter 4
Multivariate distributions

4.1 Recommended reading

Casella, G. and R.L. Berger, *Statistical Inference*. Chapter 4, Sections 4.1, 4.3, 4.5 and 4.6.

4.2 Learning outcomes

On completion of this chapter, you should be able to:

- define the terms 'joint' and 'marginal'
- generate the marginal distribution function of a random variable from a joint distribution function involving the variable
- construct a table providing the joint and marginal mass functions for a discrete bivariate distribution
- determine whether a function of two random variables is a valid mass or density function
- derive the marginal mass or density function from a joint mass or density function (even when the support is an 'awkward' shape)
- calculate probabilities from a bivariate mass or density function
- calculate expectations for functions of several variables
- calculate joint moments from a bivariate mass or density function
- explain the distinction between pairwise and mutual independence
- explain the distinction between uncorrelated and independent random variables
- determine that the variance matrix of a random vector is non-negative definite
- derive the density function of a transformation of continuous random variables
- derive the distribution of the sum of random variables
- construct a standard bivariate normal distribution
- derive the density function of a linear transformation of independent normal random variables.

4.3 Introduction

We extend our previous work on univariate distributions to consider how to model multiple random variables using a multivariate distribution. For example, if we collected data on 10 attributes of an individual – such as height, weight, age, income etc. – the observations would be recorded as $x_1, x_2, \ldots, x_{10}$, which can be modelled as realisations of the random variables $X_1, X_2, \ldots, X_{10}$. In practice, the random variables would not be expected to be independent of each other (for example, height and weight would likely be positively correlated), and so we use multivariate distributions to capture any dependencies between variables. Before proceeding, you may wish to review Chapter 5 of **ST104b Statistics 2** which introduced many of the concepts presented here for the case of discrete bivariate distributions.

4.4 Bivariate joint and marginal distributions

We begin with definitions for the bivariate case, when $n = 2$, and then proceed to give general definitions for the n-dimensional case (for $n \geq 2$).

Joint cumulative distribution function (cdf)

Let X and Y be two random variables. The **joint cumulative distribution function** is a function $F_{X,Y} : \mathbb{R} \times \mathbb{R} \to [0, 1]$ such that:

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y).$$

The joint cdf has the following properties.

1. $F_{X,Y}(-\infty, y) = \lim_{x \to -\infty} F_{X,Y}(x, y) = 0.$
   $F_{X,Y}(x, -\infty) = \lim_{y \to -\infty} F_{X,Y}(x, y) = 0.$
   $F_{X,Y}(\infty, \infty) = \lim_{x \to \infty, y \to \infty} F_{X,Y}(x, y) = 1.$
2. Right continuous in x: $\lim_{h \to 0^+} F_{X,Y}(x + h, y) = F_{X,Y}(x, y).$
   Right continuous in y: $\lim_{h \to 0^+} F_{X,Y}(x, y + h) = F_{X,Y}(x, y).$
3. For any y, the function $F_{X,Y}(x, y)$ is non-decreasing in x.
   For any x, the function $F_{X,Y}(x, y)$ is non-decreasing in y.

Our interest lies in the probability that X and Y take values in a particular (Borel) subset of the plane $\mathbb{R} \times \mathbb{R} \equiv \mathbb{R}^2$. The simplest subset to consider is the rectangular region $A = \{(x, y) \in \mathbb{R}^2 : x_1 < x \leq x_2 \text{ and } y_1 < y \leq y_2\}$. Hence:

$$\begin{aligned} P(A) &= P(x_1 < X \leq x_2, y_1 < Y \leq y_2) \\ &= F_{X,Y}(x_2, y_2) - F_{X,Y}(x_1, y_2) - (F_{X,Y}(x_2, y_1) - F_{X,Y}(x_1, y_1)). \end{aligned}$$

Marginal cumulative distribution functions

Let $F_{X,Y}$ be the joint cdf of the random variables X and Y. The **marginal cdfs** are then the cdfs of the individual random variables X and Y, such that:

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y) = F_{X,Y}(x,\infty) \quad \text{and} \quad F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x,y) = F_{X,Y}(\infty,y).$$

Note that the marginal cdfs are derived from the joint cdf, but we cannot derive the joint cdf from the marginal cdfs. This is because the joint cdf comprises additional information beyond that contained in the marginal cdfs; specifically information about the *association* between the random variables, i.e. information about their *dependence structure*.

Activity 4.1 Provide a general, i.e. n-variable, version of the properties of bivariate joint distribution functions.

Activity 4.2 If $F_X(x)$ and $F_Y(y)$ are distribution functions of random variables X and Y, for which of the following definitions can $G(x,y)$ be a valid joint distribution function?

(Recall that $\max(a,b)$ and $\min(a,b)$ mean the larger of a and b and the smaller of a and b, respectively.)

(a) $G(x,y) = F_X(x) + F_Y(y)$.

(b) $G(x,y) = F_X(x)\,F_Y(y)$.

(c) $G(x,y) = \max(F_X(x), F_Y(y))$.

(d) $G(x,y) = \min(F_X(x), F_X(y))$.

4.4.1 Bivariate joint and marginal mass functions

Joint probability mass function

Let X and Y be two discrete random variables. Their **joint probability mass function** is a function $p_{X,Y} : \mathbb{R}\times\mathbb{R} \to [0,1]$ such that:

$$p_{X,Y}(x,y) = P(X = x, Y = y) \quad \forall\, x,y \in \mathbb{R}.$$

In general:

$$P(x_1 < X \le x_2, y_1 < Y \le y_2) = \sum_{x_1<x\le x_2} \sum_{y_1<y\le y_2} p_{X,Y}(x,y).$$

Marginal probability mass function

Let X and Y be two discrete random variables with the ranges $\{x_1, x_2, \ldots\}$ and $\{y_1, y_2, \ldots\}$, respectively. The **marginal probability mass functions** are, respectively:

$$p_X(x) = \sum_{y \in \{y_1, y_2, \ldots\}} p_{X,Y}(x, y) \quad \text{and} \quad p_Y(y) = \sum_{x \in \{x_1, x_2, \ldots\}} p_{X,Y}(x, y).$$

4.4.2 Bivariate joint and marginal density functions

Joint probability density function

Let X and Y be two jointly continuous random variables. The **joint probability density function** is an integrable function $f_{X,Y} : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ such that:

$$F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v) \, \mathrm{d}u \, \mathrm{d}v \quad \forall \; x, y \in \mathbb{R}.$$

Note that the joint density function can be written as:

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(u, v)}{\partial u \, \partial v}\Big|_{u=x, v=y}.$$

The joint pdf has the following properties.

1. $f_{X,Y}(x, y) \geq 0$ for any $x, y \in \mathbb{R}$.
2. Normalisation, such that:
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y = 1.$$
3. The probability of a rectangular region is:
$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y.$$
4. For any (Borel) set $B \subseteq \mathbb{R}^2$, the probability that (X, Y) takes values in B is:
$$P(B) = \iint_B f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y.$$

A word on probabilities

Probability is the measure of events with respect to the measure of the entire sample space, which is 1 by definition.

For the one-dimensional case, events are typically intervals of $\mathbb{R}$ such that their probabilities are proportional to their *length*. For the two-dimensional case, events are regions in the plane $\mathbb{R}^2$ such that their probabilities are proportional to their *area*. For the three-dimensional case, events are regions in the space $\mathbb{R}^3$ such that their probabilities are proportional to their *volume*.

Lengths, areas and volumes are weighted by the frequencies of the outcomes which are part of the events under consideration, hence these are areas, volumes and four-dimensional volumes under the respective density functions.

Marginal probability density function

Let X and Y be two jointly continuous random variables. These have **marginal probability density functions** which are integrable functions $f_X : \mathbb{R} \to [0, \infty)$ and $f_Y : \mathbb{R} \to [0, \infty)$ such that:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}y \quad \forall\, x \in \mathbb{R}$$

and:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}x \quad \forall\, y \in \mathbb{R}.$$

Consequently, the respective marginal cdfs may be obtained as:

$$F_X(x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(u, y)\, \mathrm{d}y\, \mathrm{d}u \quad \forall\, x \in \mathbb{R}$$

and:

$$F_Y(y) = \int_{-\infty}^{y} \int_{-\infty}^{\infty} f_{X,Y}(x, v)\, \mathrm{d}x\, \mathrm{d}v \quad \forall\, y \in \mathbb{R}.$$

Example 4.1 Consider the function:

$$f_{X,Y}(x, y) = \begin{cases} k(x^2 + y^2) & \text{for } 0 < x < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Under which conditions is $f_{X,Y}(x, y)$ a valid joint probability density function for X and Y?

(b) Compute $P(X < Y^2)$.

Solution

(a) We need two conditions to hold in order to have a valid joint pdf:

i. $f(x, y) \geq 0$

ii. the integral must be 1.

The first holds in this case for any $k \geq 0$. The second condition gives us the value of k. We have:

$$\begin{aligned}\int_0^1 \int_0^y k(x^2+y^2)\,\mathrm{d}x\,\mathrm{d}y = \int_0^1 k\left[\frac{x^3}{3}+xy^2\right]_0^y \mathrm{d}y &= k\int_0^1 \frac{4y^3}{3}\,\mathrm{d}y \\ &= \frac{4k}{3}\left[\frac{y^4}{4}\right]_0^1 \\ &= \frac{k}{3}\end{aligned}$$

hence $k = 3$.

(b) $P(X < Y^2)$ is computed as $\iint_A f_{X,Y}\,\mathrm{d}x\,\mathrm{d}y$, where:

$$A = \{(x,y) \in \mathbb{R}^2 \,:\, 0 < x < y^2,\ 0 < y < 1\}$$

hence:

$$\begin{aligned}\int_0^1 \int_0^{y^2} 3(x^2+y^2)\,\mathrm{d}x\,\mathrm{d}y = \int_0^1 \left[x^3+3y^2x\right]_0^{y^2} \mathrm{d}y &= \int_0^1 (y^6+3y^4)\,\mathrm{d}y \\ &= \left[\frac{y^7}{7}+\frac{3y^5}{5}\right]_0^1 \\ &= \frac{26}{35}.\end{aligned}$$

Example 4.2 For the following joint probability density function:

$$f_{X,Y}(x,y) = \begin{cases} k\mathrm{e}^{-\alpha x}\mathrm{e}^{-\beta y} & \text{for } 0 < x < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

determine the value of k.

Solution

To determine the value of k we integrate over the whole support. We have:

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = k\int_0^{\infty}\int_0^y \mathrm{e}^{-\alpha x}\mathrm{e}^{-\beta y}\,\mathrm{d}x\,\mathrm{d}y &= k\int_0^{\infty}\left[-\frac{1}{\alpha}\,\mathrm{e}^{-\alpha x}\mathrm{e}^{-\beta y}\right]_0^y \mathrm{d}y \\ &= \frac{k}{\alpha}\int_0^{\infty}(\mathrm{e}^{-\beta y}-\mathrm{e}^{-(\alpha+\beta)y})\,\mathrm{d}y \\ &= \frac{k}{\alpha}\left[-\frac{1}{\beta}\,\mathrm{e}^{-\beta y}+\frac{1}{\alpha+\beta}\,\mathrm{e}^{-(\alpha+\beta)y}\right]_0^{\infty} \\ &= \frac{k}{\alpha}\left(\frac{1}{\beta}-\frac{1}{\alpha+\beta}\right) \\ &= \frac{k}{\beta(\alpha+\beta)}\end{aligned}$$

hence $k = \beta(\alpha + \beta)$.

Activity 4.3 A fair coin is tossed three times. Let X be the number of heads and let Y be a random variable which is 1 if a head occurs on the first and third tosses, and 0 otherwise.

(a) Write down a table summarising the joint mass function of X and Y.

(b) Determine the marginal mass functions. Are they as might be expected?

Activity 4.4 Suppose X and Y have the joint density function:

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

(a) Determine the value of k.

(b) Find the marginal density functions $f_X(x)$ and $f_Y(y)$.

(c) Calculate $P(X < Y)$.

Activity 4.5 Suppose X and Y have the joint density function:

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

(a) Show that $f_{X,Y}$ is a valid joint density function.

(b) Find the marginal density functions $f_X(x)$ and $f_Y(y)$.

(c) Calculate $P(Y < X + 1/2)$.

4.4.3 Independence of two random variables

In the univariate setting we are interested in key random variable attributes such as *location* (i.e. the mean, $\mathrm{E}(X)$) and *scale* (i.e. the variance, $\mathrm{Var}(X)$). However, in the multivariate setting we are also interested in measuring the *dependence* between the random variables.

Joint cdf of independent random variables

Two random variables X and Y are **independent**, denoted $X \perp\!\!\!\perp Y$, if and only if the events $\{X \le x\}$ and $\{Y \le y\}$ are independent for all $x, y \in \mathbb{R}$. That is:

$$P(X \le x, Y \le y) = P(X \le x)\, P(Y \le y) \quad \Leftrightarrow \quad F_{X,Y}(x,y) = F_X(x)\, F_Y(y).$$

Joint pmf/pdf of independent random variables

Two random variables X and Y are independent if and only if for all $x, y \in \mathbb{R}$ we have:

$$p_{X,Y}(x,y) = p_X(x)\, p_Y(y) \quad \text{and} \quad f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$$

for the discrete and continuous cases, respectively.

The above joint cdf and pmf/pdf requirements are necessary and sufficient conditions for independence. The following result is a necessary (but not sufficient) condition.

Expectation and independence

If two random variables X and Y are independent, then:

$$\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y).$$

Furthermore, if g_1 and g_2 are well-behaved functions, then $g_1(X)$ and $g_2(Y)$ are also independent random variables, hence:

$$\mathrm{E}(g_1(X)\, g_2(Y)) = \mathrm{E}(g_1(X))\,\mathrm{E}(g_2(Y)).$$

4.5 Multivariate generalisations

4.5.1 n-variate cdf, pmf and pdf

We now consider a multivariate generalisation of the previous bivariate cases. Fortunately, for n random variables $X_1, X_2, \ldots, X_n$, we have analogous definitions as follows.

1. The joint cdf is a function $F_{X_1,X_2,\ldots,X_n} : \mathbb{R}^n \to [0,1]$ such that:

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n).$$

2. The marginal cdfs are the functions:

$$F_{X_j}(x_j) = F_{X_1,X_2,\ldots,X_n}(\infty, \ldots, \infty, x_j, \infty, \ldots, \infty) \quad \text{for any } j = 1, 2, \ldots, n.$$

3. The marginal pmf is:

$$p_{X_j}(x_j) = \sum_{x_1} \cdots \sum_{x_{j-1}} \sum_{x_{j+1}} \cdots \sum_{x_n} p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) \quad \text{for any } j = 1, 2, \ldots, n.$$

4. The marginal pdf is:

$$f_{X_j}(x_j) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_{j-1}\, \mathrm{d}x_{j+1} \cdots \mathrm{d}x_n$$

for any $j = 1, 2, \ldots, n$.

5. For a well-behaved function $g : \mathbb{R}^n \to \mathbb{R}$, then:

$$\mathrm{E}(g(X_1, X_2, \ldots, X_n)) = \begin{cases} \sum_{x_1} \cdots \sum_{x_n} g(x_1, x_2, \ldots, x_n)\, p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, x_2, \ldots, x_n)\, f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)\, \mathrm{d}x_1\, \mathrm{d}x_2 \cdots \mathrm{d}x_n \end{cases}$$

for the discrete and continuous cases, respectively.

4.5.2 Independence of n random variables

We now consider a multivariate generalisation of the previous bivariate cases. Fortunately, for n random variables $X_1, X_2, \ldots, X_n$, we have analogous definitions as follows.

1. The random variables $X_1, X_2, \ldots, X_n$ are mutually independent if and only if the events $\{X_1 \leq x_1\}$, $\{X_2 \leq x_2\}$, …, $\{X_n \leq x_n\}$ are independent for all choices of $x_1, x_2, \ldots, x_n \in \mathbb{R}$. That is:

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1)\, F_{X_2}(x_2) \cdots F_{X_n}(x_n) = \prod_{i=1}^{n} F_{X_i}(x_i).$$

2. The random variables $X_1, X_2, \ldots, X_n$ are mutually independent if and only if for $x_1, x_2, \ldots, x_n$ we have:

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)\, f_{X_2}(x_2) \cdots f_{X_n}(x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

3. If the random variables $X_1, X_2, \ldots, X_n$ are mutually independent, then:

$$\mathrm{E}(X_1 X_2 \cdots X_n) = \mathrm{E}(X_1)\, \mathrm{E}(X_2) \cdots \mathrm{E}(X_n) = \prod_{i=1}^{n} \mathrm{E}(X_i)$$

and if $g_1, g_2, \ldots, g_n$ are well-behaved functions, then $g_1(X_1), g_2(X_2), \ldots, g_n(X_n)$ are also mutually independent random variables, hence:

$$\mathrm{E}(g_1(X_1)\, g_2(X_2) \cdots g_n(X_n)) = \mathrm{E}(g_1(X_1))\, \mathrm{E}(g_2(X_2)) \cdots \mathrm{E}(g_n(X_n)) = \prod_{i=1}^{n} \mathrm{E}(g_i(X_i)).$$

4.5.3 Identical distributions

Identically distributed random variables

The random variables $X_1, X_2, \ldots, X_n$ are **identically distributed** if and only if their distribution functions are identical, that is:

$$F_{X_1}(x) = F_{X_2}(x) = \cdots = F_{X_n}(x)$$

for all $x \in \mathbb{R}$.

If $X_1, X_2, \ldots, X_n$ are identically distributed, then we will often just use the letter X to denote a random variable which has the distribution common to all of them. So the distribution function of X is:

$$P(X \leq x) = F_X(x) = F_{X_1}(x) = F_{X_2}(x) = \cdots = F_{X_n}(x).$$

If $X_1, X_2, \ldots, X_n$ are independent and identically distributed, we may sometimes denote this as $\{X_i\} \overset{\text{i.i.d.}}{\sim} F_X(x)$.

Activity 4.6 Suppose $\{X_1, X_2, \ldots, X_n\}$ is a sequence of n independent and identically distributed standard normal random variables. Find an expression for the joint density function of $X_1, X_2, \ldots, X_n$.

4.6 Measures of pairwise dependence

We proceed to consider measures of pairwise dependence between two random variables X and Y. The concepts of covariance and correlation were introduced in **ST104b Statistics 2**, and are extended here to the case of bivariate continuous distributions.

Covariance

Let X and Y be two random variables. The **covariance** of X and Y is defined as:

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

The covariance function has the following properties.

1. Symmetry: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
2. Bilinearity:

   $$\text{Cov}(X_1 + X_2, Y_1 + Y_2) = \text{Cov}(X_1, Y_1) + \text{Cov}(X_1, Y_2) + \text{Cov}(X_2, Y_1) + \text{Cov}(X_2, Y_2)$$

   and for any $a, b \in \mathbb{R}$ we have:

   $$\text{Cov}(aX, bY) = ab\,\text{Cov}(X, Y).$$

3. It is related to the variance:

   $$\text{Var}(X) = \text{Cov}(X, X)$$

   and for linear combinations of X and Y, we have:

   $$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2 \times \text{Cov}(X, Y).$$

4. If X and Y are independent, then $\text{Cov}(X, Y) = 0$.

Example 4.3 For the following joint probability density function:

$$f_{X,Y}(x,y) = \begin{cases} 8xy & \text{for } 0 < y < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

calculate $\text{Cov}(X,Y)$.

Solution

We first determine the marginal density function of X by integrating out y. We have:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \int_0^x 8xy\,\mathrm{d}y = \Big[4xy^2\Big]_0^x = 4x^3.$$

In full we have:

$$f_X(x) = \begin{cases} 4x^3 & \text{for } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Hence:

$$\text{E}(X) = \int_{-\infty}^{\infty} x\, f_X(x)\,\mathrm{d}x = \int_0^1 4x^4\,\mathrm{d}x = \left[\frac{4x^5}{5}\right]_0^1 = \frac{4}{5}.$$

Similarly derive $f_Y(y)$ to obtain $\text{E}(Y) = 8/15$. Next we compute:

$$\text{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int_0^1\int_0^x 8x^2y^2\,\mathrm{d}y\,\mathrm{d}x = \int_0^1 \frac{8x^5}{3}\,\mathrm{d}x = \left[\frac{8x^6}{18}\right]_0^1 = \frac{4}{9}.$$

Hence:

$$\text{Cov}(X,Y) = \text{E}(XY) - \text{E}(X)\,\text{E}(Y) = \frac{4}{9} - \frac{4}{5}\times\frac{8}{15} = \frac{4}{225}.$$

Correlation coefficient

Let X and Y be two random variables. The **correlation coefficient** of X and Y is defined as:

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}.$$

Recall from **ST104b Statistics 2** that correlation measures the strength of the *linear* association between two random variables. It is a scaled version of covariance, such that:

$$|\text{Corr}(X,Y)| \leq 1.$$

Proof: Define the random variable $Z = Y - rX$, where r is a real number. Hence:

$$\text{Var}(Z) = \text{Var}(Y) - 2r\,\text{Cov}(X,Y) + r^2\text{Var}(X) = h(r).$$

Viewing the variance of Z as a function of r, we see that $h(r)$ is a quadratic equation with roots:

$$r = \frac{\text{Cov}(X,Y) \pm \sqrt{(\text{Cov}(X,Y))^2 - \text{Var}(X)\,\text{Var}(Y)}}{\text{Var}(X)}.$$

Since $h(r)$ is defined as a variance, then $h(r) \geq 0$. Hence $h(r)$ has at most one real root which implies that the term under the square root must be at most zero:

$$(\mathrm{Cov}(X,Y))^2 - \mathrm{Var}(X)\,\mathrm{Var}(Y) \leq 0 \quad \Rightarrow \quad \frac{(\mathrm{Cov}(X,Y))^2}{\mathrm{Var}(X)\,\mathrm{Var}(Y)} \leq 1$$

hence:

$$|\mathrm{Corr}(X,Y)| \leq 1.$$

■

Furthermore, $|\mathrm{Corr}(X,Y)| = 1$ if and only if there exists a perfect linear relationship between X and Y, i.e. there exist real numbers $a \neq 0$ and b such that:

$$P(Y = aX + b) = 1$$

that is there is a linear relationship between the variables. Hence if $\mathrm{Corr}(X,Y) = 1$ then $a > 0$, and if $\mathrm{Corr}(X,Y) = -1$ then $a < 0$.

Proof: Since $Y = rX + k$, then:

$$\mathrm{Var}(Y) = r^2\mathrm{Var}(X) \quad \text{and} \quad \mathrm{Cov}(X,Y) = r\,\mathrm{Var}(X).$$

Therefore:

$$\mathrm{Corr}(X,Y) = \frac{r\,\mathrm{Var}(X)}{\sqrt{r^2\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{r}{|r|}.$$

Hence $\mathrm{Corr}(X,Y) = 1$ if $r > 0$ and $\mathrm{Corr}(X,Y) = -1$ if $r < 0$. ■

Independent vs. uncorrelated

Two random variables X and Y are **uncorrelated**, i.e. $\mathrm{Corr}(X,Y) = 0$, if and only if:

$$\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y).$$

Hence:

$$X \perp\!\!\!\perp Y \quad \Rightarrow \quad \mathrm{Corr}(X,Y) = 0$$

but not vice versa, i.e. being uncorrelated is a necessary, but not sufficient, condition for two random variables to be independent. Note that correlation implies *linear* dependence only.

Example 4.4 For any well-behaved functions g_1 and g_2, independence implies that:

$$\mathrm{E}(g_1(X)\,g_2(Y)) = \mathrm{E}(g_1(X))\,\mathrm{E}(g_2(Y)).$$

Suppose X and Y are discrete random variables with joint mass function:

$$p_{X,Y}(x,y) = \begin{cases} 1/4 & \text{for } x = 0 \text{ and } y = 1 \\ 1/4 & \text{for } x = 0 \text{ and } y = -1 \\ 1/4 & \text{for } x = 1 \text{ and } y = 0 \\ 1/4 & \text{for } x = -1 \text{ and } y = 0 \\ 0 & \text{otherwise.} \end{cases}$$

We have that $\mathrm{E}(XY) = 0$ and $\mathrm{E}(X) = \mathrm{E}(Y) = 0$ (check these!) and hence:

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = 0 - 0 \times 0 = 0$$

so X and Y are uncorrelated. Suppose we let $g_1(X) = X^2$ and $g_2(Y) = Y^2$, then:

$$\mathrm{E}(g_1(X)\, g_2(Y)) = \mathrm{E}(X^2 Y^2) = 0$$

however:

$$\mathrm{E}(g_1(X))\,\mathrm{E}(g_2(Y)) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \neq 0.$$

Hence X and Y are not independent.

Example 4.5 Let $X \sim N(0, 1)$, hence $\mathrm{E}(X^k) = 0$ for odd k, and suppose $Y = X^2$. Immediately we deduce that X and Y are dependent, since if you know X then you know Y. Also, if you know Y then you know the absolute value of X. The covariance is:

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = \mathrm{E}(X^3) - 0 \times \mathrm{E}(Y) = \mathrm{E}(X^3) = 0.$$

Therefore, $\mathrm{Corr}(X, Y) = 0$ and so X and Y are uncorrelated, but they are not independent; they are only *linearly* independent. The linear correlation coefficient does not reveal anything about the quadratic dependence of Y on X.

Activity 4.7 Using the definition of covariance:

$$\mathrm{Cov}(X, Y) = \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y)))$$

show that:

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y).$$

4.7 Joint moments and mgfs for two random variables

We now define expectations of functions of random variables, of which joint moments (and joint central moments) are special cases. This then gives rise to the joint moment generating function, and by extension the joint cumulant generating function.

Expectation of a function of two random variables

Let X and Y be two random variables with joint pmf, $p_{X,Y}$, or joint pdf, $f_{X,Y}$. If g is a well-behaved function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, then:

$$\mathrm{E}(g(X, Y)) = \begin{cases} \sum_y \sum_x g(x, y)\, p_{X,Y}(x, y) & \text{(discrete case)} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)\, f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y & \text{(continuous case).} \end{cases}$$

Joint moments

Let X and Y be two random variables with joint pmf, $p_{X,Y}$, or joint pdf, $f_{X,Y}$. The (r,s)th **joint moment** is:

$$\mu_{r,s} = \mathrm{E}(X^r Y^s) = \begin{cases} \sum_y \sum_x x^r y^s p_{X,Y}(x,y) & \text{(discrete case)} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^r y^s f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y & \text{(continuous case).} \end{cases}$$

4

Joint central moments

Let X and Y be two random variables with joint pmf, $p_{X,Y}$, or joint pdf, $f_{X,Y}$. The (r,s)th **joint central moment** is:

$$\begin{aligned} \mu'_{r,s} &= \mathrm{E}((X-\mathrm{E}(X))^r (Y-\mathrm{E}(Y))^s) \\ &= \begin{cases} \sum_y \sum_x (x-\mu_X)^r (y-\mu_Y)^s p_{X,Y}(x,y) & \text{(discrete case)} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)^r (y-\mu_Y)^s f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y & \text{(continuous case).} \end{cases} \end{aligned}$$

Joint moments and joint central moments have the following properties.

1. The mean of X is $\mathrm{E}(X) = \mu_{1,0}$.
2. The rth moment of X is $\mathrm{E}(X^r) = \mu_{r,0}$.
3. The variance of X is $\sigma_X^2 = \mathrm{E}((X-\mathrm{E}(X))^2) = \mu'_{2,0}$.
4. The rth central moment of X is $\mathrm{E}((X-\mathrm{E}(X))^r) = \mu'_{r,0}$.
5. The covariance of X and Y is $\mathrm{Cov}(X,Y) = \mathrm{E}((X-\mathrm{E}(X))(Y-\mathrm{E}(Y))) = \mu'_{1,1}$.
6. The correlation of X and Y is $\mathrm{Corr}(X,Y) = \mu'_{1,1}/\sqrt{\mu'_{2,0}\,\mu'_{0,2}}$.

Example 4.6 Suppose X and Y have the joint density function given by:

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

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

(b) Find $\mathrm{E}(XY)$ and $\mathrm{Var}(XY)$.

Solution

(a) We use the joint density function to get the required expectations. We have:

$$\mathrm{E}(X) = \int_0^1 \int_0^y 2x\,\mathrm{d}x\,\mathrm{d}y = \int_0^1 \Big[x^2\Big]_0^y \mathrm{d}y = \int_0^1 y^2\,\mathrm{d}y = \left[\frac{y^3}{3}\right]_0^1 = \frac{1}{3}.$$

Also:

$$\mathrm{E}(Y) = \int_0^1 \int_x^1 2y \,\mathrm{d}y \,\mathrm{d}x = \int_0^1 \Big[y^2\Big]_x^1 \mathrm{d}x = \int_0^1 (1-x^2)\,\mathrm{d}x = \left[x - \frac{x^3}{3}\right]_0^1 = \frac{2}{3}$$

and:

$$\mathrm{E}(Y^2) = \int_0^1 \int_x^1 2y^2 \mathrm{d}y \,\mathrm{d}x = \int_0^1 \left[\frac{2y^3}{3}\right]_x^1 \mathrm{d}y = \frac{2}{3}\int_0^1 (1-x^3)\,\mathrm{d}x = \frac{2}{3}\left[x - \frac{x^4}{3}\right]_0^1 = \frac{1}{2}.$$

Therefore:

$$\mathrm{Var}(X) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = \frac{1}{18}.$$

(b) We have:

$$\mathrm{E}(XY) = \int_0^1 \int_0^y 2xy \,\mathrm{d}x \,\mathrm{d}y = \int_0^1 \Big[x^2 y\Big]_0^y \mathrm{d}y = \int_0^1 y^3 \mathrm{d}y = \left[\frac{y^4}{4}\right]_0^1 = \frac{1}{4}$$

and:

$$\mathrm{E}(X^2Y^2) = \int_0^1 \int_0^y 2x^2y^2 \mathrm{d}x \,\mathrm{d}y = \int_0^1 \left[\frac{2}{3}\,x^3y^2\right]_0^y \mathrm{d}y = \int_0^1 \frac{2}{3}\,y^5 \mathrm{d}y = \left[\frac{2y^6}{18}\right]_0^1 = \frac{1}{9}$$

hence:

$$\mathrm{Var}(XY) = \mathrm{E}(X^2Y^2) - (\mathrm{E}(XY))^2 = \frac{1}{9} - \left(\frac{1}{4}\right)^2 = \frac{7}{144}.$$

Joint moment generating function

Let X and Y be two random variables with joint pmf, $p_{X,Y}$, or joint pdf, $f_{X,Y}$. The **joint moment generating function** is a function $M_{X,Y} : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ such that:

$$M_{X,Y}(t,u) = \mathrm{E}(\mathrm{e}^{tX+uY}) = \begin{cases} \sum\limits_y \sum\limits_x \mathrm{e}^{tx+uy} p_{X,Y}(x,y) & \text{(discrete case)} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{e}^{tx+uy} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y & \text{(continuous case).} \end{cases}$$

Joint moment generating functions have the following properties.

1. Taylor expansion:

$$M_{X,Y}(t,u) = \mathrm{E}\left(\sum_{i=0}^{\infty} \frac{(tX)^i}{i!} \sum_{j=0}^{\infty} \frac{(uY)^j}{j!}\right) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \mathrm{E}(X^iY^j)\frac{t^iu^j}{i!\,j!}.$$

2. The (r,s)th joint moment is the coefficient of $t^r u^s/(r!\,s!)$ in the Taylor expansion.

3. The (r,s)th joint moment is obtained by differentiation:

$$\mu_{r,s} = \mathrm{E}(X^rY^s) = M_{X,Y}^{(r,s)}(0,0) = \frac{\partial^{r+s}}{\partial t^r \partial u^s} M_{X,Y}(t,u)\Big|_{t=0,u=0}.$$

4. The marginal moment generating functions are:

$$M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = M_{X,Y}(t,0) \quad \text{and} \quad M_Y(u) = \mathrm{E}(\mathrm{e}^{uY}) = M_{X,Y}(0,u).$$

5. If X and Y are independent, then:

$$M_{X,Y}(t,u) = M_X(t)\, M_Y(u).$$

Joint cumulant generating function and joint cumulants

Let $K_{X,Y}(t,u) = \log M_{X,Y}(t,u)$ be the **joint cumulant generating function**. The (r,s)th **joint cumulant**, $\kappa_{r,s}$, is defined as the coefficient of $(t^r u^s)/(r!\, s!)$ in the Taylor expansion of $K_{X,Y}$. Hence:

$$\mathrm{Cov}(X,Y) = \kappa_{1,1} \quad \text{and} \quad \mathrm{Corr}(X,Y) = \frac{\kappa_{1,1}}{\sqrt{\kappa_{2,0}\,\kappa_{0,2}}}.$$

A non-examinable proof can be found in Appendix A.

4.7.1 Joint moments and mgfs of n random variables

We now consider a multivariate generalisation of the previous bivariate cases. Fortunately, for n random variables $X_1, X_2, \ldots, X_n$, we have analogous definitions as follows.

1. Joint moments for the discrete and continuous cases:

$$\begin{aligned}\mu_{r_1,r_2,\ldots,r_n} &= \mathrm{E}(X_1^{r_1} X_2^{r_2} \cdots X_n^{r_n}) \\ &= \begin{cases} \sum_{x_1} \cdots \sum_{x_n} x_1^{r_1} x_2^{r_2} \cdots x_n^{r_n} p_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1^{r_1} x_2^{r_2} \cdots x_n^{r_n} f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)\,\mathrm{d}x_1\,\mathrm{d}x_2 \cdots \mathrm{d}x_n. \end{cases}\end{aligned}$$

2. Joint central moments are:

$$\mu'_{r_1,r_2,\ldots,r_n} = \mathrm{E}((X_1 - \mathrm{E}(X_1))^{r_1} (X_2 - \mathrm{E}(X_2))^{r_2} \cdots (X_n - \mathrm{E}(X_n))^{r_n}).$$

3. The joint moment generating function is:

$$M_{X_1,X_2,\ldots,X_n}(t_1,t_2,\ldots,t_n) = \mathrm{E}(\mathrm{e}^{t_1X_1 + t_2X_2 + \cdots + t_nX_n})$$

and the coefficient of $t_1^{r_1} t_2^{r_2} \cdots t_n^{r_n}/(r_1!\, r_2! \cdots r_n!)$ in the Taylor expansion of $M_{X_1,X_2,\ldots,X_n}$ is $\mathrm{E}(X_1^{r_1} X_2^{r_2} \cdots X_n^{r_n})$.

4. If $X_1, X_2, \ldots, X_n$ are independent, then:

$$M_{X_1,X_2,\ldots,X_n}(t_1,t_2,\ldots,t_n) = M_{X_1}(t_1)\, M_{X_2}(t_2) \cdots M_{X_n}(t_n) = \prod_{i=1}^{n} M_{X_i}(t_i).$$

5. The joint cumulant generating function is:

$$K_{X_1,X_2,\ldots,X_n}(t_1,t_2,\ldots,t_n) = \log(M_{X_1,X_2,\ldots,X_n}(t_1,t_2,\ldots,t_n))$$

and the $(r_1,r_2,\ldots,r_n)$th joint cumulant is defined as the coefficient of $(t_1^{r_1}t_2^{r_2}\cdots t_n^{r_n})/(r_1!\,r_2!\cdots r_n!)$ in the Taylor expansion of $K_{X_1,X_2,\ldots,X_n}$.

Activity 4.8 Suppose $\{X_1,X_2,\ldots,X_n\}$ is a sequence of n independent random variables with cumulant generating functions $K_{X_1},K_{X_2},\ldots,K_{X_n}$. Find an expression for the joint cumulant generating function $K_{X_1,X_2,\ldots,X_n}$ in terms of the individual cumulant generating functions.

4.8 Random vectors

Random vectors allow us to simplify notation when we consider n random variables. Expectations are elementwise and we must remember that the variance of a vector is a matrix, known as the variance–covariance matrix.

Random vector

A **random vector** is an n-dimensional vector of random variables such that:

$$\mathbf{X} = (X_1,X_2,\ldots,X_n)^T : \Omega \to \mathbb{R}^n.$$

The cdf, pmf/pdf, and mgf of a random vector are the joint cdf, joint pmf/pdf, and joint mgf of $X_1,X_2,\ldots,X_n$, so, for any $\mathbf{x} = (x_1,x_2,\ldots,x_n)$, we have:

- $F_{\mathbf{X}}(\mathbf{x}) = F_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$
- $f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$
- $M_{\mathbf{X}}(\mathbf{t}) = M_{X_1,X_2,\ldots,X_n}(t_1,t_2,\ldots,t_n)$, for $\mathbf{t} = (t_1,t_2,\ldots,t_n) \in \mathbb{R}^n$.

Expectation of a random vector

The **expectation of a random vector** is a vector of the expectations, i.e. it is taken element by element:

$$\mathrm{E}(\mathbf{X}) = \begin{pmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_n) \end{pmatrix}.$$

For jointly continuous random variables we have:

$$\begin{aligned} \mathrm{E}(\mathbf{X}) &= \int_{\mathbb{R}^n} \mathbf{x}\, f_{\mathbf{X}}(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_1x_2\cdots x_n\, f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)\,\mathrm{d}x_1\cdots\mathrm{d}x_n. \end{aligned}$$

Variance–covariance matrix

For n random variables $X_1, X_2, \ldots, X_n$ we know the variance of each and the covariance of each pair. All of this information can be summarised in just one object, the **variance–covariance matrix**, defined as:

$$\boldsymbol{\Sigma} = \mathrm{Var}(\mathbf{X}) = \mathrm{E}((\mathbf{X} - \mathrm{E}(\mathbf{X}))(\mathbf{X} - \mathrm{E}(\mathbf{X}))^T)$$

where $\mathbf{X}$ is an $n \times 1$ column vector. Hence its transpose, $\mathbf{X}^T$, is a $1 \times n$ row vector, and $\boldsymbol{\Sigma}$ is an $n \times n$ matrix.

Taking expectations element by element of this matrix we have:

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

The variance–covariance matrix is symmetric, and if the variables are uncorrelated then it is a diagonal matrix. If the variables are also identically distributed then $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$, where σ^2 is the (identical) variance of each random variable and $\mathbf{I}_n$ is the n-dimensional identity matrix. Finally, as the univariate variance is always positive, in this case we have that $\boldsymbol{\Sigma}$ is a non-negative definite matrix, i.e. we have:

$$\mathbf{b}^T \boldsymbol{\Sigma} \mathbf{b} \geq 0 \quad \forall\ \mathbf{b} \in \mathbb{R}^n.$$

Example 4.7 Suppose $n = 2$ and assume that $\mathrm{E}(X) = \mathrm{E}(Y) = 0$, then:

$$\begin{aligned} \boldsymbol{\Sigma} = \mathrm{E}\left(\begin{pmatrix} X \\ Y \end{pmatrix} \begin{pmatrix} X & Y \end{pmatrix} \right) &= \mathrm{E}\left(\begin{pmatrix} X^2 & XY \\ YX & Y^2 \end{pmatrix} \right) \\ &= \begin{pmatrix} \mathrm{E}(X^2) & \mathrm{E}(XY) \\ \mathrm{E}(YX) & \mathrm{E}(Y^2) \end{pmatrix} \\ &= \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(Y, X) & \mathrm{Var}(Y) \end{pmatrix}. \end{aligned}$$

Activity 4.9 Suppose that $\mathbf{X}$ is a random vector. Using the definition of the variance–covariance matrix $\mathrm{Var}(\mathbf{X}) = \mathrm{E}((\mathbf{X} - \mathrm{E}(\mathbf{X}))(\mathbf{X} - \mathrm{E}(\mathbf{X}))^T)$, show that:

$$\mathrm{Var}(\mathbf{X}) = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \cdots & \cdots & \mathrm{Var}(X_n) \end{pmatrix}.$$

Activity 4.10 Consider the random vectors $\mathbf{X} = (X_1, X_2, \ldots, X_m)^T$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$.

(a) Provide an expression for the entries in $\text{Cov}(\mathbf{X}, \mathbf{Y})$.

(b) Show that $\text{Cov}(\mathbf{Y}, \mathbf{X}) = \text{Cov}(\mathbf{X}, \mathbf{Y})^T$.

4.9 Transformations of continuous random variables

Recall that if Y is given by a strictly monotone function of a continuous random variable X, we can derive an expression for the density function of Y in terms of the density function of X (Section 3.11.2). We now extend this idea to functions of several continuous random variables starting with the bivariate case.

Definition: one-to-one and onto

Consider a function $\mathbf{g}$ with domain D and range R. We say that $\mathbf{g}$ is:

- **one-to-one** if:
$$\mathbf{g}(\mathbf{x}_1) = \mathbf{g}(\mathbf{x}_2) \quad \Rightarrow \quad \mathbf{x}_1 = \mathbf{x}_2$$
for all $\mathbf{x}_1, \mathbf{x}_2 \in D$.
- **onto** for all $\mathbf{y} \in R$ if we can find $\mathbf{x} \in D$ such that $\mathbf{y} = \mathbf{g}(\mathbf{x})$.

4.9.1 Bivariate transformations

We are interested in transforming one pair of random variables into another pair of random variables. Consider the pairs (U, V) and (X, Y). Suppose that X and Y are both functions of U and V such that:

$$X = g_1(U, V) \quad \text{and} \quad Y = g_2(U, V).$$

We will make extensive use of the inverse transformations:

$$U = h_1(X, Y) \quad \text{and} \quad V = h_2(X, Y).$$

We refer to the overall transformation as $\mathbf{g}$ so $(X, Y) = \mathbf{g}(U, V)$ and the inverse as $\mathbf{h}$ such that:

$$(U, V) = \mathbf{g}^{-1}(X, Y) = \mathbf{h}(X, Y).$$

Suppose $\mathbf{g}$ is a well-behaved function $\mathbf{g} : \mathbb{R}^2 \to \mathbb{R}^2$. More specifically, suppose that, for domain $D \subseteq \mathbb{R}^2$, the function $\mathbf{g}$ is a one-to-one map onto the range $R \subseteq \mathbb{R}^2$.

Change of variables formula

If (U, V) is a pair of continuous random variables with support D and $(X, Y) = \mathbf{g}(U, V)$, the joint density function of X and Y is given by the **change of variables formula**:

$$f_{X,Y}(x,y) = \begin{cases} f_{U,V}(\mathbf{h}(x,y))\,|J_{\mathbf{h}}(x,y)| & \text{for } (x,y) \in R \\ 0 & \text{otherwise} \end{cases}$$

where $J_{\mathbf{h}}(x, y)$ is the **Jacobian** of the inverse transformation:

$$\begin{aligned} J_{\mathbf{h}}(x,y) &= \begin{vmatrix} \dfrac{\partial}{\partial x} h_1(x,y) & \dfrac{\partial}{\partial x} h_2(x,y) \\ \dfrac{\partial}{\partial y} h_1(x,y) & \dfrac{\partial}{\partial y} h_2(x,y) \end{vmatrix} \\ &= \frac{\partial}{\partial x} h_1(x,y) \frac{\partial}{\partial y} h_2(x,y) - \frac{\partial}{\partial x} h_2(x,y) \frac{\partial}{\partial y} h_1(x,y). \end{aligned}$$

Note that the Jacobian can be expressed in terms of the Jacobian of the original transformation. If:

$$J_{\mathbf{g}}(u,v) = \begin{vmatrix} \dfrac{\partial}{\partial u} g_1(u,v) & \dfrac{\partial}{\partial u} g_2(u,v) \\ \dfrac{\partial}{\partial v} g_1(u,v) & \dfrac{\partial}{\partial v} g_2(u,v) \end{vmatrix}$$

then:

$$J_{\mathbf{h}}(x,y) = (J_{\mathbf{g}}(h_1(x,y), h_2(x,y)))^{-1}.$$

There are several equivalent statements which can be formed for the change of variables formula. For (x, y) in the range of $\mathbf{g}$, then:

$$\begin{aligned} f_{X,Y}(x,y) &= f_{U,V}(h_1(x,y), h_2(x,y))\,|J_{\mathbf{h}}(x,y)| \\ &= f_{U,V}(\mathbf{g}^{-1}(x,y))\,|(J_{\mathbf{g}}(\mathbf{g}^{-1}(x,y)))^{-1}| \\ &= f_{U,V}(g_1^{-1}(x,y), g_2^{-1}(x,y))\,|(J_{\mathbf{g}}(g_1^{-1}(x,y), g_2^{-1}(x,y)))^{-1}|. \end{aligned}$$

In order to apply the change of variables formula, follow the steps outlined below.

1. Determine the transformation, i.e. find g_1 and g_2 such that:

   $$x = g_1(u,v) \quad \text{and} \quad y = g_2(u,v).$$

   Make sure you specify the domain and range of the transformation.

2. Invert the transformation, i.e. find h_1 and h_2 such that:

   $$u = h_1(x,y) \quad \text{and} \quad v = h_2(x,y).$$

3. Derive the Jacobian by finding the partial derivatives of h_1 and h_2 with respect to x and y such that:

   $$J_{\mathbf{h}}(x,y) = \begin{vmatrix} \dfrac{\partial}{\partial x} h_1(x,y) & \dfrac{\partial}{\partial x} h_2(x,y) \\ \dfrac{\partial}{\partial y} h_1(x,y) & \dfrac{\partial}{\partial y} h_2(x,y) \end{vmatrix}.$$

Remember that it is the *absolute value* of the Jacobian, i.e. $|J_{\mathbf{h}}(x,y)|$ which is used in the change of variables formula.

4. Construct the joint density function of X and Y as:

$$f_{X,Y}(x,y) = f_{U,V}(h_1(x,y), h_2(x,y))\,|J_{\mathbf{h}}(x,y)|.$$

Make sure you specify the support of the joint density function.

Example 4.8 Suppose that U and V are continuous random variables with joint density function $f_{U,V}$. We seek the density function of the product UV in terms of $f_{U,V}$. We begin by defining a bivariate transformation $\mathbf{g} : \mathbb{R}^2 \to \mathbb{R}^2$ as:

$$g_1(u,v) = uv \quad \text{and} \quad g_2(u,v) = v.$$

Let $X = g_1(U,V) = UV$ and $Y = g_2(U,V) = V$. We now follow the above steps to determine the joint density function $f_{X,Y}$ from which we will be able to derive the density function of $X = UV$ by integrating $f_{X,Y}$.

1. The transformation can be written as:

$$x = uv \quad \text{and} \quad y = v.$$

If the domain of this transformation is the whole of $\mathbb{R}^2$, then the range is also the whole of $\mathbb{R}^2$.

2. The inverse transformation $\mathbf{h}$ is:

$$u = \frac{x}{y} \quad \text{and} \quad v = y.$$

3. The Jacobian is:

$$J_{\mathbf{h}}(x,y) = \begin{vmatrix} \partial u/\partial x & \partial v/\partial x \\ \partial u/\partial y & \partial v/\partial y \end{vmatrix} = \begin{vmatrix} 1/y & 0 \\ -x/y^2 & 1 \end{vmatrix} = \frac{1}{y}$$

and so the absolute value of the Jacobian is $1/|y|$.

4. The joint density function of X and Y is:

$$f_{X,Y}(x,y) = f_{U,V}(x/y, y)\,\frac{1}{|y|}.$$

We are now in a position to obtain the density function of $X = UV$, which is:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \int_{-\infty}^{\infty} f_{U,V}(x/y, y)\,\frac{1}{|y|}\,\mathrm{d}y.$$

Example 4.9 Consider the transformation:

$$x(u,v) = u \quad \text{and} \quad y(u,v) = \rho u + \sqrt{(1-\rho^2)}v$$

where $|\rho| < 1$.

(a) Show that the Jacobian of the inverse transformation is $1/\sqrt{(1-\rho^2)}$.

(b) Let U and V be independent standard normal random variables. Write down the joint density function of U and V.

(c) Define transformed variables $X = x(U, V)$ and $Y = y(U, V)$. Use the change of variables formula to derive the joint density function of X and Y.

Solution

(a) The inverse transformation is:

$$u = x \quad \text{and} \quad v = \frac{y - \rho x}{\sqrt{1-\rho^2}}.$$

Hence the Jacobian is:

$$\begin{vmatrix} \partial u/\partial x & \partial v/\partial x \\ \partial u/\partial y & \partial v/\partial y \end{vmatrix} = \begin{vmatrix} 1 & -\rho/\sqrt{1-\rho^2} \\ 0 & 1/\sqrt{1-\rho^2} \end{vmatrix} = \frac{1}{\sqrt{1-\rho^2}}.$$

(b) Since U and V are independent:

$$f_{U,V}(u,v) = f_U(u)\, f_V(v) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-u^2/2} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-v^2/2} = \frac{1}{2\pi} \exp\left(-\frac{1}{2}(u^2+v^2)\right).$$

(c) Applying the change of variables formula, we have:

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left(\frac{x^2 + (y-\rho x)^2}{1-\rho^2}\right)\right) \frac{1}{\sqrt{1-\rho^2}} \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)(x^2 - 2\rho xy + y^2)}\right). \end{aligned}$$

Example 4.10 If the joint density function of X and Y is:

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

find the density function of $U = X/Y$.

Solution

Note U only takes values between 0 and 1. First we identify A_u for $u \in (0,1)$. As $A_u = \{0 < x < y < 1,\ x/y \le u\}$, it is the triangle with the three sides: $x = 0$, $y = 1$ and $x = uy$. The area of A_u is $1 \times u/2 = u/2$. Next:

$$F_U(u) = P(U \le u) = \int_{A_u} f_{X,Y}(x,y)\, \mathrm{d}x\, \mathrm{d}y = 2 \int_{A_u} 1\, \mathrm{d}x\, \mathrm{d}y = 2 \times \frac{u}{2}$$

for $u \in (0,1)$. Hence:

$$f_U(u) = F'(u) = 1$$

for $u \in (0,1)$, and so U has a continuous uniform distribution.

A word on notation

We have used $\mathbf{g}$ and $\mathbf{h}$ to denote a transformation and its inverse. Often these functions are treated implicitly, i.e. given the inverse transformation $u = h_1(x,y)$ and $v = h_2(x,y)$ the Jacobian may be written as:

$$J_{\mathbf{h}}(x,y) = \begin{vmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial v}{\partial x} \\ \dfrac{\partial u}{\partial y} & \dfrac{\partial v}{\partial y} \end{vmatrix}.$$

This provides a useful way of remembering the Jacobian – in this context we are viewing u and v as functions of x and y.

Activity 4.11 Suppose U and V are continuous random variables with the joint density function $f_{U,V}$. Derive an expression for the density function of the ratio U/V in terms of $f_{U,V}$.

Activity 4.12 Suppose U and V are independent and identically distributed exponential random variables. Determine the density function of U/V.

4.9.2 Multivariate transformations

We now generalise to the multivariate case of n random variables. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ be a continuous random vector and let $g : \mathbb{R}^n \to \mathbb{R}^n$ be a well-behaved function. We will assume that, if $D \subseteq \mathbb{R}^n$ is the support of $\mathbf{X}$, then $\mathbf{g}$ is a one-to-one mapping from D onto the range $R \subseteq \mathbb{R}^n$. We continue to make extensive use of the inverse transformation $\mathbf{h}(\mathbf{y}) = \mathbf{g}^{-1}(\mathbf{y})$, and on occasion consider individual components of vectors such as:

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \quad \text{and} \quad \mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_n(\mathbf{x}))^T.$$

Note that, for $i = 1, 2, \ldots, n$, each g_i is a function of n variables, i.e. $g_i : \mathbb{R}^n \to \mathbb{R}$, so we could write:

$$\mathbf{g}(\mathbf{x}) = (g_1(x_1, x_2, \ldots, x_n), g_2(x_1, x_2, \ldots, x_n), \ldots, g_n(x_1, x_2, \ldots, x_n))^T.$$

We now define a random vector $\mathbf{Y} = \mathbf{g}(\mathbf{X})$. The joint density function of $\mathbf{Y}$ is:

$$f_{\mathbf{Y}}(\mathbf{y}) = \begin{cases} f_{\mathbf{X}}(\mathbf{h}(\mathbf{Y}))\,|J_{\mathbf{h}}(\mathbf{y})| & \text{for } \mathbf{y} \in R \\ 0 & \text{otherwise.} \end{cases}$$

The Jacobian is:

$$J_{\mathbf{h}}(\mathbf{y}) = \left|\frac{\partial}{\partial \mathbf{y}}\,\mathbf{h}(\mathbf{y})\right| = \begin{vmatrix} \frac{\partial}{\partial y_1}\,h_1(\mathbf{y}) & \frac{\partial}{\partial y_1}\,h_2(\mathbf{y}) & \cdots & \frac{\partial}{\partial y_1}\,h_n(\mathbf{y}) \\ \frac{\partial}{\partial y_2}\,h_1(\mathbf{y}) & \frac{\partial}{\partial y_2}\,h_2(\mathbf{y}) & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial}{\partial y_n}\,h_1(\mathbf{y}) & \cdots & \cdots & \frac{\partial}{\partial y_n}\,h_n(\mathbf{y}) \end{vmatrix}.$$

The Jacobian can also be written as:

$$J_{\mathbf{h}}(\mathbf{y}) = (J_{\mathbf{g}}(\mathbf{h}(\mathbf{y})))^{-1} \quad \text{where} \quad J_{\mathbf{g}}(\mathbf{x}) = \left|\frac{\partial}{\partial \mathbf{x}}\,\mathbf{g}(\mathbf{x})\right|.$$

Example 4.11 A simple function of a random vector which often arises is the linear transformation:

$$\mathbf{Y} = \mathbf{A}\mathbf{X}$$

where $\mathbf{A}$ is non-singular, i.e. $|\mathbf{A}| \neq 0$, which ensures that $\mathbf{A}^{-1}$ is well-defined. We have that $J_{\mathbf{g}}(\mathbf{x}) = |\mathbf{A}|$ and hence $J_{\mathbf{h}}(\mathbf{y}) = |\mathbf{A}|^{-1}$. Therefore:

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{||\mathbf{A}||}\, f_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y}).$$

where $||\mathbf{A}||$ is the absolute value of the determinant of $\mathbf{A}$.

Activity 4.13 Suppose that $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ is a vector of independent standard normal random variables and that $\mathbf{A}$ is an $n \times n$ non-singular matrix. Provide an expression for the density function of $\mathbf{Y}$, where $\mathbf{Y} = \mathbf{A}\mathbf{X}$.

4.10 Sums of random variables

We first consider the bivariate case, and then generalise to n random variables.

Moments of a sum

Let X and Y be random variables. We have:

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$$

and:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \times \mathrm{Cov}(X, Y).$$

Applying the linearity of expectations and the binomial expansion, for $r \in \mathbb{N}$ we have:

$$\mathrm{E}((X + Y)^r) = \sum_{i=0}^{r} \binom{r}{i} \mathrm{E}(X^i Y^{r-i}).$$

Probability mass/density function of a sum

Let X and Y be jointly distributed random variables. Define $Z = X + Y$.

In the discrete case, the probability mass function of Z is:

$$p_Z(z) = \sum_u p_{X,Y}(u, z-u).$$

In the continuous case, the probability density function of Z is:

$$\int_{-\infty}^{\infty} f_{X,Y}(u, z-u)\,\mathrm{d}u.$$

In the continuous case we simply change variables where $X = U$ and $Y = Z - U$.
In the discrete case, note that:

$$\{X + Y = z\} = \bigcup_u \{X = u \cap Y = z - u\}$$

and, because this is a sum of mutually exclusive events, for any u we have:

$$P(X + Y = z) = \sum_u P(X = u \cap Y = z - u).$$

Probability distribution of a sum of independent random variables

Let X and Y be independent random variables. Define $Z = X + Y$.

In the discrete case, the probability mass function of Z is:

$$p_Z(z) = \sum_u p_X(u)\, p_Y(z-u).$$

In the continuous case, the probability density function of Z is:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(u)\, f_Y(z-u)\,\mathrm{d}u.$$

This operation is known as *convolution*:

$$f_Z = f_X * f_Y \quad \Leftrightarrow \quad \int_{-\infty}^{\infty} f_X(u)\, f_Y(z-u)\,\mathrm{d}u.$$

Convolution is commutative, i.e. $f_X * f_Y = f_Y * f_X$.

Generating functions of the sum of independent random variables

Let X and Y be independent random variables. Define $Z = X + Y$. The moment generating function of Z is:

$$M_Z(t) = M_X(t)\, M_Y(t)$$

and the cumulant generating function is:

$$K_Z(t) = K_X(t) + K_Y(t).$$

Example 4.12 Let X and Y be independent exponential random variables, such that $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\theta)$, with $\lambda \neq \theta$. The pdf of $Z = X + Y$ is:

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(u)\, f_Y(z-u)\, \mathrm{d}u \\ &= \int_0^z \lambda \mathrm{e}^{-\lambda u} \theta \mathrm{e}^{-\theta(z-u)}\, \mathrm{d}u \\ &= \lambda\theta \mathrm{e}^{-\theta z} \left[-\frac{\mathrm{e}^{-(\lambda-\theta)u}}{\lambda-\theta} \right]_0^z \\ &= \frac{\lambda\theta}{\lambda-\theta} (\mathrm{e}^{-\theta z} - \mathrm{e}^{-\lambda z}) \end{aligned}$$

for $z \geq 0$, and 0 otherwise.

Note the domain of integration is $[0, z]$. Indeed, since both X and Y are positive random variables, we have that U and $Z - U$ must also be positive, hence we require $0 < U \leq Z$.

While in principle we could make use of the moment generating functions, in this instance we obtain a function of t which does not have an expression which resembles that of a known distribution.

Example 4.13 Let X and Y be independent normal random variables, such that $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$. In order to compute the pdf of $Z = X + Y$ we use the cumulant generating functions:

$$K_X(t) = \mu_X t + \frac{\sigma_X^2 t^2}{2} \quad \text{and} \quad K_Y(t) = \mu_Y t + \frac{\sigma_Y^2 t^2}{2}$$

such that:

$$K_Z(t) = (\mu_X + \mu_Y)t + \frac{(\sigma_X^2 + \sigma_Y^2)t^2}{2}.$$

By uniqueness of cumulant generating functions we have that:

$$Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

Activity 4.14 Suppose the random variables X and Y have the joint density function:

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

Derive the density function of $Z = X + Y$.

Activity 4.15 Suppose X and Y are positive continuous random variables and let $Z = X + Y$. Show that:

$$f_Z(z) = \begin{cases} \int_0^z f_{X,Y}(u, z-u)\,\mathrm{d}u & \text{for } z \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

We now generalise to the multivariate case for n independent random variables $X_1, X_2, \ldots, X_n$. Define $S = \sum_{i=1}^{n} X_i$.

1. The pmf of S is:
$$p_S = p_{X_1} * p_{X_2} * \cdots * p_{X_n}$$
and the pdf of S is:
$$f_S = f_{X_1} * f_{X_2} * \cdots * f_{X_n}.$$

2. The mgf of S is:
$$M_S(t) = \prod_{i=1}^{n} M_{X_i}(t).$$

3. If $X_1, X_2, \ldots, X_n$ are also identically distributed, then they have a common mgf $M_X(t)$, hence:
$$f_S = \underbrace{f * f * \cdots * f}_{n \text{ times}}, \quad M_S(t) = (M_X(t))^n \quad \text{and} \quad K_S(t) = nK_X(t).$$
Recall that the acronym i.i.d. is used for 'independent and identically distributed'.

Example 4.14 Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli(π) random variables. We have that the mgf is:
$$M_X(t) = 1 - \pi + \pi \mathrm{e}^t$$
hence the sum $S = \sum_{i=1}^{n} X_i$ has mgf:
$$M_S(t) = \prod_{i=1}^{n} (1 - \pi + \pi \mathrm{e}^t) = (1 - \pi + \pi \mathrm{e}^t)^n.$$
By uniqueness of the mgf, we conclude that $S \sim \mathrm{Bin}(n, \pi)$.

Example 4.15 Let $X_1, X_2, \ldots, X_n$ be independent normal random variables such that $X_i \sim N(\mu_i, \sigma_i^2)$. For fixed constants $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$, we have:

$$S = \sum_{i=1}^{n}(a_i X_i + b_i) \sim N\left(\sum_{i=1}^{n}(a_i\mu_i + b_i), \sum_{i=1}^{n} a_i^2\sigma_i^2\right).$$

Furthermore, if $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, then:

$$S = \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2).$$

Example 4.16 $X_1, X_2, \ldots, X_n$ is a sequence of i.i.d. random variables with $\mathrm{E}(X) = \mu < \infty$. Let $S_n = \sum_{i=1}^{n} X_i$ and $Y_n = S_n/n$, which note is the sample mean.

(a) Show that the moment generating function of Y_n is:

$$M_{Y_n}(t) = M_{S_n}(t/n)$$

and hence show that the cumulant generating function is:

$$K_{Y_n}(t) = n\,K_X(t/n)$$

where M_{S_n} is the moment generating function of S_n and K_X is the cumulant generating function of X.

Hint: You will need an expression for M_{S_n} in terms of M_X.

(b) Show that $K_{Y_n}(t) \to \mu t$ as $n \to \infty$. Draw the appropriate conclusions about convergence of Y_n.

Solution

(a) We have:

$$M_{Y_n}(t) = \mathrm{E}(\mathrm{e}^{tY_n}) = \mathrm{E}(\mathrm{e}^{tS_n/n}) = M_{S_n}(t/n).$$

Using the fact that $X_1, X_2, \ldots, X_n$ are i.i.d. then:

$$M_{S_n}(t) = \mathrm{E}(\mathrm{e}^{t(X_1+X_2+\cdots+X_n)}) = (\mathrm{E}(\mathrm{e}^{tX}))^n = (M_X(t))^n.$$

Hence:

$$K_{Y_n}(t) = \log M_{Y_n}(t) = \log M_{S_n}(t/n) = \log(M_X(t/n))^n = nK_X(t/n).$$

(b) We have:

$$\begin{aligned} K_{Y_n}(t) = nK_X(t/n) &= n\left(\kappa_1\frac{t}{n} + \text{terms of order } \left(\frac{1}{n}\right)^2 \text{ and higher}\right) \\ &= \kappa_1 t + \text{terms of order } \frac{1}{n} \text{ and higher} \\ &\to \kappa_1 t \end{aligned}$$

as $n \to \infty$. We know that $\kappa_1 = \mu$, the mean of X. We conclude that the sample mean converges in distribution to the mean as $n \to \infty$.

Example 4.17 Let X and Y be independent random variables with the Laplace distribution for which the moment generating function (see Example 3.43) is:

$$M_X(t) = M_Y(t) = \frac{1}{1-t^2}$$

for $|t| < 1$.

(a) Let $U = X + Y$. Derive the moment generating function of U.

(b) Let $V = X - Y$. Show that U and V are identically distributed.

(c) Find the moment generating function of U and V. Show that U and V are uncorrelated, but not independent.

Solution

(a) Since X and Y are independent, we have:

$$M_U(t) = \mathrm{E}(\mathrm{e}^{tU}) = \mathrm{E}(\mathrm{e}^{t(X+Y)}) = \mathrm{E}(\mathrm{e}^{tX})\,\mathrm{E}(\mathrm{e}^{tY}) = M_X(t)\,M_Y(t) = \left(\frac{1}{1-t^2}\right)^2$$

for $|t| < 1$.

(b) Since X and Y are independent, we have:

$$M_V(t) = \mathrm{E}(\mathrm{e}^{tV}) = \mathrm{E}(\mathrm{e}^{t(X-Y)}) = \mathrm{E}(\mathrm{e}^{tX})\,\mathrm{E}(\mathrm{e}^{-tY}) = M_X(t)\,M_Y(-t) = \left(\frac{1}{1-t^2}\right)^2$$

for $|t| > 1$. Since $M_U(t) = M_V(t)$, U and V are identically distributed, by uniqueness of moment generating functions.

(c) Since X and Y are independent, we have:

$$\begin{aligned} M_{U,V}(s,t) = \mathrm{E}(\mathrm{e}^{sU+tV}) &= \mathrm{E}(\mathrm{e}^{s(X+Y)+t(X-Y)}) \\ &= \mathrm{E}(\mathrm{e}^{(s+t)X}\mathrm{e}^{(s-t)Y}) \\ &= \mathrm{E}(\mathrm{e}^{(s+t)X})\,\mathrm{E}(\mathrm{e}^{(s-t)Y}) \\ &= M_X(s+t)\,M_Y(s-t) \\ &= \left(\frac{1}{1-(s+t)^2}\right)\left(\frac{1}{1-(s-t)^2}\right). \end{aligned}$$

Since:

$$M_U(s)\,M_V(t) = \left(\frac{1}{1-s^2}\right)^2\left(\frac{1}{1-t^2}\right)^2 \neq M_{U,V}(s,t)$$

then U and V are not independent. However:

$$\begin{aligned}\mathrm{Cov}(U,V) &= \mathrm{E}((U-\mathrm{E}(U))(V-\mathrm{E}(V))) \\ &= \mathrm{E}(UV) \\ &= \mathrm{E}(X+Y)\,\mathrm{E}(X-Y) \\ &= \mathrm{E}(X^2-Y^2) \\ &= \mathrm{E}(X^2)-\mathrm{E}(Y^2) \\ &= 0\end{aligned}$$

so U and V are uncorrelated.

Special examples of sums of independent random variables are the following.

1. Binomial: If $X \sim \mathrm{Bin}(n_1,\pi)$ and $Y \sim \mathrm{Bin}(n_2,\pi)$, then $Z \sim \mathrm{Bin}(n_1+n_2,\pi)$. Also, if:

$$X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bin}(m,\pi) \quad \Rightarrow \quad S \sim \mathrm{Bin}(nm,\pi) \quad \text{for } i=1,2,\ldots,n.$$

2. Poisson: If $X \sim \mathrm{Pois}(\lambda_1)$ and $Y \sim \mathrm{Pois}(\lambda_2)$, then $Z \sim \mathrm{Pois}(\lambda_1+\lambda_2)$. Also, if:

$$X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Pois}(\lambda) \quad \Rightarrow \quad S \sim \mathrm{Pois}(n\lambda) \quad \text{for } i=1,2,\ldots,n.$$

3. Gamma: If $X \sim \mathrm{Gamma}(\alpha_1,\beta)$ and $Y \sim \mathrm{Gamma}(\alpha_2,\beta)$, then $Z \sim \mathrm{Gamma}(\alpha_1+\alpha_2,\beta)$. Also, if:

$$X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(\lambda) \quad \Rightarrow \quad S \sim \mathrm{Gamma}(n,\lambda) \quad \text{for } i=1,2,\ldots,n.$$

Example 4.18 Let $X_1, X_2, \ldots, X_n$ be independent Poisson random variables with $\mathrm{E}(X_i)=\mu_i$ for $i=1,2,\ldots,n$. Find the probability distribution of $Y=\sum_{i=1}^{n} X_i$.

Solution

We have:

$$\begin{aligned}M_Y(t) = \mathrm{E}(\mathrm{e}^{t(X_1+X_2+\cdots+X_n)}) &= \prod_{i=1}^{n} \mathrm{E}(\mathrm{e}^{tX_i}) \\ &= \prod_{i=1}^{n} \exp(\mu_i(\mathrm{e}^t-1)) \\ &= \exp((\mu_1+\mu_2+\cdots+\mu_n)(\mathrm{e}^t-1)).\end{aligned}$$

Hence $Y \sim \mathrm{Pois}(\mu_1+\mu_2+\cdots+\mu_n)$.

Example 4.19 Assume that X and Y are independent exponential random variables with the same parameter λ. Derive the moment generating function of $U = X + Y$. What can we say about the distribution of U?

Solution

Since $M_X(t) = M_Y(t) = \lambda/(\lambda - t)$, for $t < \lambda$, noting the independence of X and Y we have:

$$\begin{aligned} M_U(t) = \mathrm{E}(\mathrm{e}^{tU}) &= \mathrm{E}(\mathrm{e}^{t(X+Y)}) \\ &= \mathrm{E}(\mathrm{e}^{tX})\,\mathrm{E}(\mathrm{e}^{tY}) \\ &= \left(\frac{\lambda}{\lambda - t}\right)^2 \end{aligned}$$

which is the mgf of a Gamma$(2, \lambda)$ random variable.

Example 4.20 Suppose $X \sim \text{Gamma}(\alpha, \beta)$, such that:

$$f_X(x) = \frac{1}{\Gamma(\alpha)}\,\beta^\alpha x^{\alpha-1}\mathrm{e}^{-\beta x}$$

for $x \geq 0$, and 0 otherwise, where:

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

(a) Find the moment generating function of X.

(b) Let $n \in \mathbb{N}$ and suppose $Y_1, Y_2, \ldots, Y_n$ are independent random variables each following an exponential distribution with parameter $\beta > 0$ (i.e. $\mathrm{E}(Y_i) = 1/\beta$ for $i = 1, 2, \ldots, n$). Use moment generating functions to show that $\sum_{i=1}^{n} Y_i$ follows a Gamma(n, β) distribution. State clearly any properties of moment generating functions you use.

Solution

(a) We have:

$$\begin{aligned} M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) &= \int_0^\infty \mathrm{e}^{tx}\frac{1}{\Gamma(\alpha)}\,\beta^\alpha x^{\alpha-1}\mathrm{e}^{-\beta x}\,\mathrm{d}x \\ &= \int_0^\infty \frac{1}{\Gamma(\alpha)}\beta^\alpha x^{\alpha-1}\mathrm{e}^{-(\beta-t)x}\,\mathrm{d}x \\ &= \frac{\beta^\alpha}{(\beta-t)^\alpha}\int_0^\infty \frac{1}{\Gamma(\alpha)}(\beta-t)^\alpha x^{\alpha-1}\mathrm{e}^{-(\beta-t)}\,\mathrm{d}x \\ &= \left(\frac{\beta}{\beta - t}\right)^\alpha \end{aligned}$$

since the integral is that of a Gamma$(\alpha, \beta - t)$ density function.

(b) Let $X = \sum_{i=1}^{n} Y_i$, where $Y_i \sim \text{Exp}(\beta)$. Since $\{Y_i\}$ is a sequence of i.i.d. random variables, then:

$$\begin{aligned} M_X(t) = \text{E}(\text{e}^{tX}) = \text{E}(\text{e}^{t(Y_1+Y_2+\cdots+Y_n)}) &= \text{E}(\text{e}^{tY_1}\text{e}^{tY_2}\cdots\text{e}^{tY_n}) \\ &= \prod_{i=1}^{n} \text{E}(\text{e}^{tY_i}) \\ &= \prod_{i=1}^{n} \text{E}(\text{e}^{tY}) \\ &= (M_Y(t))^n \\ &= \left(\frac{\beta}{\beta - t}\right)^n. \end{aligned}$$

By uniqueness of moment generating functions, $X \sim \text{Gamma}(n, \beta)$.

Activity 4.16 Show that if $X_1, X_2, \ldots, X_n$ is a sequence of independent and identically distributed random variables and $S = \sum_{i=1}^{n} X_i$, then:

$$M_S(t) = (M_X(t))^n \quad \text{and} \quad K_S(t) = nK_X(t).$$

Activity 4.17 Let S be the sum of n binomial distributions. Under what conditions does S also have a binomial distribution?

Activity 4.18 Suppose $X \sim \text{Gamma}(\alpha_1, \beta_1)$ and $Y \sim \text{Gamma}(\alpha_2, \beta_2)$ are independent random variables. Provide an expression for the moment generating function of $Z = X + Y$. Under what conditions does Z have a gamma distribution?

4.11 Multivariate normal distribution

We consider the bivariate case. We require a bivariate version of the normal distribution. Given two standard normal random variables, we can build a bivariate normal distribution which depends only on the correlation between the variables.

4.11.1 Standard bivariate normal distribution

Let U and V be standard normal random variables, and for some correlation coefficient $|\rho| < 1$, define $X = U$ and $Y = \rho U + \sqrt{1-\rho^2}V$. We then have the following.

1. $X \sim N(0,1)$ and $Y \sim N(0,1)$.
2. $\text{Corr}(X, Y) = \rho$.

3. The joint pdf is that of a standard bivariate normal random variable and depends only on the parameter ρ. The joint pdf is:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right).$$

The random vector $\mathbf{X} = (X,Y)^T$ is normally distributed, written as:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

or $\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma}_{X,Y})$, where $\mathbf{\Sigma}_{X,Y}$ is the 2×2 variance–covariance matrix.

4. The joint mgf is:

$$M_{X,Y}(s,t) = \exp\left(\frac{s^2 + 2\rho st + t^2}{2}\right).$$

Activity 4.19 Prove the four results above.

4.11.2 Bivariate normal for independent random variables

Let U and V be i.i.d. standard normal random variables. Since $\text{Corr}(U,V) = \rho = 0$, the joint pdf and mgf are:

$$f_{U,V}(u,v) = \frac{1}{2\pi} \mathrm{e}^{-(u^2+v^2)^2/2} \quad \text{and} \quad M_{U,V}(s,t) = \mathrm{e}^{(s^2+t^2)^2/2}.$$

The random vector $(U,V)^T$ is normally distributed with variance–covariance matrix:

$$\mathbf{\Sigma}_{U,V} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Activity 4.20 Show that if U and V are independent standard normal random variables, then the joint density and joint moment generating functions of U and V are given by, respectively:

$$f_{U,V}(u,v) = \frac{1}{2\pi} \mathrm{e}^{-(u^2+v^2)/2} \quad \text{and} \quad M_{U,V}(s,t) = \mathrm{e}^{(s^2+t^2)/2}.$$

4.11.3 Computing the joint pdf

Given $X = U$ and $Y = \rho U + \sqrt{1-\rho^2}V$, we have to compute $f_{X,Y}(x,y)$ given $f_{U,V}(u,v)$. Given the function $\mathbf{h} : \mathbb{R}^2 \to \mathbb{R}^2$ such that $\mathbf{h}(X,Y) = (U,V)$ and the domain of $\mathbf{h}$ is $D \subseteq \mathbb{R}^2$ and it is in one-to-one correspondence with the support of (U,V), we have the rule:

$$f_{X,Y}(x,y) = \begin{cases} f_{U,V}(\mathbf{h}(x,y)) \, |J_{\mathbf{h}}(x,y)| & \text{for } (x,y) \in D \\ 0 & \text{otherwise} \end{cases}$$

where:

$$J_{\mathbf{h}}(x,y) = \begin{vmatrix} \dfrac{\partial}{\partial x} h_1(x,y) & \dfrac{\partial}{\partial x} h_2(x,y) \\ \dfrac{\partial}{\partial y} h_1(x,y) & \dfrac{\partial}{\partial y} h_2(x,y) \end{vmatrix}.$$

Here $D = \mathbb{R}^2$ and we have:

$$u = h_1(x, y) = x \quad \text{and} \quad v = h_2(x, y) = \frac{y - \rho x}{\sqrt{1 - \rho^2}}$$

hence:

$$J_{\mathbf{h}}(x, y) = \begin{vmatrix} 1 & -\rho/\sqrt{1-\rho^2} \\ 0 & 1/\sqrt{1-\rho^2} \end{vmatrix} \quad \Rightarrow \quad |J_{\mathbf{h}}(x, y)| = \frac{1}{\sqrt{1 - \rho^2}}.$$

Therefore:

$$f_{X,Y}(x, y) = f_{U,V}\left(x, \frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) \frac{1}{\sqrt{1 - \rho^2}}.$$

4

4.11.4 Generic bivariate normal distribution

Let $X \sim N(0, 1)$ and $Y \sim N(0, 1)$. If $X^* = \mu_X + \sigma_X X$ and $Y^* = \mu_Y + \sigma_Y Y$ then $X^* \sim N(\mu_X, \sigma_X^2)$ and $Y^* \sim N(\mu_Y, \sigma_Y^2)$ with $\text{Corr}(X^*, Y^*) = \rho$. The joint pdf is:

$$f_{X^*,Y^*}(x^*, y^*) = f_{X,Y}\left(\frac{x^* - \mu_X}{\sigma_X}, \frac{y^* - \mu_Y}{\sigma_Y}\right) \frac{1}{\sigma_X \sigma_Y}.$$

A generic jointly normal random vector is distributed as:

$$\begin{pmatrix} X^* \\ Y^* \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \sigma_Y^2 \end{pmatrix}\right).$$

4.11.5 Joint normality and independence

Independent normal random variables are jointly normally distributed. However, a pair of jointly normally distributed random variables need not necessarily be independent.

While it is true that the marginals of a multivariate normal distribution are also normal, it is not true in general that given two normal random variables their joint distribution is normal.

In general, random variables may be uncorrelated but highly dependent, but if a random vector has a multivariate normal distribution then any two or more of its components which are uncorrelated are independent, this implies that any two or more of its components which are pairwise independent are independent.

However, it is not true that two random variables which are marginally normally distributed and uncorrelated are independent. It is possible for two random variables to be distributed jointly in such a way that each one alone is marginally normally distributed, and they are uncorrelated, but they are not independent.

Example 4.21 Let $X \sim N(0, 1)$, and define:

$$Y = \begin{cases} X & \text{for } |X| > c \\ -X & \text{for } |X| < c \end{cases}$$

where c is a positive number to be specified. If c is very small, then $\text{Corr}(X, Y) \approx 1$. If c is very large, then $\text{Corr}(X, Y) \approx -1$. Since the correlation is a continuous

function of c, there is some particular value of c which makes the correlation 0. This value is approximately 1.54, in which case X and Y are uncorrelated, but they are clearly not independent, since X completely determines Y. Furthermore, Y is normally distributed. Indeed, its distribution is the same as that of X. We use cdfs:

$$\begin{aligned} P(Y \leq x) &= P((|X| < c \cap -X < x) \cup (|X| > c \cap X < x)) \\ &= P((|X| < c \cap X > -x)) + P((|X| > c \cap X < x)) \\ &= P((|X| < c \cap X < x)) + P((|X| > c \cap X < x)) \end{aligned}$$

where the last row depends on the fact that for a symmetric distribution $P(X < x) = P(X > -x)$. Since the events $\{|X| < c\}$ and $\{|X| > c\}$ are a partition of the sample space which is $\mathbb{R}$, then:

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

hence Y is a standard normal random variable as well.

Finally, note that the sum $X + Y$ for $c = 1.54$ has a substantial probability (about 0.88) of it being equal to 0, whereas the normal distribution, being a continuous distribution, has no discrete part, i.e. it does not concentrate more than zero probability at any single point. Consequently, X and Y are not jointly normally distributed, even though they are marginally normally distributed.

4.12 A reminder of your learning outcomes

On completion of this chapter, you should be able to:

- define the terms 'joint' and 'marginal'
- generate the marginal distribution function of a random variable from a joint distribution function involving the variable
- construct a table providing the joint and marginal mass functions for a discrete bivariate distribution
- determine whether a function of two random variables is a valid mass or density function
- derive the marginal mass or density function from a joint mass or density function (even when the support is an 'awkward' shape)
- calculate probabilities from a bivariate mass or density function
- calculate expectations for functions of several variables
- calculate joint moments from a bivariate mass or density function
- explain the distinction between pairwise and mutual independence
- explain the distinction between uncorrelated and independent random variables

- determine that the variance matrix of a random vector is non-negative definite
- derive the density function of a transformation of continuous random variables
- derive the distribution of the sum of random variables
- construct a standard bivariate normal distribution
- derive the density function of a linear transformation of independent normal random variables.

4.13 Sample examination questions

Solutions can be found in Appendix C.

1. The joint probability density function of the random variables X and Y is given by:

$$f_{X,Y}(x,y) = \begin{cases} a(x+y) & \text{for } 0 < x < 2,\ 0 < y < 2 \\ 0 & \text{otherwise.} \end{cases}$$

   (a) Find the value of a.

   (b) Find the marginal cumulative distribution function $F_X(x)$, and the mean $\mathrm{E}(X)$, of X.

2. The joint probability density function of the random variables X and Y is given by:

$$f_{X,Y}(x,y) = \begin{cases} axy^2 & \text{for } 0 < x < y < 2 \\ 0 & \text{otherwise.} \end{cases}$$

   (a) Show that $a = 5/16$.

   (b) Let $U = X + Y$, $V = (Y + X)/(Y - X)$. Show that the joint probability density function of U, V is given by:

$$f_{U,V}(u,v) = \frac{5u^4}{256v^5}(v-1)(v+1)^2$$

   for $0 < u < 4v/(1+v)$ and $v > 1$, and 0 otherwise.

   (c) Find an expression for $\sqrt{Y/(Y-X)}$ in terms of U, V. Given that the marginal density function of V is:

$$f_V(v) = \frac{4(v-1)}{(v+1)^3}$$

   for $v > 1$, and 0 otherwise, show that:

$$\mathrm{E}\left(\sqrt{\frac{Y}{Y-X}}\right) = \frac{8}{3}.$$

3. The joint probability density function of the random variables X and Y is given by:

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{axy}{(x+y)^{3/2}} & \text{for } 0 < x < y < 2 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Let $U = X/(X+Y)$ and $V = X+Y$. Show that the joint probability density function of U, V is given by:

$$f_{U,V}(u,v) = au(1-u)v^{3/2}$$

for $0 < v < 2/(1-u)$ and $0 < u < 1/2$, and 0 otherwise.

(b) You are given that:

$$\int_0^{1/2} u(1-u)^{-3/2}\,\mathrm{d}u = 3\sqrt{2} - 4.$$

Show that:

$$a = \frac{5}{48 - 32\sqrt{2}}.$$

(c) You are given that:

$$\int_0^{1/2} u^2(1-u)^{-5/2}\,\mathrm{d}u = \frac{16 - 11\sqrt{2}}{3}.$$

Find $\mathrm{E}(X)$.

Chapter 5
Conditional distributions

5.1 Recommended reading

Casella, G. and R.L. Berger, *Statistical Inference*. Chapter 4, Sections 4.2 and 4.4.

5.2 Learning outcomes

On completion of this chapter, you should be able to:

- derive conditional mass and density functions from joint mass and density functions
- state the relationships between joint, marginal and conditional distributions, and use these relationships to solve problems
- calculate conditional expectations of a function of a random variable
- apply the law of iterated expectations
- prove that variance can be decomposed as the sum of the expected value of the conditional variance and the variance of the conditional expectation
- derive and use conditional moment generating functions and conditional cumulant generating functions
- solve problems involving hierarchies, mixtures and random sums.

5.3 Introduction

In multivariate settings, the value of one random variable may be related to the value of one or more other variables. Conditional probabilities allow us to improve our knowledge of one random variable based on the information available to us about one or more other variables.

5.4 Discrete and continuous conditional distributions

We begin by defining the conditional cumulative distribution function. We treat $\{X = x\}$ as an event and, provided $P(X = x) > 0$, we can then condition on the event $\{X = x\}$ as usual.

Conditional cumulative distribution function

Let X and Y be random variables such that $P(X = x) > 0$. The distribution of Y conditional on $X = x$ is given by the **conditional cumulative distribution function**:

$$F_{Y|X}(y \,|\, x) = P(Y \leq y \,|\, X = x).$$

This distribution may vary for different values of X, hence this defines a family of distributions.

When X and Y are discrete random variables the conditional (probability) mass functions return conditional probabilities as their values. Note that the relationship between the conditional (probability) mass function and conditional (cumulative) distribution function is the same as for the unconditional case.

Conditional probability mass function and distribution function

Let X and Y be discrete random variables such that $P(X = x) > 0$. The **conditional probability mass function** of Y given $X = x$ is:

$$p_{Y|X}(y \,|\, x) = P(Y = y \,|\, X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)} = \frac{p_{X,Y}(x, y)}{\sum\limits_y p_{X,Y}(x, y)}$$

such that the conditional cdf is:

$$F_{Y|X}(y \,|\, x) = \sum_{y_i \leq y} p_{Y|X}(y_i \,|\, x) = \sum_{y_i \leq y} \frac{p_{X,Y}(x, y)}{p_X(x)} = \frac{P(Y \leq y, X = x)}{P(X = x)}$$

$$= P(Y \leq y \,|\, X = x).$$

When X and Y are continuous random variables, we similarly define the conditional (probability) density function and the conditional (cumulative) distribution function.

Conditional probability density function and distribution function

Let X and Y be jointly continuous random variables such that $f_X(x) > 0$. The **conditional probability density function** of Y given $X = x$ is:

$$f_{Y|X}(y \,|\, x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{f_{X,Y}(x, y)}{\int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}y}$$

such that the conditional cdf is:

$$F_{Y|X}(y \,|\, x) = \int_{-\infty}^{y} \frac{f_{X,Y}(x, v)}{f_X(x)}\,\mathrm{d}v.$$

Recall that if X is a continuous random variable then $P(X = x) = 0$, hence the conditional density function does *not* return a conditional probability.

Activity 5.1 Consider the game in Activity 3.11. X is the number of heads from tossing a coin three times and Y takes the value 1 if the first and third throws are heads, and 0 otherwise. Suppose we are told that the first throw is a head. Write down the mass functions of X and Y conditional on this event.

Activity 5.2 Suppose that X and Y are jointly continuous random variables with joint density function $f_{X,Y}$ and the marginal density function of X is f_X. Let x be a real number with $f_X(x) > 0$. Show that:

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

is a valid (conditional) density function.

Activity 5.3 Suppose the random variables X and Y are jointly normal with $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ and $\text{Corr}(X, Y) = \rho$. Show that:

$$Y \,|\, X = x \sim N\left(\mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X), \sigma_Y^2(1 - \rho^2)\right).$$

Activity 5.4 Suppose the random variables X and Y have the joint density function:

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

Derive expressions for $f_{Y|X}(y \,|\, x)$ and $f_{X|Y}(x \,|\, y)$.

We have the following relationships connecting joint, marginal and conditional distributions, all of which are direct implications of Bayes' theorem.

1. The joint probability mass function is:

   $$p_{X,Y}(x, y) = p_{Y|X}(y \,|\, x)\, p_X(x)$$

   and the joint probability density function is:

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

   Note that by symmetry:

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

2. The marginal probability mass function is:

   $$p_Y(y) = \sum_x p_{Y|X}(y \,|\, x)\, p_X(x)$$

   and the marginal probability density function is:

   $$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y \,|\, x)\, f_X(x)\, \mathrm{d}x.$$

3. Reverse conditioning (provided $f_Y(y) > 0$):

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

When X and Y are independent random variables, the conditional distributions are the same as the respective marginal distributions.

Conditional distributions and independence

Suppose X and Y are independent random variables. For the cdfs we have:

$$F_{Y|X}(y \,|\, x) = F_Y(y) \;\; \forall\, x, y \in \mathbb{R} \quad \text{and} \quad F_{X|Y}(x \,|\, y) = F_X(x) \;\; \forall\, x, y \in \mathbb{R}.$$

For the density functions we have:

$$f_{Y|X}(y \,|\, x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x)\, f_Y(y)}{f_X(x)} = f_Y(y) \quad \forall\, x, y \in \mathbb{R}$$

and:

$$f_{X|Y}(x \,|\, y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_X(x)\, f_Y(y)}{f_Y(x)} = f_X(x) \quad \forall\, x, y \in \mathbb{R}.$$

Similarly, for the mass functions.

Example 5.1 Let X denote the number of hurricanes which form in a given year, and let Y denote the number of these which make landfall. Suppose each hurricane has a probability π of making landfall independent of other hurricanes. Given the number of hurricanes x, then Y can be thought of as the number of successes in x independent and identically distributed Bernoulli trials. We can write this as $Y \,|\, X = x \sim \text{Bin}(x, \pi)$. Suppose we also have that $X \sim \text{Pois}(\lambda)$. We can now determine the distribution of Y (noting that $X \geq Y$). We have:

$$\begin{aligned}
p_Y(y) = \sum_{x=y}^{\infty} p_{Y|X}(y \,|\, x)\, p_X(x) &= \sum_{x=y}^{\infty} \frac{x!}{y!\,(x-y)!}\, \pi^y (1-\pi)^{x-y}\, \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} \\
&= \frac{\mathrm{e}^{-\lambda}\lambda^y \pi^y}{y!} \sum_{x=y}^{\infty} \frac{(\lambda(1-\pi))^{x-y}}{(x-y)!} \\
&= \frac{\mathrm{e}^{-\lambda}\lambda^y \pi^y}{y!} \sum_{j=0}^{\infty} \frac{(\lambda(1-\pi))^{j}}{j!} \\
&= \frac{\mathrm{e}^{-\lambda}\lambda^y \pi^y}{y!}\, \mathrm{e}^{\lambda(1-\pi)} \\
&= \frac{\mathrm{e}^{-\lambda\pi}(\lambda\pi)^y}{y!}.
\end{aligned}$$

Therefore, $Y \sim \text{Pois}(\lambda\pi)$. Hence $\text{E}(Y) = \lambda\pi$ and $\text{Var}(Y) = \lambda\pi$.

Example 5.2 Suppose that the number of insurance claims which a particular policyholder makes in one year has a Poisson distribution with mean Λ, and that over the large population of policyholders Λ has a gamma distribution with parameters α and β. Find the probability of x claims in one year by a policyholder chosen at random from the population of policyholders.

Solution

The density function of λ, for $\lambda > 0$, is:

$$f_\Lambda(\lambda) = \frac{1}{\Gamma(\alpha)}\,\beta^\alpha \lambda^{\alpha-1}\mathrm{e}^{-\beta\lambda}$$

where α and β are positive. If X is the number of claims made by a policyholder, the marginal distribution of X is given by:

$$\begin{aligned}
f_X(x) &= \int_0^\infty f_{X|\Lambda}(x \mid \lambda)\, f_\Lambda(\lambda)\,\mathrm{d}\lambda \\
&= \int_0^\infty \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}\,\frac{\beta^\alpha\lambda^{\alpha-1}\mathrm{e}^{-\beta\lambda}}{\Gamma(\alpha)}\,\mathrm{d}\lambda \\
&= \frac{\Gamma(\alpha+x)\beta^\alpha}{(1+\beta)^{\alpha+x}x!\,\Gamma(\alpha)}\int_0^\infty \frac{(1+\beta)^{\alpha+x}}{\Gamma(\alpha+x)}\,\lambda^{\alpha+x-1}\mathrm{e}^{-\lambda(1+\beta)}\,\mathrm{d}\lambda \\
&= \frac{\Gamma(\alpha+x)\beta^\alpha}{(1+\beta)^{\alpha+x}x!\,\Gamma(\alpha)} \\
&= \frac{(\alpha+x-1)!}{x!\,(\alpha-1)!}\left(\frac{\beta}{1+\beta}\right)^\alpha\left(\frac{1}{1+\beta}\right)^x \\
&= \binom{x+\alpha-1}{\alpha-1}\left(\frac{\beta}{1+\beta}\right)^\alpha\left(\frac{1}{1+\beta}\right)^x
\end{aligned}$$

for $x = 0, 1, 2, \ldots$, when α is an integer such that $\alpha \geq 1$. Note that the integrand is a Gamma$(\alpha + x, 1 + \beta)$ density function, hence integrates to 1. The final result is a negative binomial distribution (second version, see Section 3.7.6) with parameters $r = \alpha$ and $\pi = \beta/(1 + \beta)$.

Activity 5.5 Let Y represent the number of insurance claims made by an insurance customer in one year. Assume that $Y \mid X = x \sim \text{Pois}(x)$, where $X \sim \text{Gamma}(\alpha, \beta)$. What is the probability that the customer makes y claims in a year?

Activity 5.6 Let Y represent the number of minutes that a bus is late arriving at a bus stop. The lateness of the bus depends on the volume of traffic, X. Specifically, $Y \mid X = x \sim \text{Exp}(x)$, where $X \sim \text{Gamma}(\alpha, \beta)$. Derive the density function of Y.

5.5 Conditional expectations, moments and mgfs

5.5.1 Conditional expectations

Conditional expectations for discrete and continuous random variables are obtained by using the appropriate conditional distribution. The definitions and results are analogous to the unconditional cases. One important distinction is that while unconditional expected values are simply numbers, conditional expectations are *random variables*.

Conditional expectation

Let X and Y be two random variables. The **conditional expectation** of Y given $X = x$ is:

$$\mathrm{E}(Y \mid X = x) = \begin{cases} \sum_y y\, p_{Y|X}(y \mid x) & \text{(discrete case)} \\ \int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x)\, \mathrm{d}y & \text{(continuous case).} \end{cases}$$

If we were to consider all possible values of X then we define a new random variable $\mathrm{E}(Y \mid X)$, i.e. the conditional expectation of Y given X. This represents the best assessment of Y given our knowledge of X. All properties of expectations still hold.

Law of iterated expectations

As $\mathrm{E}(Y \mid X)$ is a random variable its expectation may be taken, which is:

$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y \mid X)).$$

Proof: In the continuous case:

$$\begin{aligned} \mathrm{E}(\mathrm{E}(Y \mid X)) = \mathrm{E}(g(X)) &= \int_{-\infty}^{\infty} \mathrm{E}(Y \mid X = x)\, f_X(x)\, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x)\, \mathrm{d}y \right) f_X(x)\, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, \frac{f_{X,Y}(x,y)}{f_X(x)}\, f_X(x)\, \mathrm{d}y\, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, f_{X,Y}(x,y)\, \mathrm{d}y\, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, \mathrm{d}x\, \mathrm{d}y \\ &= \int_{-\infty}^{\infty} y\, f_Y(y)\, \mathrm{d}y \\ &= \mathrm{E}(Y). \end{aligned}$$

■

A useful consequence of the law of iterated expectations is that we are able to compute

$\mathrm{E}(Y)$ without explicitly requiring the marginal distribution of Y, since:

$$\mathrm{E}(Y) = \begin{cases} \sum_x \mathrm{E}(Y \mid X = x)\, p_X(x) & \text{(discrete case)} \\ \int_{-\infty}^{\infty} \mathrm{E}(Y \mid X = x)\, f_X(x)\, \mathrm{d}x & \text{(continuous case).} \end{cases}$$

5.5.2 Conditional moments

We first consider how to determine the conditional expectation of a function of a random variable.

Conditional expectations of functions of random variables

Suppose g is a well-behaved, real-valued function. The **conditional expectation** of $g(Y)$ given $X = x$ is defined as:

$$\mathrm{E}(g(Y) \mid X = x) = \begin{cases} \sum_y g(y)\, p_{Y|X}(y \mid x) & \text{(discrete case)} \\ \int_{-\infty}^{\infty} g(y)\, f_{Y|X}(y \mid x)\, \mathrm{d}y & \text{(continuous case).} \end{cases}$$

The conditional expectation of $g(Y)$ given X is denoted $\mathrm{E}(g(Y) \mid X)$ and is also a random variable.

A useful consequence is that any function of X can be treated as a constant with respect to expectations which are conditional on X. In general, for well-behaved functions g_1 and g_2 we have:

$$\mathrm{E}(g_1(X)\, g_2(Y) \mid X) = g_1(X)\, \mathrm{E}(g_2(Y) \mid X).$$

Example 5.3 When conditioning on X, we have:

$$\mathrm{E}(XY \mid X) = X\, \mathrm{E}(Y \mid X).$$

Also, since $\mathrm{E}(Y \mid X)$ is a function of X, then:

$$\mathrm{E}(\mathrm{E}(Y \mid X)\, Y \mid X) = \mathrm{E}(Y \mid X)\, \mathrm{E}(Y \mid X) = (\mathrm{E}(Y \mid X))^2.$$

Note that any conditional expectation is a function of the random variable being conditioned on.

Example 5.4 Let X and Y be two random variables with joint probability density function:

$$f_{X,Y}(x, y) = \begin{cases} x\mathrm{e}^{-xy}\mathrm{e}^{-x} & \text{for } x > 0 \text{ and } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

We want to find the conditional expectation $\mathrm{E}(Y \mid X)$. We begin by calculating the marginal density function of X, which is:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}y = \int_0^{\infty} x\mathrm{e}^{-xy}\mathrm{e}^{-x}\, \mathrm{d}y = \mathrm{e}^{-x}\Big[-\mathrm{e}^{-xy}\Big]_0^{\infty} = \mathrm{e}^{-x}$$

for $x > 0$, and 0 otherwise. Hence:

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{x\mathrm{e}^{-xy}\mathrm{e}^{-x}}{\mathrm{e}^{-x}} = x\mathrm{e}^{-xy}$$

for $y > 0$, and 0 otherwise. Therefore:

$$Y \mid X = x \sim \mathrm{Exp}(x).$$

We conclude that:

$$\mathrm{E}(Y \mid X = x) = \frac{1}{x} \quad \Rightarrow \quad \mathrm{E}(Y \mid X) = \frac{1}{X}.$$

5

Example 5.5 Let X and Y be two random variables with joint probability density function:

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

(a) Find $f_X(x)$ and hence $\mathrm{E}(X)$, $\mathrm{E}(X^2)$ and $\mathrm{Var}(X)$.

(b) Find $f_{Y|X}(y \mid x)$ and hence $\mathrm{E}(Y \mid X)$.

(c) Evaluate $\mathrm{E}(Y)$ and $\mathrm{Cov}(X,Y)$.

Solution

(a) We have:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \int_0^x 6y\,\mathrm{d}y = \Big[3y^2\Big]_0^x = 3x^2.$$

In full:

$$f_X(x) = \begin{cases} 3x^2 & \text{for } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Hence:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f_X(x)\,\mathrm{d}x = \int_0^x 3x^3\,\mathrm{d}x = \left[\frac{3}{4}x^4\right]_0^1 = \frac{3}{4}$$

and:

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x = \int_0^x 3x^4\mathrm{d}x = \left[\frac{3}{5}x^5\right]_0^1 = \frac{3}{5}$$

hence:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{3}{5} - \left(\frac{3}{4}\right)^2 = \frac{3}{80}.$$

(b) The conditional density function is given by:

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{6y}{3x^2} = \frac{2y}{x^2}.$$

In full:

$$f_{Y|X}(y \mid x) = \begin{cases} 2y/x^2 & \text{for } 0 < y < x \\ 0 & \text{otherwise.} \end{cases}$$

Using this conditional density function, we can now calculate the conditional expectation:

$$\mathrm{E}(Y \mid X = x) = \int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x)\, \mathrm{d}y = \int_0^x \frac{2y^2}{x^2}\, \mathrm{d}y = \left[\frac{2y^3}{3x^2}\right]_0^x = \frac{2}{3}\, x.$$

Hence:

$$\mathrm{E}(Y \mid X) = \frac{2}{3}\, X.$$

(c) We could work out $\mathrm{E}(Y)$ by finding the marginal density function $f_Y(y)$ and then integrating. However, using iterated expectations avoids additional integration. We have:

$$\begin{aligned} \mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y \mid X)) = \mathrm{E}\left(\frac{2}{3}\, X\right) &= \frac{2}{3}\, \mathrm{E}(X) \\ &= \frac{2}{3} \times \frac{3}{4} \\ &= \frac{1}{2}. \end{aligned}$$

Now we may exploit iterated expectations to calculate $\mathrm{Cov}(X, Y)$ by noting that:

$$\begin{aligned} \mathrm{E}(XY) = \mathrm{E}(\mathrm{E}(XY \mid X)) = \mathrm{E}(X\, \mathrm{E}(Y \mid X)) = \mathrm{E}\left(\frac{2}{3}\, X^2\right) \\ &= \frac{2}{3} \times \frac{3}{5} \\ &= \frac{2}{5}. \end{aligned}$$

Hence:

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathrm{E}(XY) - \mathrm{E}(X)\, \mathrm{E}(Y) \\ &= \frac{2}{5} - \frac{3}{4} \times \frac{1}{2} \\ &= \frac{1}{40}. \end{aligned}$$

Example 5.6 The random variables X and Y have the joint probability density function:

$$f_{X,Y}(x, y) = \begin{cases} k(1 - y) & \text{for } 0 < x < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find k.

(b) Evaluate $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

(c) Derive the conditional density function $f_{Y|X}(y\,|\,x)$ and the conditional expectation $\mathrm{E}((1-Y)\,|\,X)$. Hence evaluate $\mathrm{E}(Y)$.

(d) Evaluate $P(Y < 2X)$.

Solution

(a) We have:

$$\begin{aligned} 1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y &= \int_0^1\int_0^y k(1-y)\,\mathrm{d}x\,\mathrm{d}y \\ &= k\int_0^1 \Big[x(1-y)\Big]_0^y\,\mathrm{d}y \\ &= k\int_0^1 (y-y^2)\,\mathrm{d}y \\ &= k\left[\frac{y^2}{2}-\frac{y^3}{3}\right]_0^1 \\ &= \frac{k}{6} \end{aligned}$$

hence $k = 6$.

(b) We have:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \int_x^1 6(1-y)\,\mathrm{d}y = \Big[6y-3y^2\Big]_x^1 = 3(1-x)^2.$$

In full:

$$f_X(x) = \begin{cases} 3(1-x)^2 & \text{for } 0<x<1 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\,f_X(x)\,\mathrm{d}x = \int_0^1 3x(1-x)^2\,\mathrm{d}x = \frac{1}{4}.$$

Also:

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x = \int_0^1 3x^2(1-x)^2\,\mathrm{d}x = \frac{1}{10}$$

hence:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{1}{10} - \left(\frac{1}{4}\right)^2 = \frac{3}{80}.$$

(c) We have:

$$f_{Y|X}(y\,|\,x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{6(1-y)}{3(1-x)^2} = \begin{cases} 2(1-y)/(1-x)^2 & \text{for } x<y<1 \\ 0 & \text{otherwise.} \end{cases}$$

We have:

$$\mathrm{E}((1-Y)\,|\,X=x) = \int_x^1 (1-y) f_{Y|X}(y\,|\,x)\,\mathrm{d}y = \int_x^1 \frac{2(1-y)^2}{(1-x)^2}\,\mathrm{d}y = \frac{2}{3}(1-x).$$

Therefore:

$$\mathrm{E}((1-Y)\,|\,X) = \frac{2}{3}(1-X).$$

So:

$$1-\mathrm{E}(Y\,|\,X) = \frac{2}{3}(1-X) \quad \Rightarrow \quad \mathrm{E}(Y\,|\,X) = \frac{1}{3}+\frac{2}{3}X$$

and applying the law of iterated expectations:

$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y\,|\,X)) = \frac{1}{2}.$$

(d) We have:

$$P(Y<2X) = \int_0^1 \int_{y/2}^y 6(1-y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^1 3(1-y)y\,\mathrm{d}y = \frac{1}{2}.$$

Example 5.7 Let X and Y be continuous random variables with joint density function:

$$f_{X,Y}(x,y) = \begin{cases} kxy & \text{for } 0<x<y<2 \\ 0 & \text{otherwise} \end{cases}$$

for some $k>0$.

(a) Find k.

(b) Are X and Y independent? Explain your answer.

(c) Find $\mathrm{E}(X\,|\,Y)$ and $\mathrm{E}(X)$.

(d) Determine $\mathrm{Cov}(X,Y)$.

Solution

(a) We have:

$$\begin{aligned} 1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^2\int_0^y kxy\,\mathrm{d}x\,\mathrm{d}y &= \int_0^2 ky\int_0^y x\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^2 ky\left[\frac{x^2}{2}\right]_0^y \mathrm{d}y \\ &= \frac{k}{2}\int_0^2 y^3\mathrm{d}y \\ &= \frac{k}{2}\left[\frac{y^4}{4}\right]_0^2 \\ &= 2k \end{aligned}$$

hence $k = 1/2$.

(b) The marginal density functions are:

$$f_X(x) = \int_x^2 \frac{1}{2}\, xy \,\mathrm{d}y = \frac{x}{2}\int_x^2 y \,\mathrm{d}y = \frac{x}{2}\left[\frac{y^2}{2}\right]_x^2 = x - \frac{x^3}{4}$$

and:

$$f_Y(y) = \int_0^y \frac{1}{2}\, xy \,\mathrm{d}x = \frac{y}{2}\int_0^y x \,\mathrm{d}x = \frac{y}{2}\left[\frac{x^2}{2}\right]_0^y = \frac{y^3}{4}.$$

Since:

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

then X and Y are not independent.

(c) For the conditional expectation of X we first need the conditional density function, which is:

$$f_{X|Y}(x \,|\, y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{xy/2}{y^3/4} = \frac{2x}{y^2}$$

for $0 < x < y < 2$, and 0 otherwise. Therefore:

$$\begin{aligned}
\mathrm{E}(X \,|\, Y = y) = \int_{-\infty}^{\infty} x\, f_{X|Y}(x \,|\, y)\,\mathrm{d}x &= \int_0^y x\,\frac{2x}{y^2}\,\mathrm{d}x \\
&= \frac{2}{y^2}\int_0^y x^2\,\mathrm{d}x \\
&= \frac{2}{y^2}\left[\frac{x^3}{3}\right]_0^y \\
&= \frac{2}{3}y.
\end{aligned}$$

Hence $\mathrm{E}(X \,|\, Y) = 2Y/3$. Now, applying the law of iterated expectations:

$$\mathrm{E}(X) = \mathrm{E}(\mathrm{E}(X \,|\, Y)) = \frac{2}{3}\,\mathrm{E}(Y)$$

and:

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} y\, f_Y(y)\,\mathrm{d}y = \int_0^2 y\,\frac{y^3}{4}\,\mathrm{d}y = \int_0^2 \frac{y^4}{4}\,\mathrm{d}y = \left[\frac{y^5}{20}\right]_0^2 = \frac{8}{5}$$

hence $\mathrm{E}(X) = (2/3) \times (8/5) = 16/15$.

(d) We have:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) \\
&= \mathrm{E}(\mathrm{E}(XY \,|\, Y)) - \mathrm{E}(\mathrm{E}(X \,|\, Y))\,\mathrm{E}(Y) \\
&= \mathrm{E}(Y\,\mathrm{E}(X \,|\, Y)) - \frac{2}{3}\,(\mathrm{E}(Y))^2 \\
&= \frac{2}{3}\,(\mathrm{E}(Y^2) - (\mathrm{E}(Y))^2).
\end{aligned}$$

Since:

$$\mathrm{E}(Y^2) = \int_{-\infty}^{\infty} y^2 f_Y(y)\,\mathrm{d}y = \int_0^2 y^2 \frac{y^3}{4}\,\mathrm{d}y = \int_0^2 \frac{y^5}{4}\,\mathrm{d}y = \left[\frac{y^6}{24}\right]_0^2 = \frac{16}{6}$$

then:

$$\mathrm{Cov}(X,Y) = \frac{2}{3}\left(\frac{16}{6} - \left(\frac{8}{5}\right)^2\right) = \frac{16}{225}.$$

Example 5.8 Consider the joint pdf:

$$f_{X,Y}(x,y) = \begin{cases} x+y & \text{for } 0<x<1 \text{ and } 0<y<1 \\ 0 & \text{otherwise.} \end{cases}$$

This is a valid joint pdf because it is a positive real-valued function and it is normalised since:

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y &= \int_0^1\int_0^1 (x+y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^1 \left[\frac{x^2}{2} + xy\right]_0^1 \mathrm{d}y \\
&= \int_0^1 \left(\frac{1}{2} + y\right)\mathrm{d}y \\
&= \left[\frac{y}{2} + \frac{y^2}{2}\right]_0^1 \\
&= 1.
\end{aligned}$$

The joint cdf is:

$$\begin{aligned}
F_{X,Y}(x,y) &= \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\,\mathrm{d}u\,\mathrm{d}v \\
&= \int_0^y\int_0^x (u+v)\,\mathrm{d}u\,\mathrm{d}v \\
&= \int_0^y \left[\frac{u^2}{2} + uv\right]_0^x \mathrm{d}v \\
&= \int_0^y \left(\frac{x^2}{2} + xv\right)\mathrm{d}v \\
&= \left[\frac{x^2 v}{2} + \frac{xv^2}{2}\right]_0^y \\
&= \frac{xy(x+y)}{2} \quad \text{for } 0<x<1,\ 0<y<1.
\end{aligned}$$

In full we have:

$$F_{X,Y}(x,y) = \begin{cases} xy(x+y)/2 & \text{for } 0 < x < 1 \text{ and } 0 < y < 1 \\ x(x+1)/2 & \text{for } 0 < x < 1 \text{ and } y \geq 1 \\ y(y+1)/2 & \text{for } x \geq 1 \text{ and } 0 < y < 1 \\ 1 & \text{for } x \geq 1 \text{ and } y \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The marginal pdf of X is:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \int_0^1 (x+y)\,\mathrm{d}y = x + \frac{1}{2}$$

for $0 < x < 1$, and 0 otherwise.

Probabilities such as $P(2X < Y)$ may be computed. First, define the event $A = \{(x,y) : 0 < x < y/2,\ 0 < y < 1\}$. Hence:

$$\begin{aligned} P(2X < Y) = P(A) &= \iint_B f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^1 \int_0^{y/2} (x+y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^1 \left[\frac{y^2}{8} + \frac{y^2}{2}\right] \mathrm{d}y \\ &= \left[\frac{y^3}{24} + \frac{y^3}{6}\right]_0^1 \\ &= \frac{5}{24}. \end{aligned}$$

Similarly, we could have defined the event $B = \{(x,y) : 0 < x < 1/2,\ 2x < y < 1\}$ and computed $P(B)$.

The (r,s)th joint moment is:

$$\begin{aligned} \mathrm{E}(X^r Y^s) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^r y^s f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^1 \int_0^1 x^r y^s (x+y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^1 \left(\frac{y^s}{r+2} + \frac{y^{s+1}}{r+1}\right) \mathrm{d}y \\ &= \left[\frac{y^{s+1}}{(r+2)(s+1)} + \frac{y^{s+2}}{(r+1)(s+2)}\right]_0^1 \\ &= \frac{1}{(r+2)(s+1)} + \frac{1}{(r+1)(s+2)}. \end{aligned}$$

Hence $\mathrm{E}(XY) = 1/3$, $\mathrm{E}(X) = \mathrm{E}(Y) = 7/12$, $\mathrm{E}(X^2) = 5/12$, hence $\mathrm{Var}(X) = 11/144$. Also:

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = \frac{1}{3} - \frac{49}{144} = -\frac{1}{144}$$

and so $\mathrm{Corr}(X, Y) = -1/11$, hence X and Y are not independent.

We can also deduce that X and Y are not independent by observing that:

$$f_X(x)\, f_Y(y) = xy + \frac{x+y}{2} + \frac{1}{4} \neq f_{X,Y}(x, y).$$

The conditional pdf of Y given $X = x$ is:

$$f_{Y|X}(y \,|\, x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \begin{cases} (x+y)/(x+1/2) & \text{for } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The conditional expectation of Y given $X = x$ is:

$$\begin{aligned} \mathrm{E}(Y \,|\, X = x) &= \int_{-\infty}^{\infty} y\, f_{Y|X}(y \,|\, x)\, \mathrm{d}y \\ &= \int_0^1 y\, \frac{x+y}{x+1/2}\, \mathrm{d}y \\ &= \frac{1}{x+1/2} \left[\frac{xy^2}{2} + \frac{y^3}{3} \right]_0^1 \\ &= \frac{3x+2}{6x+3}. \end{aligned}$$

Applying the law of iterated expectations, we have:

$$\begin{aligned} \mathrm{E}(\mathrm{E}(Y \,|\, X = x)) &= \int_0^1 \frac{3x+2}{6x+3} \left(x + \frac{1}{2} \right) \mathrm{d}x \\ &= \frac{1}{6} \int_0^1 (3x+2)\, \mathrm{d}x \\ &= \frac{1}{6} \left(\frac{3}{2} + 2 \right) \\ &= \frac{7}{12} \\ &= \mathrm{E}(Y). \end{aligned}$$

Example 5.9 Consider the random variables X and Y with joint density function:

$$f_{X,Y}(x, y) = \begin{cases} \beta(\alpha+\beta)\mathrm{e}^{-\alpha x}\mathrm{e}^{-\beta y} & \text{for } 0 < x < y < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Write down $f_{Y|X}(y \,|\, x)$. Hence find $\mathrm{E}(Y \,|\, X)$, $\mathrm{Var}(\mathrm{E}(Y \,|\, X))$, and $\mathrm{Cov}(X, Y)$.

Solution

To perform the conditioning we start by evaluating the marginal density of X. We have:

$$\begin{aligned} f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y &= \int_x^{\infty} \beta(\alpha+\beta)\mathrm{e}^{-\alpha x}\mathrm{e}^{-\beta y}\,\mathrm{d}y \\ &= \Big[-(\alpha+\beta)\mathrm{e}^{-\alpha x}\mathrm{e}^{-\beta y}\Big]_x^{\infty} \\ &= (\alpha+\beta)\mathrm{e}^{-(\alpha+\beta)x} \end{aligned}$$

for $x > 0$, and 0 otherwise. Therefore, the required conditional density function is:

$$f_{Y|X}(y\,|\,x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \begin{cases} \beta\mathrm{e}^{-\beta(y-x)} & \text{for } x < y < \infty \\ 0 & \text{otherwise.} \end{cases}$$

For the conditional expectation, we have (using integration by parts):

$$\begin{aligned} \mathrm{E}(Y\,|\,X=x) = \int_{-\infty}^{\infty} y\, f_{Y|X}(y\,|\,x)\,\mathrm{d}y &= \int_x^{\infty} \beta y\mathrm{e}^{-\beta(y-x)}\,\mathrm{d}y \\ &= \beta\mathrm{e}^{\beta x}\left[-y\frac{\mathrm{e}^{-\beta y}}{\beta}\right]_x^{\infty} - \int_x^{\infty}\mathrm{e}^{-\beta y}\,\mathrm{d}y \\ &= x + \frac{1}{\beta} \end{aligned}$$

hence:

$$\mathrm{E}(Y\,|\,X) = X + \frac{1}{\beta}.$$

Therefore:

$$\mathrm{Var}(\mathrm{E}(Y\,|\,X)) = \mathrm{Var}(X) = \frac{1}{(\alpha+\beta)^2}$$

since $X \sim \mathrm{Exp}(\alpha+\beta)$. So $\mathrm{E}(X) = 1/(\alpha+\beta)$ and:

$$\mathrm{E}(Y) = \frac{1}{\beta} + \mathrm{E}(X) = \frac{1}{\beta} + \frac{1}{\alpha+\beta} = \frac{\alpha+2\beta}{(\alpha+\beta)\beta}.$$

Also:

$$\begin{aligned} \mathrm{E}(XY) = \mathrm{E}(X\,\mathrm{E}(Y\,|\,X)) &= \mathrm{E}\left(X\left(\frac{1}{\beta}+X\right)\right) \\ &= \mathrm{E}\left(\frac{X}{\beta}\right) + \mathrm{E}(X^2) \\ &= \frac{1}{\beta}\,\mathrm{E}(X) + \mathrm{Var}(X) + (\mathrm{E}(X))^2 \\ &= \frac{1}{\beta(\alpha+\beta)} + \frac{1}{(\alpha+\beta)^2} + \frac{1}{(\alpha+\beta)^2} \\ &= \frac{\alpha+3\beta}{\beta(\alpha+\beta)^2} \end{aligned}$$

hence:

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = \frac{\alpha+3\beta}{\beta(\alpha+\beta)^2} - \frac{\alpha+2\beta}{\beta(\alpha+\beta)^2} = \frac{1}{(\alpha+\beta)^2}.$$

We now define conditional moments, which themselves are conditional expectations. Conditional moments are functions of the random variables being conditioned on, and hence are also random variables.

Conditional moments and conditional central moments

For two random variables X and Y, the rth **conditional moment** of Y given X is:

$$\mathrm{E}(Y^r \mid X)$$

and the rth **conditional central moment** of Y given X is:

$$\mathrm{E}((Y - \mathrm{E}(Y \mid X))^r \mid X).$$

Setting $r = 2$, the second conditional central moment is a special case, i.e. the conditional variance.

Conditional variance

Let X and Y be two random variables. The **conditional variance** is defined as:

$$\begin{aligned}\mathrm{Var}(Y \mid X = x) &= \mathrm{E}((Y - \mathrm{E}(Y \mid X = x))^2 \mid X = x) \\ &= \begin{cases} \sum_y (y - \mathrm{E}(Y \mid X = x))^2 p_{Y|X}(y \mid x) & \text{(discrete case)} \\ \int_{-\infty}^{\infty} (y - \mathrm{E}(Y \mid X = x))^2 f_{Y|X}(y \mid x)\,\mathrm{d}y & \text{(continuous case).} \end{cases}\end{aligned}$$

The conditional variance of Y given X is denoted $\mathrm{Var}(Y \mid X)$ and is also a random variable. Furthermore:

$$\mathrm{Var}(Y \mid X) = \mathrm{E}(Y^2 \mid X) - (\mathrm{E}(Y \mid X))^2.$$

As $\mathrm{Var}(Y \mid X)$ is a random variable, we may have interest in its expected value. This is:

$$\mathrm{E}(\mathrm{Var}(Y \mid X)) = \mathrm{Var}(Y) - \mathrm{Var}(\mathrm{E}(Y \mid X)).$$

Proof: By using the law of iterated expectations:

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 \\ &= \mathrm{E}(\mathrm{E}(Y^2 \mid X)) - (\mathrm{E}(\mathrm{E}(Y \mid X)))^2 \\ &= \mathrm{E}(\mathrm{Var}(Y \mid X) + (\mathrm{E}(Y \mid X))^2) - (\mathrm{E}(\mathrm{E}(Y \mid X)))^2 \\ &= \mathrm{E}(\mathrm{Var}(Y \mid X)) + \mathrm{E}((\mathrm{E}(Y \mid X))^2) - (\mathrm{E}(\mathrm{E}(Y \mid X)))^2 \\ &= \mathrm{E}(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(\mathrm{E}(Y \mid X)).\end{aligned}$$

Rearranging returns the result. ■

We now highlight the useful result of the decomposition of variance.

Decomposition of variance

For random variables X and Y, the variance of Y is given by:

$$\mathrm{Var}(Y) = \mathrm{E}(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(\mathrm{E}(Y \mid X)).$$

It follows that:

$$\mathrm{Var}(Y) \geq \mathrm{E}(\mathrm{Var}(Y \mid X))$$

i.e. the unconditional variance is in general larger than the expected value of the conditional variance. Hence if X contains some useful information about Y, then conditioning on X makes the uncertainty about the value of Y smaller. Equality holds when $\mathrm{Var}(\mathrm{E}(Y \mid X)) = 0$, i.e. when $\mathrm{E}(Y \mid X)$ is no longer random, which is when X contains no information about Y, i.e. when X and Y are independent.

Example 5.10 In Example 5.1 we deduced that $Y \sim \mathrm{Pois}(\lambda\pi)$. Hence $\mathrm{E}(Y) = \lambda\pi$ and $\mathrm{Var}(Y) = \lambda\pi$. However, these results could alternatively be derived without having to determine $p_Y(y)$.

Since $Y \mid X = x \sim \mathrm{Bin}(x, \pi)$, then:

$$\mathrm{E}(Y \mid X = x) = X\pi \quad \text{and} \quad \mathrm{Var}(Y \mid X = x) = X\pi(1-\pi).$$

Given that $X \sim \mathrm{Pois}(\lambda)$, we now apply the law of iterated expectations which gives:

$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y \mid X = x)) = \mathrm{E}(X)\,\pi = \lambda\pi$$

and:

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{E}(\mathrm{Var}(Y \mid X = x)) + \mathrm{Var}(\mathrm{E}(Y \mid X = x)) \\ &= \mathrm{E}(X)\,\pi(1-\pi) + \mathrm{Var}(X\pi) \\ &= \lambda\pi(1-\pi) + \lambda\pi^2 \\ &= \lambda\pi.\end{aligned}$$

Example 5.11 Let X and Y be two random variables with joint probability density function:

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

(a) Find the conditional variance $\mathrm{Var}(Y \mid X)$.

(b) Hence evaluate $\mathrm{Var}(Y)$.

Solution

(a) We start by calculating $\mathrm{Var}(X = x)$. We have that:

$$\mathrm{Var}(Y \mid X = x) = \mathrm{E}(Y^2 \mid X = x) - (\mathrm{E}(Y \mid X = x))^2.$$

In Example 5.5 we found that $\mathrm{E}(Y \mid X = x) = 2x/3$ and derived $f_{Y|X}(y \mid x)$. We can find $\mathrm{E}(Y^2 \mid X = x)$ using this conditional density function:

$$\mathrm{E}(Y^2 \mid X = x) = \int_{-\infty}^{\infty} y^2 f_{Y|X}(y \mid x)\,\mathrm{d}y = \int_0^x \frac{2y^3}{x^2}\,\mathrm{d}y = \left[\frac{2y^4}{x^2}\right]_0^x = \frac{x^2}{2}.$$

Therefore:

$$\mathrm{Var}(Y \mid X = x) = \frac{x^2}{18} - \left(\frac{2x}{3}\right)^2 = \frac{x^2}{18}$$

and so:

$$\mathrm{Var}(Y \mid X) = \frac{X^2}{18}.$$

(b) We have:

$$\mathrm{Var}(X) = \mathrm{E}(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(\mathrm{E}(Y \mid X)).$$

Substituting previously calculated expressions yields:

$$\mathrm{Var}(Y) = \mathrm{E}\left(\frac{X^2}{18}\right) + \mathrm{Var}\left(\frac{2X}{3}\right) = \frac{1}{18} \times \frac{3}{5} + \left(\frac{2}{3}\right)^2 \times \frac{3}{80} = \frac{1}{20}.$$

5.5.3 Conditional moment generating functions

The conditional moment generating function is defined as a conditional expectation.

Conditional moment generating function

The **conditional moment generating function** of Y given $X = x$ is defined as:

$$M_{Y|X}(u \mid x) = \mathrm{E}(\mathrm{e}^{uY} \mid X = x) = \begin{cases} \sum\limits_y \mathrm{e}^{uy} p_{Y|X}(y \mid x) & \text{(discrete cases)} \\ \int_{-\infty}^{\infty} \mathrm{e}^{uy} f_{Y|X}(y \mid x)\,\mathrm{d}y & \text{(continuous case).} \end{cases}$$

Since this is a conditional expectation, it is a random variable.

Note that from $M_{Y|X}(u \mid x)$ we can derive the joint moment generating function of X and Y, and the marginal moment generating function of Y. We have:

$$M_{X,Y}(t, u) = \mathrm{E}(\mathrm{e}^{tX+uY}) = \mathrm{E}(\mathrm{e}^{tX} M_{Y|X}(u \mid X))$$

and:

$$M_Y(u) = M_{X,Y}(0, u) = \mathrm{E}(M_{Y|X}(u \mid X))$$

as a consequence of the law of iterated expectations.

Example 5.12 Consider again Example 5.1. Alternatively, we may use the moment generating functions. We have:

$$M_X(t) = \exp(\lambda(\mathrm{e}^t - 1)) \quad \text{and} \quad M_{Y|X}(u \,|\, X) = (1 - \pi + \pi\mathrm{e}^u)^X$$

hence:

$$\begin{aligned} M_Y(u) = \mathrm{E}(M_{Y|X}(u \,|\, X)) &= \mathrm{E}((1 - \pi + \pi\mathrm{e}^u)^X) \\ &= \mathrm{E}(\exp(\ln(1 - \pi + \pi\mathrm{e}^u)^X)) \\ &= \mathrm{E}(\exp(X \ln(1 - \pi + \pi\mathrm{e}^u))) \\ &= M_X(\ln(1 - \pi + \pi\mathrm{e}^u)) \\ &= \exp(\lambda(1 - \pi + \pi\mathrm{e}^u - 1)) \\ &= \exp(\lambda\pi(\mathrm{e}^u - 1)) \end{aligned}$$

which is the moment generating function of a Poisson distribution with parameter $\lambda\pi$.

Example 5.13 The number of eggs, N, laid by an insect follows a Poisson distribution with mean λ. For each egg, the probability that the egg hatches is π, and eggs hatch independently. Let H be the number of eggs which hatch.

(a) Write down an expression for the conditional moment generating function $M_{H|N}(u \,|\, N)$.

(b) Derive the moment generating function of H.

(c) What is the distribution of H?

Solution

(a) We know that $N \sim \text{Pois}(\lambda)$ and we model H as a sum of N independent random variables distributed as Bernoulli with probability of success (egg hatches) π. Hence $H \,|\, N = n \sim \text{Bin}(n, \pi)$. Since $H \,|\, N = n$ has a binomial distribution:

$$M_{H|N}(t \,|\, N = n) = (1 - \pi + \pi\mathrm{e}^t)^n$$

and so the conditional mgf is given by:

$$M_{H|N}(t \,|\, R) = (1 - \pi + \pi\mathrm{e}^t)^N.$$

(b) We now apply iterated expectations:

$$\begin{aligned} M_K(t) = \mathrm{E}(M_{H|N}(t \,|\, N)) &= \mathrm{E}((1 - \pi + \pi\mathrm{e}^t)^N) \\ &= \sum_{n=0}^{\infty} (1 - \pi + \pi\mathrm{e}^t)^r \, \frac{\mathrm{e}^{-\lambda}\lambda^n}{n!} \\ &= \mathrm{e}^{-\lambda} \exp(\lambda(1 - \pi + \pi\mathrm{e}^t)) \\ &= \exp(\lambda\pi(\mathrm{e}^t - 1)). \end{aligned}$$

(c) This is the mgf of a Poisson-distributed random variable with parameter $\lambda\pi$.

Activity 5.7 Let X and Y be random variables and we define the conditional variance of Y given X as:

$$\mathrm{Var}(Y \mid X) = \mathrm{E}((Y - \mathrm{E}(Y \mid X))^2 \mid X).$$

Show that:

$$\mathrm{Var}(Y \mid X) = \mathrm{E}(Y^2 \mid X) - \mathrm{E}((Y \mid X)^2).$$

Activity 5.8 Let X and Y be random variables with conditional moment generating function $M_{Y|X}(u \mid X)$. Show that:

(a) $M_Y(u) = \mathrm{E}(M_{Y|X}(u \mid X))$

(b) $M_{X,Y}(t, u) = \mathrm{E}(\mathrm{e}^{tX} M_{Y|X}(u \mid X))$

where M_Y is the marginal moment generating function of Y and $M_{X,Y}$ is the joint moment generating function of X and Y.

Activity 5.9 Let X and Y be random variables with the joint density function:

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

Determine $\mathrm{E}(X)$, $\mathrm{E}(X^2)$, $f_{Y|X}$ and $\mathrm{E}(Y \mid X)$. Hence calculate $\mathrm{E}(Y)$ and $\mathrm{Cov}(X, Y)$.

5.6 Hierarchies and mixtures

Suppose that we are interested in a random variable Y with a distribution which depends on another random variable, say X. This is called a *hierarchical model* and Y has a *mixture distribution*. In the first instance we do not know the marginal distribution of Y directly, but we know the conditional distribution of Y given $X = x$ and the marginal distribution of X.

The key results which are necessary for characterising hierarchies such as Y are:

1. $\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y \mid X))$
2. $\mathrm{Var}(Y) = \mathrm{E}(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(\mathrm{E}(Y \mid X))$
3. $f_Y(y) = \mathrm{E}(f_{Y|X}(y \mid X))$
4. $M_Y(u) = \mathrm{E}(M_{Y|X}(u \mid X))$.

Example 5.14 Here we consider Poisson mixing. Suppose $Y \mid \Lambda = \lambda \sim \text{Pois}(\lambda)$, for some positive random variable Λ. Hence:

$$\text{E}(Y \mid \Lambda) = \text{Var}(Y \mid \Lambda) = \Lambda.$$

Therefore:

$$\text{E}(Y) = \text{E}(\Lambda) \quad \text{and} \quad \text{Var}(Y) = \text{E}(\Lambda) + \text{Var}(\Lambda).$$

Example 5.15 Consider the following lottery game involving two urns.

- Urn A: five balls, numbered 1 to 5.
- Urn B: five balls, four green and one red.

Tickets cost £1 and players select two numbers between 1 and 5 (repetition not permitted). The lottery draw consists of two balls selected at random without replacement from Urn A, and one ball selected at random from Urn B. For a given ticket, there are three possible outcomes:

1. numbers on the ticket do not match the draw from Urn A $\Rightarrow$ win £0
2. numbers on the ticket match those drawn from Urn A and a green ball is drawn from Urn B $\Rightarrow$ win £1
3. numbers on the ticket match those drawn from Urn A and the red ball is drawn from Urn B $\Rightarrow$ win £$(z+1)$.

(a) Determine the probability of each of the three possible outcomes.

(b) Define R to be a random variable denoting the return from playing the game with a single ticket. Write down the probability mass function of R and evaluate the expected return $\text{E}(R)$. What value must z take for the expected return to be zero?

Now suppose that, rather than taking a fixed value z, the size of the top prize is determined by a random variable Z.

(c) Derive an expression for $M_{R|Z=z}(t \mid z)$, i.e. the moment generating function of the random variable $R \mid Z = z$.

(d) Write down an expression for the conditional moment generating function $M_{R|Z}(t \mid Z)$. Derive an expression for the moment generating function of R in terms of the moment generating function of Z.

(e) Provide an expression for the kth moment of R in terms of the kth moment of Z.

Solution

(a) The total number of combinations is $\binom{5}{2} = 10$, so defining M to be the event that the numbers match those on the ticket, G to be a green ball drawn from Urn B, and R to be the red ball drawn from Urn B, we have:

$$P(M^c) = \frac{9}{10}, \quad P(M \cap G) = \frac{1}{10} \times \frac{4}{5} = \frac{2}{25} \quad \text{and} \quad P(M \cap R) = \frac{1}{10} \times \frac{1}{5} = \frac{1}{50}.$$

(b) The return R can take three values: -1, 0 or z. The probability mass function is then:

$$p_R(r) = \begin{cases} 9/10 & \text{for } r = -1 \\ 2/25 & \text{for } r = 0 \\ 1/50 & \text{for } r = z \\ 0 & \text{otherwise.} \end{cases}$$

Hence:

$$\mathrm{E}(R) = \sum_r r\, p_R(r) = -1 \times \frac{9}{10} + 0 \times \frac{2}{25} + z \times \frac{1}{50} = \frac{z - 45}{50}.$$

Hence $z = 45$ will yield an expected return of zero.

(c) We have:

$$M_{R|Z=z}(t \,|\, z) = \mathrm{E}(\mathrm{e}^{tR} \,|\, Z = z) = \frac{9}{10}\,\mathrm{e}^{-t} + \frac{2}{25} + \frac{1}{50}\,\mathrm{e}^{tz}.$$

(d) We have:

$$M_{R|Z}(t \,|\, Z) = \frac{2}{25} + \frac{9}{10}\,\mathrm{e}^{-t} + \frac{1}{50}\,\mathrm{e}^{tZ}.$$

Hence:

$$M_R(t) = \mathrm{E}(M_{R|Z}(t \,|\, Z)) = \frac{2}{25} + \frac{9}{10}\,\mathrm{e}^{-t} + \frac{1}{50}\,M_Z(t).$$

(e) $\mu_{Z,k}$ is the kth moment of Z. We can use the fact that:

$$\mathrm{e}^{-t} = 1 - t + \frac{t^2}{2} - \cdots$$

and:

$$M_Z(t) = 1 + \mu_{Z,1} t + \mu_{Z,2}\,\frac{t^2}{2} + \cdots$$

and compare coefficients to show that the kth moment of R is:

$$\mu_{R,k} = \frac{1}{50}\,\mu_{Z,k} + (-1)^k\,\frac{9}{10}.$$

Activity 5.10 Show that if $Y \,|\, \Lambda = \lambda \sim \mathrm{Pois}(\lambda)$, then $M_Y(t) = M_\Lambda(\mathrm{e}^t - 1)$.

Activity 5.11 Find the moment generating function of Y as defined in Activity 5.5.

Activity 5.12 Suppose there are two investment opportunities from which to choose:

A: with a return of £X which has density function f_X

B: with a return of £Y which has density function f_Y.

Suppose we select investment A with probability π, and investment B with probability $1 - \pi$. Let Z denote the return on our investment. Formulate this framework as a mixture. Proceed to provide the density function, expected value and variance of Z in terms of the density functions, expected values and variances of X and Y.

5.7 Random sums

We consider the case in which $X_1, X_2, \ldots, X_N$ is a sequence of i.i.d. random variables and:

$$Y = \sum_{i=1}^{N} X_i$$

where N is also a random variable (instead of the usual fixed constant n) which is independent of each X_i.

The random variable Y is called a **random sum** and can be viewed as a mixture such that $Y \mid N = n$ is a sum of random variables. Therefore, previous results regarding sums of random variables are applicable to $Y \mid N = n$, to which we then apply our definitions of conditional expectation, conditional variance and conditional moment generating function.

Conditional results for random sums

Let $\{X_i\}$ be a sequence of i.i.d. random variables with mean $\mathrm{E}(X)$ and variance $\mathrm{Var}(X)$, for all i. Suppose that N is a random variable taking only positive integer values and define $Y = \sum_{i=1}^{N} X_i$, then:

1. $\mathrm{E}(Y \mid N) = N\,\mathrm{E}(X)$
2. $\mathrm{Var}(Y \mid N) = N\,\mathrm{Var}(X)$
3. $M_{Y|N}(t \mid N) = (M_X(t))^N$
4. $K_{Y|N}(t \mid N) = NK_X(t)$.

Marginal results for random sums

Let $\{X_i\}$ be a sequence of i.i.d. random variables with mean $\mathrm{E}(X)$ and variance $\mathrm{Var}(X)$, for all i. Suppose that N is a random variable taking only positive integer values and define $Y = \sum_{i=1}^{N} X_i$, then:

1. $\mathrm{E}(Y) = \mathrm{E}(N)\,\mathrm{E}(X)$
2. $\mathrm{Var}(Y) = \mathrm{E}(N)\,\mathrm{Var}(X) + \mathrm{Var}(N)(\mathrm{E}(X))^2$
3. $M_Y(t) = M_N(\log M_X(t))$
4. $K_Y(t) = K_N(K_X(t))$.

Proof:

1. We have:
$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y \mid N)) = \mathrm{E}(N\,\mathrm{E}(X)) = \mathrm{E}(N)\,\mathrm{E}(X).$$

2. We have:
$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{E}(\mathrm{Var}(Y \mid N)) + \mathrm{Var}(\mathrm{E}(Y \mid N)) \\ &= \mathrm{E}(N\,\mathrm{Var}(X)) + \mathrm{Var}(N\,\mathrm{E}(X)) \\ &= \mathrm{E}(N)\,\mathrm{Var}(X) + (\mathrm{E}(X))^2\mathrm{Var}(N).\end{aligned}$$

3. We have:
$$\begin{aligned}M_Y(t) = \mathrm{E}(M_{Y|N}(t \mid N)) &= \mathrm{E}((M_X(t))^N) \\ &= \mathrm{E}(\exp(N\,\log M_X(t))) \\ &= M_N(\log M_X(t)).\end{aligned}$$

4. We have:
$$K_Y(t) = \log M_Y(t) = \log(M_N(\log M_X(t))) = \log(M_N(K_X(t))) = K_N(K_X(t)).$$

■

Example 5.16 Each year the value of claims made by an owner of a health insurance policy is distributed exponentially with mean μ, independent of previous years. At the end of each year with probability π the individual will cancel their policy. We want the distribution of the *total cost* of the health insurance policy for the insurer. The value of claims in year i is X_i, and the number of years in which the policy is held is N. Therefore:

$$X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}\left(\frac{1}{\mu}\right) \quad \text{and} \quad N \sim \mathrm{Geometric}(\pi).$$

The total cost for the insurer is $Y = \sum_{i=1}^{N} X_i$. Hence:

$$\mathrm{E}(Y) = \mathrm{E}(N)\,\mathrm{E}(X) = \frac{1}{\pi} \times \frac{1}{1/\mu} = \frac{\mu}{\pi}.$$

In order to obtain the distribution we use the cumulant generating functions:

$$K_X(t) = -\log(1-\mu t) \quad \text{and} \quad K_N(t) = -\log\left(1 - \frac{1}{\pi} + \frac{\mathrm{e}^{-t}}{\pi}\right)$$

and so:

$$K_Y(t) = K_N(K_X(t)) = -\log\left(1 - \frac{1}{\pi} + \frac{1-\mu t}{\pi}\right) = -\log\left(1 - \frac{\mu}{\pi}\,t\right).$$

By uniqueness of cumulant generating functions, we have that $Y \sim \mathrm{Exp}(\pi/\mu)$.

Example 5.17 Let $X_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi)$, and $N \sim \text{Pois}(\lambda)$. Consider $Y = \sum_{i=1}^{N} X_i$, hence $Y \mid N = n \sim \text{Bin}(n, \pi)$ and:

1. $\mathrm{E}(Y) = \lambda\,\mathrm{E}(X)$
2. $\mathrm{Var}(Y) = \lambda\,\mathrm{E}(X^2)$
3. $M_Y(t) = M_N(\log M_X(t)) = \mathrm{e}^{\lambda(M_X(t)-1)}$
4. $K_S(t) = \lambda(M_X(t) - 1)$.

By using the mgf of the Bernoulli distribution, $M_X(t) = 1 - \pi + \pi\mathrm{e}^t$, we get:

$$M_Y(t) = \mathrm{e}^{\lambda(M_X(t)-1)} = \mathrm{e}^{\lambda\pi(\mathrm{e}^t-1)}.$$

By uniqueness of moment generating functions, we have that $Y \sim \text{Pois}(\lambda\pi)$.

Example 5.18 The box office of a venue has several service windows. Queues form at each of these windows before the box office opens. As each person arrives they join queue A with probability π independently of the choices made by others. Starting at 8am ($\lambda = 0$) with no-one in any queue, the number who have arrived by time λ (measured in hours) is Poisson with parameter λ. At 10am ($\lambda = 2$) the box office opens and no-one else is allowed to join queue A.

(a) Show that for $0 < \lambda < 2$ the number of people in queue A is Poisson with parameter $\lambda\pi$.

(b) When the box office opens, the length of time W_i taken to serve customer i is exponentially distributed with parameter θ. If T is the length of time until queue A is empty, show that:

$$\mathrm{E}(T) = \frac{2\pi}{\theta} \quad \text{and} \quad \mathrm{Var}(T) = \frac{4\pi}{\theta^2}.$$

Hint: Instead of using the mgf $M_T(t)$ you might prefer to use the cumulant generating function $K_T(t) = \log M_T(t)$.

Solution

(a) Let X_i be a random variable which takes the value 1 if the ith person joins queue A, and 0 otherwise. Let N be the total number of people who have arrived at time λ and let S be the number of people in queue A. Hence $X_i \sim \text{Bernoulli}(\pi)$, $N \sim \text{Pois}(\lambda)$ and $S = \sum_{i=1}^{N} X_i$ so we have the random sum set up. We then use standard results. We have:

$$\begin{aligned} M_S(t) = M_N(\log M_X(t)) &= \exp(\lambda(\mathrm{e}^{\log M_X(t)} - 1)) \\ &= \exp(\lambda(M_X(t) - 1)) \\ &= \exp(\lambda(\pi \mathrm{e}^t - \pi)) \\ &= \exp(\lambda\pi(\mathrm{e}^t - 1)). \end{aligned}$$

This is the moment generating function of a Poisson distribution, hence $S \sim \text{Pois}(\lambda\pi)$.

(b) If the number of people in queue A at time $\lambda = 2$ is N then by part (a) $N \sim \text{Pois}(2\pi)$. If individual i takes time W_i to be served then the length of time for the queue to empty is $T = \sum_{i=1}^{N} W_i$, so we are back to the standard random sum set up. We are told that $W_i \sim \text{Exp}(\theta)$, so $M_W(t) = \theta/(\theta - t)$. Hence:

$$M_T(t) = \exp\left(2\pi\left(\frac{\theta}{\theta - t} - 1\right)\right) = \exp\left(\frac{2\pi t}{\theta - t}\right).$$

The cumulant generating function is then:

$$\begin{aligned} K_T(t) = \log(M_T(t)) &= \frac{2\pi t}{\theta - t} \\ &= \frac{2\pi t}{\theta}\left(1 - \frac{t}{\theta}\right)^{-1} \\ &= \frac{2\pi t}{\theta}\left(1 + \frac{t}{\theta} + \frac{t^2}{\theta^2} + \cdots\right) \\ &= \frac{2\pi}{\theta}\, t + \frac{4\pi}{\theta^2}\,\frac{t^2}{2} + \cdots. \end{aligned}$$

Note that $\text{E}(T)$ is the coefficient of t in the cumulant generating function and $\text{Var}(T)$ is the coefficient of $t^2/2$, hence:

$$\text{E}(T) = \frac{2\pi}{\theta} \quad \text{and} \quad \text{Var}(T) = \frac{4\pi}{\theta^2}$$

as required.

Example 5.19 Let:

$$Y = \begin{cases} \sum_{i=1}^{N} X_i & \text{if } N \geq 1 \\ 0 & \text{if } N = 0 \end{cases}$$

where $\{X_i\}$, for $i = 1, 2, \ldots, N$, is a sequence of i.i.d. random variables and N is a discrete random variable taking values $0, 1, 2, \ldots$ independent of $\{X_i\}$. Let:

$$K_Y(t) = \log(M_Y(t)), \quad K_N(t) = \log(M_N(t)) \quad \text{and} \quad K_X(t) = \log(M_X(t))$$

denote the cumulant generating functions of Y, N and X_i, respectively. Prove that:

$$K_Y(t) = K_N(K_X(t)).$$

Now, suppose that $P(N = i) = \pi^{i-1}(1 - \pi)$ for $i = 1, 2, \ldots$, and the density function of X_i is:

$$f_X(x) = \begin{cases} \alpha^2 x \mathrm{e}^{-\alpha x} & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Prove that Y has the density function:

$$f_Y(y) = \begin{cases} \dfrac{\alpha(1-\pi)}{2\sqrt{\pi}} \left(\mathrm{e}^{-\alpha(1-\sqrt{\pi})y} - \mathrm{e}^{-\alpha(1+\sqrt{\pi})y} \right) & \text{for } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Hint: Use the result of the mgf above and use the mgf of the sum of two exponentially distributed random variables.

Solution

We will interpret $\sum_{i=1}^{N} X_i$ as 0 for $N = 0$, and $\prod_{i=1}^{N}$ operations as giving 1 when $N = 0$. Hence:

$$\begin{aligned} M_Y(t) = \mathrm{E}(\mathrm{e}^{tY}) = \mathrm{E}(\mathrm{E}(\mathrm{e}^{tY} \mid N)) &= \mathrm{E}(\mathrm{E}(\mathrm{e}^{t\sum_{i=1}^{N} X_i} \mid N)) \\ &= \mathrm{E}\left(\prod_{i=1}^{N} \mathrm{E}(\mathrm{e}^{tX_i}) \right) \\ &= \mathrm{E}\left(\prod_{i=1}^{n} M_X(t) \right) \\ &= \mathrm{E}((M_X(t))^N) \\ &= \mathrm{E}(\mathrm{e}^{N \ln M_X(t)}) \\ &= \mathrm{E}(\mathrm{e}^{N K_X(t)}) \\ &= M_N(K_X(t)). \end{aligned}$$

Taking logs on both sides gives the result, though the form above may often be more useful.

The example has a geometric distribution for N. Therefore:

$$M_N(t) = \sum_{n=1}^{\infty} \mathrm{e}^{tn}\pi^{n-1}(1-\pi) = \mathrm{e}^t(1-\pi)\sum_{n=1}^{\infty}(\pi\mathrm{e}^t)^{n-1} = \frac{\mathrm{e}^t(1-\pi)}{1-\pi\mathrm{e}^t}.$$

X_i has a Gamma$(\alpha, 2)$ distribution with mgf:

$$M_X(t) = \frac{\alpha^2}{(\alpha-t)^2}.$$

Using the result we obtained above, we have:

$$\begin{aligned}
M_Y(t) = M_N(K_X(t)) &= \frac{\mathrm{e}^{K_X(t)}(1-\pi)}{1-\pi\mathrm{e}^{K_X(t)}} \\
&= \frac{M_X(t)(1-\pi)}{1-\pi M_X(t)} \\
&= \frac{(1-\pi)\alpha^2/(\alpha-t)^2}{1-\pi\alpha^2/(\alpha-t)^2} \\
&= \frac{(1-\pi)\alpha^2}{(\alpha-t)^2-\pi\alpha^2} \\
&= \frac{(1-\pi)\alpha^2}{((\alpha-t)-\sqrt{\pi}\alpha)((\alpha-t)+\sqrt{\pi}\alpha)} \\
&= \frac{(1-\pi)\alpha^2}{(\alpha(1-\sqrt{\pi})-t)(\alpha(1+\sqrt{\pi})-t)} \\
&= \frac{(1-\sqrt{\pi})(1+\sqrt{\pi})\alpha}{2\sqrt{\pi}}\,\frac{1}{\alpha(1-\sqrt{\pi})-t} \\
&\quad - \frac{(1-\sqrt{\pi})(1+\sqrt{\pi})\alpha}{2\sqrt{\pi}}\,\frac{1}{\alpha(1+\sqrt{\pi})-t} \\
&= \frac{1+\sqrt{\pi}}{2\sqrt{\pi}}\,\frac{\alpha(1-\sqrt{\pi})}{\alpha(1-\sqrt{\pi})-t} - \frac{1-\sqrt{\pi}}{2\sqrt{\pi}}\,\frac{\alpha(1+\sqrt{\pi})}{\alpha(1+\sqrt{\pi})-t}.
\end{aligned}$$

We can identify the mgf as a mixture of two exponential distributions with scale parameters $\alpha(1-\sqrt{\pi})$ and $\alpha(1+\sqrt{\pi})$ and mixing weights $(1+\sqrt{\pi})/(2\sqrt{\pi})$ and $-(1-\sqrt{\pi})/(2\sqrt{\pi})$, respectively. Note that one of the weights is negative. The corresponding density function is:

$$\begin{aligned}
f_Y(y) &= \frac{1+\sqrt{\pi}}{2\sqrt{\pi}}\,\alpha(1-\sqrt{\pi})\mathrm{e}^{-\alpha(1-\sqrt{\pi})y} - \frac{1-\sqrt{\pi}}{2\sqrt{\pi}}\,\alpha(1+\sqrt{\pi})\mathrm{e}^{-\alpha(1+\sqrt{\pi})y} \\
&= \frac{\alpha(1-\pi)}{2\sqrt{\pi}}\left(\mathrm{e}^{-\alpha(1-\sqrt{\pi})y} - \mathrm{e}^{-\alpha(1+\sqrt{\pi})y}\right)
\end{aligned}$$

for $y > 0$, and 0 otherwise.

Activity 5.13 Show that Example 5.1 can be formulated as a random sum.

Activity 5.14 Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables. Define the random sum $S = \sum_{i=1}^{N} X_i$, where $N \sim \text{Pois}(\lambda)$. Show that:

(a) $\text{E}(S) = \lambda\,\text{E}(X)$

(b) $\text{Var}(S) = \lambda\,\text{E}(X^2)$

(c) the kth cumulant of S is $\lambda\,\text{E}(X^k)$.

5.8 Conditioning for random vectors

Consider the two random vectors $\mathbf{X} = (X_1, X_2, \ldots, X_m)^T$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$. The joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ is the joint distribution of all variables in $\mathbf{X}$ and all variables in $\mathbf{Y}$. Hence:

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = f_{X_1,X_2,\ldots,X_m,Y_1,Y_2,\ldots,Y_n}(x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n).$$

We can now define conditional distributions accordingly.

Conditional distributions of random vectors

Let $\mathbf{X}$ and $\mathbf{Y}$ be random vectors. If $f_{\mathbf{X}}(\mathbf{x}) > 0$, then we define the conditional probability mass function as:

$$p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = \frac{p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})}{p_{\mathbf{X}}(\mathbf{x})}$$

and the conditional probability density function as:

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = \frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})}.$$

Alternatively:

$$p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x})\, p_{\mathbf{X}}(\mathbf{x})$$

and:

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x})\, f_{\mathbf{X}}(\mathbf{x}).$$

We may decompose a joint distribution of random vectors into a product of conditional distributions.

Decomposition of probability density function

Given an n-dimensional random vector $\mathbf{X}$ and given $\mathbf{x} \in \mathbb{R}^n$, then:

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= f_{X_n|X_{n-1},\ldots,X_1}(x_n \,|\, x_{n-1},\ldots,x_1)\, f_{X_{n-1}|X_{n-2},\ldots,X_1}(x_{n-1} \,|\, x_{n-2},\ldots,x_1) \\ &\quad \cdots f_{X_2|X_1}(x_2 \,|\, x_1)\, f_{X_1}(x_1) \\ &= \prod_{i=1}^{n} f_{X_j|\mathbf{X}_{j-1}}(x_j \,|\, \mathbf{x}_{j-1}) \end{aligned}$$

where the random vector $\mathbf{X}_{j-1}$ is the random vector $\mathbf{X}$ without its jth element.

Example 5.20 For the three continuous random variables X_1, X_2 and X_3, we can group them in different ways. For example:

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = f_{X_3|X_1,X_2}(x_3 \,|\, x_1,x_2)\, f_{X_1,X_2}(x_1,x_2).$$

Applying again the definition above to the joint pdf of X_1 and X_2 we have:

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = f_{X_3|X_1,X_2}(x_3 \,|\, x_1,x_2)\, f_{X_2|X_1}(x_2 \,|\, x_1)\, f_{X_1}(x_1).$$

Example 5.21 Let $X_1, X_2, \ldots, X_n$ be random variables and define the $n \times 1$ random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$. If $X_1, X_2, \ldots, X_n$ are jointly normal then $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the mean $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$ is an $n \times 1$ vector and the variance–covariance matrix $\boldsymbol{\Sigma} = \mathrm{Var}(\mathbf{X})$ is an $n \times n$ matrix whose (i,j)th entry is $\mathrm{Cov}(X_i, X_j)$. The joint density function is:

$$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-n/2} |\det \boldsymbol{\Sigma}|^{-1/2} \mathrm{e}^{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}.$$

Example 5.22 Suppose that $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)^T$, for some positive integers n and m, and $X \sim N(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}})$ and $Y \sim N(\boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\Sigma}_{\mathbf{Y}})$. If we also have $\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \boldsymbol{\Sigma}_{\mathbf{XY}} = \boldsymbol{\Sigma}_{\mathbf{YX}}^T$, then:

$$\mathrm{E}(\mathbf{Y} \,|\, \mathbf{X}) = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Sigma}_{\mathbf{YX}} \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})$$

and:

$$\mathrm{Var}(\mathbf{Y} \,|\, \mathbf{X}) = \boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{YX}} \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \boldsymbol{\Sigma}_{\mathbf{XY}}.$$

5.9 A reminder of your learning outcomes

On completion of this chapter, you should be able to:

- derive conditional mass and density functions from joint mass and density functions

- state the relationships between joint, marginal and conditional distributions, and use these relationships to solve problems
- calculate conditional expectations of a function of a random variable
- apply the law of iterated expectations
- prove that variance can be decomposed as the sum of the expected value of the conditional variance and the variance of the conditional expectation
- derive and use conditional moment generating functions and conditional cumulant generating functions
- solve problems involving hierarchies, mixtures and random sums.

5.10 Sample examination questions

Solutions can be found in Appendix C.

1. Let the joint density function for X and Y be:

$$f_{X,Y}(x,y) = \begin{cases} 8xy & \text{for } 0 < y < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

   (a) Find $\mathrm{E}(Y \mid X)$ and hence evaluate $\mathrm{E}(Y)$ and $\mathrm{E}(XY)$ using the law of iterated expectations.

   (b) Find $\mathrm{Var}(Y \mid X = 0.5)$.

2. Let X denote the score when a fair die is rolled. A fair coin is then tossed the number of times indicated by the die, and let Y denote the number of heads.

   (a) What is the distribution of $Y \mid X = x$? Specify the parameters of this distribution. Use the law of iterated expectations to find $\mathrm{E}(Y)$.

   (b) Write down an expression for $P(Y = 5 \mid X = x)$. Use the total probability formula to find $P(Y = 5)$.

3. Suppose that $Y \mid X = x \sim N(0, x)$, where $X \sim \mathrm{Exp}(\lambda)$. Find the mean, the variance and the cumulant generating function of Y.

Appendix A
Non-examinable proofs

A.1 Chapter 2 – Probability space

Proofs of basic probability properties

Lemma: Consider a probability space $(\Omega, \mathcal{F}, P)$, with $A \in \mathcal{F}$ and $B \in \mathcal{F}$, i.e. A and B are events.

i. $P(A^c) = 1 - P(A)$.

Proof: By definition of A^c we have $A \cap A^c = \emptyset$ and $A \cup A^c = \Omega$. Hence A and A^c are mutually exclusive events whose union is the sample space. Hence:

$$P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$$

by property iii. of measure. We have that $P(\Omega) = 1$ by the additional property of probability measure. Combining, we have that:

$$P(A) + P(A^c) = 1$$

and the result follows immediately. ■

ii. If $A \subseteq B$, then $P(B \setminus A) = P(B) - P(A)$.

Proof: Instead of partitioning the sample space, we partition the set B. Since $A \subseteq B$, we can write $B = A \cup (B \setminus A)$. By definition, $A \cap (B \setminus A) = \emptyset$. Therefore:

$$P(B) = P(A) + P(B \setminus A)$$

by property iii. of measure and the result follows immediately. ■

iii. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof: We have established $P(B \setminus A) = P(B) - P(A)$. $A \cup B$ can be decomposed into the mutually exclusive sets A and $B \setminus A$, such that:

$$A \cup B = A \cup (B \setminus A).$$

Hence by the properties of measures:

$$P(A \cup B) = P(A) + P(B \setminus A).$$

By definition, $B \setminus A = B \setminus (A \cap B)$ and, since $A \cap B \subseteq B$ then we can use part ii. above to give:

$$P(B \setminus A) = P(B \setminus (A \cap B)) = P(B) - P(A \cap B).$$

■

Corollary: If $A \subseteq B$, then $P(A) \leq P(B)$.

Proof: This is an immediate consequence of part i. above. If $A \subseteq B$, then $P(A) = P(B) - P(B \setminus A)$. However, by the properties of measures, $P(B \setminus A) \geq 0$ and hence $P(A) \leq P(B)$. ■

Proof of Boole inequality

Consider a probability space $(\Omega, \mathcal{F}, P)$ with $A_1, A_2, \ldots \in \mathcal{F}$, then:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i).$$

Intuitively, when we add up the probabilities of overlapping events, the overlaps get included repeatedly so the resulting probability is larger than the probability of the union.

Proof: We partition the event of interest, $\bigcup_{i=1}^{\infty} A_i$, recursively by removing from each event the part which overlaps with events we have already considered. This yields:

$$B_1 = A_1, \quad B_2 = A_2 \setminus B_1, \quad B_3 = A_3 \setminus (B_1 \cup B_2), \quad \ldots, \quad B_i = A_i \setminus \bigcup_{j=1}^{i-1} B_j, \ldots.$$

$B_1, B_2, \ldots \in \mathcal{F}$ are mutually exclusive with:

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i.$$

The proof is completed by noting that $B_i \subseteq A_i$, hence $P(B_i) \leq P(A_i)$ for all i and hence:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i) \leq \sum_{i=1}^{\infty} P(A_i).$$

■

Proofs of probability limits for sequences of sets

i. If $\{A_i : i = 1, 2, \ldots\}$ is an increasing sequence of sets $A_1 \subseteq A_2 \subseteq \cdots$, then:

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcup_{i=1}^{\infty} A_i\right).$$

Proof: Define $B_1 = A_1$ and $B_i = A_i \setminus A_{i-1}$ for $i \geq 2$. This ensures that the B_is are mutually exclusive and that $A_n = \bigcup_{i=1}^{n} B_i$ for any n. Therefore, $\{B_1, B_2, \ldots, B_n\}$ forms a partition of A_n and:

$$P(A_n) = \sum_{i=1}^{n} P(B_i).$$

It is also clear from the definition of the B_is that:

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i.$$

Putting everything together results in:

$$\lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i) = \sum_{i=1}^{\infty} P(B_i) = P\left(\bigcup_{i=1}^{\infty} B_i\right) = P\left(\bigcup_{i=1}^{\infty} A_i\right).$$

Note $P\left(\bigcup_{i=1}^{\infty} A_i\right)$ is well-defined since, by definition of a σ-algebra, $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$. ■

ii. If $\{A_i : i = 1, 2, \ldots\}$ is a decreasing sequence of sets $A_1 \supseteq A_2 \supseteq \cdots$, then:

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcap_{i=1}^{\infty} A_i\right).$$

Proof: If $A_1 \supseteq A_2 \supseteq \cdots$ then $A_1^c \subseteq A_2^c \subseteq \cdots$. Applying part i. above, we have:

$$\lim_{n\to\infty} P(A_n^c) = P\left(\bigcup_{i=1}^{\infty} A_i^c\right) = P\left(\left(\bigcap_{i=1}^{\infty} A_i\right)^c\right) = 1 - P\left(\bigcap_{i=1}^{\infty} A_i\right).$$

The result follows by noting that $P(A_n^c) = 1 - P(A_n)$. ■

Counting cards

1. Introduction

This gives some additional information related to the examples of playing cards used in Example 2.11 to illustrate counting rules. First, it explains the terminology of playing cards, in case you are unfamiliar with it. Second, it shows how to calculate the numbers of all the standard types of 5-card hands. These provide further practice in counting rules.

2. Playing cards: Basic terminology

The standard deck of playing cards contains 52 cards. Each card is labelled in two ways, with one of 13 ranks (numbered 1, 2, ..., 13) and one of 4 suits (hearts ♡, diamonds ♢, spades ♠ and clubs ♣). Each combination of rank and suit appears once and only once, so there are $4 \times 13 = 52$ distinct cards.

We will consider examples where one or more cards are drawn from the deck. We assume that the deck is thoroughly shuffled, so that the order in which cards are drawn can be treated as random and all orderings are equally probable.

In particular, we consider the case of drawing 5 cards from the deck. The selection is without replacement, and the order in which the cards are drawn is ignored, so the selections are treated as unordered sets. In several card games these sets or 'hands' of 5 distinct cards are classified as follows, in descending order of value:

- Straight flush: 5 cards of the same suit and with successive ranks (for example 1–5, all clubs; note that 10–11–12–13–1 also count as successive).
- Four-of-a-kind: any hand with 4 cards of the same rank (for example 8–8–8–8–2).
- Full house: 3 cards of one rank and 2 cards of another rank (for example 1–1–1–3–3).
- Flush: 5 cards of the same suit (but not of successive ranks).
- Straight: 5 cards with successive ranks (but not of the same suit).
- Three-of-a-kind: 3 cards of the same rank and 2 cards of different ranks (for example 2–2–2–4–6).
- Two pairs: 2 cards of one rank, 2 cards of another rank, and 1 card of a third rank (for example 3–3–4–4–10).
- One pair: 2 cards of the same rank, and 3 cards of 3 different ranks (for example 5–5–6–7–9).
- High card: none of the above, i.e. 5 cards of different, non-successive ranks and not of the same suit.

3. Total numbers of different hands

In total, there are:

$$\binom{52}{5} = \frac{52!}{5! \times 47!} = \frac{52 \times 51 \times 50 \times 49 \times 48}{5 \times 4 \times 3 \times 2 \times 1} = 2{,}598{,}960$$

different hands, i.e. different unordered subsets of 5 different cards selected without replacement from 52 cards.

The numbers of different types of hands are calculated as follows.

Straight flush: The straight can start at ranks 1–10, and it can be of any of 4 suits. Therefore, $10 \times 4 = 40$.

Four-of-a-kind: The rank of the four can be chosen in 13 ways, and the fifth card can be any of the remaining 48. Therefore, $13 \times 48 = 624$.

Full house: The rank of the three can be chosen in 13 ways. There are four cards of this rank so the three can be chosen in $\binom{4}{3} = 4$ ways. The rank of the two can be any of the remaining 12 ranks, and the two cards of this rank can be chosen in $\binom{4}{2}$ ways. Therefore:

$$13 \times \binom{4}{3} \times 12 \times \binom{4}{2} = 3{,}744.$$

Flush: The suit can be chosen in 4 ways. There are $\binom{13}{5}$ ways of choosing five cards from the thirteen of the same suit. Of these, we subtract the 10 that are successive (because those would give a straight flush). Therefore, $4 \times \left[\binom{13}{5} - 10\right] = 5{,}108$.

Straight: The straight can start at 10 values. The suits of the five successive cards can be chosen in 4^5 ways, minus the 4 which consist of one suit only (the straight flushes). Therefore, $10 \times (4^5 - 4) = 10{,}200$.

Three-of-a-kind: The rank of the three can be chosen in 13 ways. The three cards out of the four with this rank can be chosen in $\binom{4}{3} = 4$ ways. The two different ranks of the remaining cards can be chosen in $\binom{12}{2}$ ways, and the suit of each of these cards in 4 ways. Therefore:

$$13 \times \binom{4}{3} \times \binom{12}{2} \times 4 \times 4 = 54{,}912.$$

Two pairs: The ranks of the two pairs can be chosen in $\binom{13}{2}$ ways. For each pair, the suits can be chosen in $\binom{4}{2} = 6$ ways. The remaining card can be any of the 44 cards not of either of these ranks. Therefore, $\binom{13}{2} \times 6 \times 6 \times 44 = 123{,}552$.

One pair: The rank of the pair can be chosen in 13 ways, and the suits of the cards in the pair in $\binom{4}{2} = 6$ ways. The ranks of the remaining three cards must all be different, so they can be chosen in $\binom{12}{3}$ ways. Each of these 3 cards can be of any of the 4 suits. Therefore, $13 \times 6 \times \binom{12}{3} \times 4^3 = 1{,}098{,}240$.

High card: These should be all the rest, i.e. the total number of possible hands (2,598,960) minus all of the above. That is, 1,302,540.

The number of high-card hands can also be calculated directly as follows. There are $\binom{13}{5} = 1{,}287$ ways of choosing five different ranks for the five cards. From these, we need to subtract the 10 choices that are successive and would hence constitute a straight. The suits for the five cards can be chosen in $4^5 = 1{,}024$ ways, from which we subtract the 4 which are all of the same suit, i.e. flushes. Multiplying these together gives:

$$\left[\binom{13}{5} - 10\right] \times \left[4^5 - 4\right] = 1{,}302{,}540.$$

In summary, the numbers of all types of 5-card hands, and their probabilities, are as follows:

Hand	Number	Probability
Straight flush	40	0.000015
Four-of-a-kind	624	0.00024
Full house	3,744	0.00144
Flush	5,108	0.0020
Straight	10,200	0.0039
Three-of-a-kind	54,912	0.0211
Two pairs	123,552	0.0475
One pair	1,098,240	0.4226
High card	1,302,540	0.5012
Total	2,598,960	1.0

Proof of multinomial coefficients

Consider n objects where n_1 are of type 1, n_2 are of type 2, ..., n_r are of type r. The number of ways of arranging these n objects is:

$$(n_1, n_2, \ldots, n_r)! = \frac{n!}{n_1!\, n_2! \cdots n_r!}$$

where $\sum_{i=1}^{r} n_i = n$.

Proof: Suppose that we know the number of ways of arranging the objects, that is, we know $(n_1, n_2, \ldots, n_r)!$. If we treated all the objects of type 1 as being distinct this would increase the number of arrangements by a factor of $n_1!$. In general, if we treat the objects of type j as being distinct, this increases the number of arrangements by a factor of $n_j!$. If we treat all of the objects as being distinct, we know the number of arrangements is ${}^{n}P_n = n!$. Therefore, by the rule of product:

$$n_1!\, n_2! \cdots n_r!\, (n_1, n_2, \ldots, n_r)! = n!.$$

The result follows by rearrangement. ■

A.2 Chapter 3 – Random variables and univariate distributions

Properties of distribution functions

i. F_X is a non-decreasing function, i.e. if $x < y$ then $F_X(x) \leq F_X(y)$.

Proof: If $x < y$, then $A_x \subseteq A_y$. Hence $P(A_x) \leq P(A_y)$ and so $F_X(x) \leq F_X(y)$. ■

ii. $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.

Proof: Let $\{x_n : n = 1, 2, \ldots\}$ be a sequence of real numbers such that $x_n \to -\infty$ as $n \to \infty$. Without loss of generality, assume that $\{x_n\}$ is a decreasing sequence, i.e. $x_n \geq x_{n+1}$ for $n = 1, 2, \ldots$. Hence $A_{x_n} \supseteq A_{x_{n+1}}$ and $\bigcap_{n=1}^{\infty} A_{x_n} = \emptyset$. Therefore:

$$\lim_{n\to\infty} F_X(x_n) = \lim_{n\to\infty} P(A_{x_n}) = P\left(\bigcap_{n=1}^{\infty} A_{x_n}\right) = P(\emptyset) = 0 \tag{1}$$

making use of the probability limit of a sequence of sets. Since (1) holds for any sequence $\{x_n\}$ such that $x_n \to -\infty$, we conclude that:

$$\lim_{x\to-\infty} F_X(x) = 0.$$

Now let $\{x_n : n = 1, 2, \ldots\}$ be a sequence of real numbers such that $x_n \to \infty$ as $n \to \infty$. Without loss of generality, assume that $\{x_n\}$ is an increasing sequence, i.e. $x_n \leq x_{n+1}$ for $n = 1, 2, \ldots$. Hence $A_{x_n} \subseteq A_{x_{n+1}}$ and $\bigcup_{n=1}^{\infty} A_{x_n} = \Omega$. Therefore:

$$\lim_{n\to\infty} F_X(x_n) = \lim_{n\to\infty} P(A_{x_n}) = P\left(\bigcup_{n=1}^{\infty} A_{x_n}\right) = P(\Omega) = 1 \tag{2}$$

again making use of the probability limit of a sequence of sets. Since (2) holds for any sequence $\{x_n\}$ such that $x_n \to \infty$, we conclude that:

$$\lim_{x\to\infty} F_X(x) = 1.$$

■

iii. F_X is right continuous, i.e. $F_X(x+) = F_X(x)$ for all $x \in \mathbb{R}$.

Proof: Suppose we have a decreasing sequence such that $x_n \downarrow x$ as $n \to \infty$. By definition, $A_x \subseteq A_{x_n}$ for all n and A_x is the largest interval for which this is true. Hence $\bigcap_{n=1}^{\infty} A_{x_n} = A_x$ and:

$$\lim_{n\to\infty} F_X(x_n) = \lim_{n\to\infty} P(A_{x_n}) = P\left(\bigcup_{n=1}^{\infty} A_{x_n}\right) = P(A_x) = F_X(x). \tag{3}$$

Since (3) holds for any sequence $\{x_n\}$ such that $x_n \downarrow x$, we conclude that $\lim_{h\downarrow 0} F_X(x+h) = x$ for all x and so F_X is right continuous. ■

Probabilities from distribution functions

For real numbers x and y, with $x < y$, we have the following.

i. $P(X > x) = 1 - F_X(x)$.

Proof: Consider the probability space $(\Omega, \mathcal{F}, P)$. We know that if $A_x \in \mathcal{F}$, then $A_x^c \in \mathcal{F}$. By definition, $A_x^c = \{\omega \in \Omega : X(\omega) > x\}$, so we can refer to $P(X > x)$. We also know that $P(A_x^c) = 1 - P(A_x)$. Hence:

$$P(X > x) = P(A_x^c) = 1 - P(A_x) = 1 - P(X \leq x) = 1 - F_X(x).$$

■

ii. $P(x < X \leq y) = F_X(y) - F_X(x)$.

Proof: Consider the probability space $(\Omega, \mathcal{F}, P)$. We can show that $\{\omega \in \Omega : x < X(\omega) \leq y\} \in \mathcal{F}$ by noting that $\{\omega \in \Omega : x < X(\omega) \leq y\} = A_y \cap A_x^c$ and that $\mathcal{F}$ is closed under intersection. We can partition $(-\infty, y]$ into the disjoint intervals $(-\infty, x]$ and $(x, y]$. As these intervals are disjoint, then:

$$P(X \leq y) = P(X \leq x) + P(x < X \leq y)$$

hence $P(x < X \leq y) = F_X(y) - F_X(x)$. ■

iii. $P(X < x) - \lim_{h\downarrow 0} F_X(x-h) = F_X(x-)$.

Proof: Let $\{x_n\}$ be an increasing sequence, such that $x_n \uparrow x$. We have that the event:

$$\{\omega \in \Omega : X(\omega) < x\} = \bigcup_{n=1}^{\infty} A_{x_n}$$

and hence:

$$\{\omega \in \Omega : X(\omega) < x\} \in \mathcal{F}.$$

Therefore:

$$P(X < x) = P\left(\bigcup_{n=1}^{\infty} A_{x_n}\right) = \lim_{n\to\infty} F_X(x_n) = \lim_{h\downarrow 0} F_X(x-h) = F_X(x-).$$

■

iv. $P(X = x) = F_X(x) - F_X(x-)$.

Proof: We can partition $(-\infty, x]$ into the disjoint intervals $(-\infty, x)$ and $[x, x]$, i.e. this 'degenerate' interval $[x, x]$ is just the point $\{x\}$. Hence:

$$P(X \leq x) = P(X < x) + P(X = x) \quad \Rightarrow \quad P(X = x) = F_X(x) - F_X(x-).$$

■

A.3 Chapter 4 – Multivariate distributions

Joint cumulant generating function and joint cumulants

Proof: First we establish that $\kappa_{2,0} = \mathrm{Var}(X)$. By setting $u = 0$ in $K_{X,Y}(t, u)$ we get:

$$K_{X,Y}(t, 0) = \log M_{X,Y}(t, 0) = \log M_X(t) = K_X(t).$$

Hence:

$$\frac{\partial^r}{\partial t^r} K_{X,Y}(t, u)\Big|_{t=0,u=0} = \frac{\partial^r}{\partial t^r} K_{X,Y}(t, 0)\Big|_{t=0} = \frac{\partial^r}{\partial t^r} K_X(t)\Big|_{t=0}.$$

We conclude that $\kappa_{r,0}$ is the rth cumulant of X and, in particular, that $\kappa_{2,0} = \mathrm{Var}(X)$.

We also need to show that $\kappa_{1,1} = \mathrm{Cov}(X, Y)$. We know that:

$$M_{X,Y}(t, u) = 1 + \mu'_{1,0}t + \mu'_{0,1}u + \mu'_{1,1}tu + \cdots .$$

Taking partial derivatives yields:

$$\frac{\partial}{\partial u} K_{X,Y}(t, u) = \frac{\mu'_{0,1} + \mu'_{1,1} + \cdots}{M_{X,Y}(t, u)}$$

and:

$$\frac{\partial^2}{\partial t\, \partial u} K_{X,Y}(t, u) = \frac{\mu'_{1,1} + \cdots}{M_{X,Y}(t, u)} - \frac{(\mu'_{0,1} + \mu'_{1,1}t + \cdots)(\mu'_{1,0} + \mu'_{1,1}u + \cdots)}{(M_{X,Y}(t, u))^2}.$$

Evaluating the derivative at $t = 0$ and $u = 0$ gives:

$$\kappa_{1,1} = \frac{\partial^2}{\partial t\, \partial u} K_{X,Y}(t, u)\Big|_{t=0,u=0} = \mu'_{1,1} - \mu'_{0,1}\mu'_{1,0}.$$

Hence:

$$\kappa_{1,1} = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = \mathrm{Cov}(X, Y).$$

■

Appendix B
Solutions to Activities

B.1 Chapter 2 – Probability space

1. (a) The balls are labelled $1, 2, \ldots, 5$. A suitable sample space is the set of all possible pairs of the five balls, with order taken into account. There are 25 such possible pairs. This gives:

$$\Omega = \{(1,1), (1,2), \ldots, (5,5)\}.$$

In the without replacement case, we remove the pairs $(1,1), (1,2), \ldots, (5,5)$ from the sample space. The sample spaces for each case can be represented as follows (each * denotes a possible outcome):

with repl.		First ball				
		1	2	3	4	5
Second ball	1	*	*	*	*	*
	2	*	*	*	*	*
	3	*	*	*	*	*
	4	*	*	*	*	*
	5	*	*	*	*	*

w/o repl.		First ball				
		1	2	3	4	5
Second ball	1		*	*	*	*
	2	*		*	*	*
	3	*	*		*	*
	4	*	*	*		*
	5	*	*	*	*	

We now mark the events as follows: $E_1 \sim \#$, $E_2 \sim \circ$ and $E_3 \sim \square$.

with repl.		First ball				
		1	2	3	4	5
Second ball	1	$\circ$	$\circ$	$\circ$	$\circ$	$\#\circ$
	2	$\circ$	$\circ$	$\circ$	$\circ$	$\#\circ$
	3	$\circ$	$\circ$	$\circ$	$\circ$	$\#\circ\square$
	4				$\square$	$\#\square$
	5			$\square$	$\square$	$\#\square$

w/o repl.		First ball				
		1	2	3	4	5
Second ball	1		$\circ$	$\circ$	$\circ$	$\#\circ$
	2	$\circ$		$\circ$	$\circ$	$\#\circ$
	3	$\circ$	$\circ$		$\circ$	$\#\circ\square$
	4					$\#\square$
	5			$\square$	$\square$	

(b) The outcomes in the sample space are equally likely so in order to calculate probabilities we simply count outcomes.

	With replacement	Without replacement
$P(E_1)$	1/5	1/5
$P(E_2)$	3/5	3/5
$P(E_3)$	6/25	1/5
$P(E_1 \cap E_2)$	3/25	3/20
$P(E_2 \cup E_3)$	4/5	3/4
$P((E_1 \cap E_3) \cup E_2^c)$	11/25	9/20
$P(E_1 \cap (E_3 \cup E_2^c))$	3/25	1/10

2. It is easiest to use a sample space which has equally likely outcomes, and then to count outcomes to find the required probabilities. Suppose we number the positions in the row from 1 to 6. A and B are equally likely to occupy any two positions which are distinct. We can use a diagram to show the outcome space and mark each outcome with the number of people between A and B. This gives the following.

Standing in a row

		Position of A					
		1	2	3	4	5	6
Position of B	1		0	1	2	3	4
	2	0		0	1	2	3
	3	1	0		0	1	2
	4	2	1	0		0	1
	5	3	2	1	0		0
	6	4	3	2	1	0	

Standing in a ring (counting clockwise)

		Position of A					
		1	2	3	4	5	6
Position of B	1		4	3	2	1	0
	2	0		4	3	2	1
	3	1	0		4	3	2
	4	2	1	0		4	3
	5	3	2	1	0		4
	6	4	3	2	1	0	

Counting the number of outcomes for each case gives:

	Standing in a row	Standing in a ring
$P(0)$	1/3	1/5
$P(1)$	4/15	1/5
$P(2)$	1/5	1/5
$P(3)$	2/15	1/5
$P(4)$	1/15	1/5

Now it is easy to generalise to the case of n positions. There are $n^2 - n = n(n-1)$ equally likely outcomes. In the case of people standing in a row, $2(n-r-1)$ of the positions have r people between A and B. So, for $r = 0, 1, 2, \ldots, n-2$, we have:

$$P(r \text{ people between } A \text{ and } B) = \frac{2(n-r-1)}{n(n-1)}.$$

If the people are standing in a ring, all numbers of people between 0 and $n-2$ have the same probability, i.e. $1/(n-1)$.

3. (a) We can work out the smallest possible σ-algebra on Ψ using the definition. If $\mathcal{B}$ is a σ-algebra then, by property i., $\emptyset \in \mathcal{B}$. By property ii., we must also have $\Psi \in \mathcal{B}$. The third property is then automatically satisfied since $\emptyset \cup \Psi = \Psi$ and $\Psi \in \mathcal{B}$. Therefore, the smallest possible σ-algebra is $\mathcal{B} = \{\emptyset, \Psi\}$.

   (b) Consider the collection of subsets $\{\emptyset, A, A^c, \Psi\}$. Properties i. and ii. of a σ-algebra are trivially satisfied. Property iii. follows since for any set $A \subseteq \Psi$ we have, $A \cup \emptyset = A$, $A \cup A^c = \Psi$ and $A \cup \Psi = \Psi$.

   (c) Consider a list of all of the elements of Ψ. For any subset we could write a string of 0s and 1s below this list, where 1 indicates a member of the subset and 0 indicates a non-member. Therefore, there is a one-to-one correspondence between subsets of Ψ and binary numbers with $n(\Psi)$ digits. We conclude that the size of the power set is $2^{n(\Psi)}$.

(d) We use the properties of a σ-algebra and de Morgan's laws:

$$\begin{aligned} A_1, A_2, \ldots \in \mathcal{B} &\Rightarrow A_1^c, A_2^c, \ldots \in \mathcal{B} && \text{(property ii.)} \\ &\Rightarrow \bigcup_{i=1}^{\infty} A_i^c \in \mathcal{B} && \text{(property iii.)} \\ &\Rightarrow \left(\bigcap_{i=1}^{\infty} A_i\right)^c \in \mathcal{B} && \text{(de Morgan's laws)} \\ &\Rightarrow \bigcap_{i=1}^{\infty} A_i \in \mathcal{B}. && \text{(property ii.)} \end{aligned}$$

4. A and B partition Ψ into four subsets which are irreducible in terms of A and B. These are $A \cap B$, $A \cap B^c$, $A^c \cap B$ and $A^c \cap B^c$ (a Venn diagram may help you to visualise this). Any σ-algebra containing A and B will also contain all possible unions of these elements. Therefore, the size of the smallest σ-algebra containing A and B is $2^4 = 16$.

5. This is just a question of checking that the requirements of the definitions are met.
   (a) This comes directly from the properties of counting.
   (b) Property ii. holds since no element can be in the empty set, $I_\emptyset(x) = 0$ for any given x. Property iii. holds since:

$$m\left(\bigcup_{i=1}^{n} A_i\right) = \begin{cases} 1 & \text{for } x \in A_i \text{ for some } i \\ 0 & \text{otherwise.} \end{cases}$$

   It is clear that:

$$\sum_{i=1}^{\infty} m(A_i) = \sum_{i=1}^{\infty} I_{A_i}(x)$$

   can be defined in precisely the same manner.

6. Another exercise in checking properties. All we really need to check (in addition to what was done for the previous activities) is that $P(\Omega) = 1$ for each defined measure.

7. To prove $B_i \in \mathcal{F}$ for all i, note that:

$$B_i = A_i \setminus \bigcup_{j=1}^{i-1} B_j = A_i \cap \left(\bigcup_{j=1}^{i-1} B_j\right)^c = A_i \cap \left(\bigcap_{j=1}^{i-1} B_j^c\right).$$

   Since σ-algebras are closed under complements and under intersections, if $B_j \in \mathcal{F}$ for $j < i$, then $\bigcap_{j=1}^{i-1} B_j^c \in \mathcal{F}$, and hence $B_i \in \mathcal{F}$. As we have $B_1 = A_1 \in \mathcal{F}$, the result is true by induction.

Next we want to show $\bigcup_{j=1}^{i} B_j = \bigcup_{j=1}^{i} A_j$ for all i. We will use induction and assume the result is true for $i < j$. Therefore:

$$B_i = A_i \setminus \bigcup_{j=1}^{i-1} B_j = A_i \setminus \bigcup_{j=1}^{i-1} A_j.$$

Hence:

$$\bigcup_{j=1}^{i} B_j = B_i \cup \left(\bigcup_{j=1}^{i-1} B_j \right) = \left(A_i \setminus \bigcup_{j=1}^{i-1} A_j \right) \cup \left(\bigcup_{j=1}^{i-1} A_j \right) = \bigcup_{j=1}^{i} A_j.$$

Finally, to show $B_i \cap B_j = \emptyset$ for any $i \neq j$ we assume, without loss of generality, that $i < j$ and use the result that if $B \subseteq C$ then $B \cap (A \setminus C) = \emptyset$. You should satisfy yourself that this result is true and work out the details of how it applies in this case.

8. The most reliable method to approaching this problem is to think in terms of a sample space of equally likely outcomes. Let A denote the event that the fraud goes undetected.

   (a) There are 10 ways of choosing the first claim to investigate and 9 ways of choosing the second. Similarly, in order for the fraud to go undetected we must choose one of the 7 real claims first and one of the 6 remaining real claims second. Therefore:

   $$n(\Omega) = 10 \times 9 = 90, \quad n(A) = 7 \times 6 = 42 \quad \text{and} \quad P(A) = \frac{42}{90} = \frac{7}{15}.$$

   (b) We have:

   $$n(\Omega) = 5 \times 5 = 25, \quad n(A) = 4 \times 3 = 12 \quad \text{and} \quad P(A) = \frac{12}{25}.$$

   (c) We have:

   $$n(\Omega) = 5 \times 5 = 25, \quad n(A) = 5 \times 2 = 10 \quad \text{and} \quad P(A) = \frac{10}{25} = \frac{2}{5}.$$

   The optimal strategy is the one with the greatest probability that the fraud goes undetected, i.e. strategy (b).

9. The result follows immediately from the fact that, if $A \subseteq C$, then as $A \cap C = A$, we have:

$$P(A \,|\, C) = \frac{P(A)}{P(C)}.$$

10. We use the fact that:

$$P(A \cap B) = P(B \,|\, A)\, P(A).$$

Note that, if $k \leq n$, then:

$$P\left(\bigcap_{i=1}^{n} A_i\right) \neq 0 \quad \text{and} \quad \bigcap_{i=1}^{k} A_i \subseteq \bigcap_{i=1}^{n} A_i \quad \Rightarrow \quad P\left(\bigcap_{i=1}^{k} A_i\right) \neq 0.$$

Suppose that the result holds for $k \leq n$, then:

$$\begin{aligned} P\left(\bigcap_{j=1}^{k+1} A_j\right) &= P\left(A_{k+1} \cap \left(\bigcap_{j=1}^{k} A_j\right)\right) \\ &= P\left(A_{k+1} \,\middle|\, \bigcap_{j=1}^{k} A_j\right) P\left(\bigcap_{j=1}^{k} A_j\right) \quad \text{(conditional probability definition)} \\ &= P\left(A_{k+1} \,\middle|\, \bigcap_{i=1}^{k} A_i\right) \prod_{j=1}^{k} P\left(A_j \,\middle|\, \bigcap_{i=1}^{j-1} A_i\right) \qquad \text{(by initial assumption)} \\ &= \prod_{j=1}^{k+1} P\left(A_j \,\middle|\, \bigcap_{i=1}^{j-1} A_i\right). \end{aligned}$$

Clearly, $P(A_1) = P(A_1 \,|\, A_0)$ so the result holds by induction.

11. In order to establish the independence of four events, we must check all the pairs, all of the combinations of three events and the final condition involving all four events. Therefore:

$$\text{number of conditions} = \binom{4}{2} + \binom{4}{3} + \binom{4}{4} = 6 + 4 + 1 = 11.$$

For n events:

$$\text{number of conditions} = \sum_{j=1}^{n} \binom{n}{j} = 2^n - (n+1).$$

12. This question involves checking a lot of conditions. Many of these are trivially satisfied. However, to answer the question properly a clear explanation of why each condition is satisfied must be given.

(a) It is left as an exercise to check that the first two conditions for a measure are met. The third holds since:

$$\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B = \bigcup_{i=1}^{\infty} (A_i \cap B)$$

and if the A_is are mutually exclusive, then so are the $(A_i \cap B)$s. We can see that $P(A \,|\, B) \leq 1$ since:

$$A \cap B \subseteq B \quad \Rightarrow \quad P(A \cap B) \leq P(B).$$

(b) The properties of a σ-algebra are clearly inherited from $\mathcal{F}$. In this case we have to be a bit careful with the notation when establishing the second property. In the usual context $A^c = \Omega \setminus A$. However, as we are interested in $\mathcal{B}$ as a σ-algebra on B, we want to show that, if $C \in \mathcal{B}$ then $B \setminus C \in \mathcal{B}$. The argument goes as follows. If $C \in \mathcal{B}$, then $C = A \cap B$ for some $A \in \mathcal{F}$. By elementary properties of sets:

$$B \setminus C = B \setminus A = B \cap A^c.$$

We know that $A^c \in \mathcal{F}$, so by definition $B \setminus C \in \mathcal{B}$.

13. (a) First note that all of these are probabilities of the sum of mutually exclusive events, so:

$$P(A_0) + P(A_1) = \frac{1}{2}, \quad P(A_1) + P(A_2) = \frac{1}{2} \quad \text{and} \quad P(A_2) + P(A_0) = \frac{2}{3}.$$

Adding up these three equations, we get:

$$2 \times (P(A_0) + P(A_1) + P(A_2)) = \frac{5}{3} \quad \Rightarrow \quad P(A_0) + P(A_1) + P(A_2) = \frac{5}{6}.$$

However, we are told that A_0, A_1 and A_2 are collectively exhaustive, so $P(A_0) + P(A_1) + P(A_2) = 1$. Therefore, it is not possible to have the probabilities suggested in the question.

(b) This is a generalisation of the previous question. The condition is on the sum of the π_rs. First note that, by mutual exclusivity, we have:

$$P\left(\bigcup_{i=r}^{r+k-1} A_{i(\text{mod } n)}\right) = \sum_{i=r}^{r+k-1} P(A_{i(\text{mod } n)}).$$

This implies that:

$$\sum_{r=0}^{n-1} \pi_t = \sum_{r=0}^{n-1} \sum_{i=r}^{r+k-1} P(A_{i(\text{mod } n)}).$$

Writing this sum out in full gives:

$$\begin{aligned}\sum_{r=0}^{n-1} \pi_r = {} & P(A_0) + P(A_1) + \cdots + P(A_{k-1}) \\ & + P(A_1) + P(A_2) + \cdots + P(A_k) \\ & \vdots \qquad \vdots \qquad \cdots \qquad \vdots \\ & + P(A_{n-1}) + P(A_0) + \cdots + P(A_{k-2}).\end{aligned}$$

Each column above contains each of $P(A_0), P(A_1), \ldots, P(A_{n-1})$ exactly once so, by exhaustiveness, sums to 1. There are k columns so we conclude that:

$$\sum_{r=0}^{n-1} \pi_r = k.$$

14. A diagram helps in this question. From a simple sketch of the sample space it should be clear that each of A_1, A_2 and A_3 has three elements and they have one element in common, namely aaa. Therefore, for $i \neq j$, we have:

$$P(A_i) = P(A_i \mid A_j) = P(A_i \mid A_j^c) = \frac{1}{3}$$

and:

$$P(A_i \cap A_j) = P(A_1 \cap A_2 \cap A_3) = \frac{1}{9}.$$

We conclude that the events A_1, A_2 and A_3 are pairwise independent, but not mutually independent.

15. (a) We can use a sample space of eight equally likely outcomes if we take into account birth order. Using M and F for boys and girls, respectively, we can write:

$$\Omega = \{MMM, MMF, MFM, MFF, FMM, FMF, FFM, FFF\}.$$

Hence:

$$A = \{MMF, MFM, MFF, FMM, FMF, FFM\}$$

also:

$$B = \{MMM, MMF, MFM, FMM\}$$

and:

$$A \cap B = \{MMF, MFM, FMM\}.$$

Counting the outcomes in the events gives:

$$P(A) = \frac{6}{8} = \frac{3}{4}, \quad P(B) = \frac{4}{8} = \frac{1}{2} \quad \text{and} \quad P(A \cap B) = \frac{3}{8}.$$

(b) For families with four children, there are 16 equally likely family outcomes. Just two of these have all the children of the same sex, so we have that $P(A) = 14/16 = 7/8$, and there is $\binom{4}{0} = 1$ family with all boys and $\binom{4}{1} = 4$ families with 1 girl. All the other families have more than 1 girl, so we have that $P(B) = 5/16$. There are four families with children of both sexes and no more than 1 girl, so $P(A \cap B) = 4/16 = 1/4$.

(c) In (a) we have $P(A \cap B) = P(A)\,P(B)$, so A and B are independent. It is hard to justify this independence intuitively. In (b) we have that A and B are not independent because:

$$P(A \cap B) = \frac{1}{4} \neq \frac{7}{8} \times \frac{5}{16} = P(A)\,P(B).$$

(d) Let K be the number of boys in a family and let the family size be N. Hence

by the total probability formula, for $k = 0, 1, 2, \ldots$, we have:

$$\begin{aligned} P(K=k) &= \sum_{n=k}^{\infty} P(K=k \,|\, N=n)\, P(N=n) \\ &= \sum_{n=k}^{\infty} \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} (1-\pi)\pi^n \\ &= (1-\pi)\pi^k \sum_{n=k}^{\infty} \binom{n}{k} \left(\frac{1}{2}\right)^n \pi^{n-k} \\ &= (1-\pi)\pi^k \left(\frac{1}{2}\right)^k \sum_{n=k}^{\infty} \binom{n}{k} \left(\frac{\pi}{2}\right)^{n-k}. \end{aligned}$$

All that is needed is the sum of the series:

$$\sum_{n=k}^{\infty} \binom{n}{k} \left(\frac{\pi}{2}\right)^{n-k}$$

which is easy for $k = 0$ since it is just a geometric series. For the general case we need a negative binomial series expansion. Our series is equivalent to (change n to $m + k$):

$$\sum_{m=0}^{\infty} \binom{m+k}{k} \left(\frac{\pi}{2}\right)^m$$

which, since $\binom{m+k}{k} = \binom{m+k}{m}$, is:

$$\sum_{m=0}^{\infty} \binom{m+k}{m} \left(\frac{\pi}{2}\right)^m$$

which is:

$$\sum_{m=0}^{\infty} \binom{m+(k+1)-1}{m} \left(\frac{\pi}{2}\right)^m.$$

Now the sum is recognisable as the negative binomial series:

$$\left(1-\frac{\pi}{2}\right)^{-(k+1)}$$

and substituting this, the required probability is:

$$(1-\pi)\,\pi^k \left(\frac{1}{2}\right)^k \left(1-\frac{\pi}{2}\right)^{-(k+1)}$$

which is:

$$\frac{2(1-\pi)\pi^k}{(2-\pi)^{k+1}}.$$

B.2 Chapter 3 – Random variables and univariate distributions

1. It is essential to distinguish between *events* and the *probabilities of events*.
   (a) $\{X \leq 2\}$ is the *event* that the absolute value of the difference between the values is at most 2.
   (b) $\{X = 0\}$ is the *event* that both dice show the same value.
   (c) $P(X \leq 2)$ is the *probability* that the absolute value of the difference between the values is at most 2.
   (d) $P(X = 0)$ is the *probability* that both dice show the same value.

2. We have the following given the definitions of the random variables X and Y.
   (a) $\{X < 3\}$.
   (b) $P(X < 3)$.
   (c) $\{Y = 1\}$.
   (d) $P(Y = 0)$.
   (e) $P(X = 6, Y = 0)$.
   (f) $P(Y < X)$.

3. Let Y denote the value of claims paid. The distribution function of Y is:

$$F_Y(x) = P(Y \leq x) = P(X \leq x \,|\, X > k) = \frac{P(X \leq x \cap X > k)}{P(X > k)} = \frac{P(k < X \leq x)}{P(X > k)}.$$

   Hence, in full:

$$F_Y(x) = \begin{cases} 0 & \text{for } x \leq k \\ \dfrac{F_X(x) - F_X(k)}{1 - F_X(k)} & \text{for } x > k. \end{cases}$$

   Let Z denote the value of claims not paid. The distribution function of Z is:

$$F_Z(x) = P(Z \leq x) = P(X \leq x \,|\, X \leq k) = \frac{P(X \leq x \cap X \leq k)}{P(X \leq k)} = \frac{P(X \leq x)}{P(X \leq k)}.$$

   Hence, in full:

$$F_Z(x) = \begin{cases} \dfrac{F_X(x)}{F_X(k)} & \text{for } x \leq k \\ 1 & \text{for } x > k. \end{cases}$$

4. This problem makes use of the following results from mathematics, concerning sums of geometric series. If $r \neq 1$, then:

$$\sum_{x=0}^{n-1} ar^x = \frac{a(1 - r^n)}{1 - r}$$

and if $|r| < 1$, then:

$$\sum_{x=0}^{\infty} ar^x = \frac{a}{1-r}.$$

(a) We first note that p_X is a positive real-valued function with respect to its support. Noting that $1 - \pi < 1$, we have:

$$\sum_{x=1}^{\infty}(1-\pi)^{x-1}\pi = \frac{\pi}{1-(1-\pi)} = 1.$$

Hence the two necessary conditions for a valid mass function are satisfied.

(b) The distribution function for the (first version of the) geometric distribution is:

$$F_X(x) = \sum_{t:t\leq x} p_X(t) = \sum_{t=1}^{x}(1-\pi)^{t-1}\pi = \frac{\pi(1-(1-\pi)^x)}{1-(1-\pi)} = 1-(1-\pi)^x.$$

In full:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 1 \\ 1-(1-\pi)^{\lfloor x \rfloor} & \text{for } x \geq 1. \end{cases}$$

5. We first note that p_X is a positive real-valued function with respect to its support. We then have:

$$\sum_{x=r}^{\infty} p_X(x) = \sum_{x=r}^{\infty}\binom{x-1}{r-1}\pi^r(1-\pi)^{x-r} = \pi^r\sum_{y=0}^{\infty}\binom{y+r-1}{r-1}(1-\pi)^y = \pi^r(1-(1-\pi))^{-r} = 1$$

where $y = x - r$. Hence the two necessary conditions for a valid mass function are satisfied.

6. We have:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f_X(x)\,\mathrm{d}x = \int_0^2 \left(x^2 - \frac{x^4}{4}\right)\mathrm{d}x = \left[\frac{x^3}{3} - \frac{x^5}{20}\right]_0^2 = \frac{16}{15}.$$

Also:

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x = \int_0^2 \left(x^3 - \frac{x^5}{4}\right)\mathrm{d}x = \left[\frac{x^4}{4} - \frac{x^6}{24}\right]_0^2 = \frac{4}{3}.$$

Hence:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{4}{3} - \left(\frac{16}{15}\right)^2 = \frac{44}{225}.$$

7. We need to evaluate:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, \mathrm{d}x = \int_0^{\infty} x\, \lambda\, \mathrm{e}^{-\lambda x}\, \mathrm{d}x.$$

We note that:

$$x\, \mathrm{e}^{-\lambda x} = \frac{x}{\mathrm{e}^{\lambda x}} = \frac{x}{1 + \lambda x + \lambda^2 x^2/2 + \cdots} = \frac{1}{1/x + \lambda + \lambda^2 x/2 + \cdots}$$

such that the numerator is fixed (the constant 1), and the denominator tends to infinity as $x \to \infty$. Hence:

$$x\, \mathrm{e}^{-\lambda x} \to 0 \quad \text{as} \quad x \to \infty.$$

Applying integration by parts, we have:

$$\mathrm{E}(X) \int_0^{\infty} x\, \lambda \mathrm{e}^{-\lambda x} \mathrm{d}x = \Big[-x\, \mathrm{e}^{-\lambda x} \Big]_0^{\infty} + \int_0^{\infty} \mathrm{e}^{-\lambda x}\, \mathrm{d}x = 0 + \left[-\frac{\mathrm{e}^{-\lambda x}}{\lambda} \right]_0^{\infty} = \frac{1}{\lambda}.$$

8. (a) We have:

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}((X - \mathrm{E}(X))^2) \\ &= \mathrm{E}(X^2 - 2X\, \mathrm{E}(X) + (\mathrm{E}(X))^2) \\ &= \mathrm{E}(X^2) - 2(\mathrm{E}(X))^2 + (\mathrm{E}(X))^2 \\ &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2. \end{aligned}$$

(b) We have:

$$\begin{aligned} \mathrm{E}(X(X-1)) - \mathrm{E}(X)\, \mathrm{E}(X-1) &= \mathrm{E}(X^2) - \mathrm{E}(X) - (\mathrm{E}(X))^2 + \mathrm{E}(X) \\ &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 \\ &= \mathrm{Var}(X). \end{aligned}$$

9. We note that:

$$x^2 \lambda \mathrm{e}^{-\lambda x} = \lambda\, \frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} \mathrm{e}^{-\lambda x}$$

and so:

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\, \mathrm{d}x = \int_0^{\infty} x^2 \lambda \mathrm{e}^{-\lambda x} \mathrm{d}x = \lambda\, \frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} \lambda^{-1} = \frac{2}{\lambda^2}.$$

Therefore:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

10. We consider the proof for a continuous random variable X. The indicator function takes the value 1 for values $\leq x$ and 0 otherwise. Hence:

$$\mathrm{E}(I_{(-\infty,x]}(X)) = \int_{-\infty}^{\infty} I_{(-\infty,x]}(t)\, f_X(t)\, \mathrm{d}t = \int_{-\infty}^{x} f_X(t)\, \mathrm{d}t = F_X(x).$$

11. (a) Let the random variable X denote the return from the game. Hence X is a discrete random variable, which can take three possible values: $-$£5 (if we do not throw a head first and last), £15 (if we throw HTH) and £25 (if we throw HHH). The probabilities associated with these values are 3/4, 1/8 and 1/8, respectively. Therefore, the expected return from playing the game is:

$$\mathrm{E}(X) = \sum_x x\, p_X(x) = -5 \times \frac{3}{4} + 15 \times \frac{1}{8} + 25 \times \frac{1}{8} = £1.25.$$

(b) The moment generating function is:

$$M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \sum_x \mathrm{e}^{tx} p_X(x) = \mathrm{e}^{-5t} \times \frac{3}{4} + \mathrm{e}^{15t} \times \frac{1}{8} + \mathrm{e}^{25t} \times \frac{1}{8} = \frac{\mathrm{e}^{-5t}}{8}(6 + \mathrm{e}^{20t} + \mathrm{e}^{30t}).$$

12. If $X \sim \mathrm{Exp}(\lambda)$, we have that:

$$M_X(t) = \frac{\lambda}{\lambda - t} = \frac{1}{1 - t/\lambda} \quad \text{for } t < \lambda.$$

Writing as a polynomial in t, we have:

$$M_X(t) = \frac{1}{1 - t/\lambda} = \sum_{i=0}^{\infty} \left(\frac{t}{\lambda}\right)^i$$

since $t/\lambda < 1$. Since the coefficient of t^r is $\mathrm{E}(X^r)/r!$ in the polynomial expansion of $M_X(t)$, for the exponential distribution we have:

$$\frac{\mathrm{E}(X^r)}{r!} = \frac{1}{\lambda^r} \quad \Rightarrow \quad \mathrm{E}(X^r) = \frac{r!}{\lambda^r}.$$

13. (a) If $X \sim \mathrm{Bernoulli}(\pi)$, then the moment generating function is:

$$M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \sum_x \mathrm{e}^{tx} p_X(x) = \sum_{x=0}^{1} \mathrm{e}^{tx} \pi^x (1-\pi)^{1-x} = (1-\pi) + \pi \mathrm{e}^t$$

and hence the cumulant generating function is:

$$K_X(t) = \log M_X(t) = \log((1-\pi) + \pi \mathrm{e}^t).$$

(b) If $X \sim \mathrm{Bin}(n, \pi)$, then the moment generating function is:

$$\begin{aligned} M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) &= \sum_x \mathrm{e}^{tx} p_X(x) \\ &= \sum_{x=0}^{n} \mathrm{e}^{tx} \binom{n}{x} \pi^x (1-\pi)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (\pi \mathrm{e}^t)^x (1-\pi)^{n-x} \\ &= ((1-\pi) + \pi \mathrm{e}^t)^n \end{aligned}$$

using the binomial expansion. Hence the cumulant generating function is:

$$K_X(t) = \log M_X(t) = \log(((1-\pi) + \pi \mathrm{e}^t)^n) = n \log((1-\pi) + \pi \mathrm{e}^t).$$

Note that if $X \sim \text{Bernoulli}(\pi)$ and $Y \sim \text{Bin}(n, \pi)$, then:

$$M_Y(t) = (M_X(t))^n \quad \text{and} \quad K_Y(t) = nK_X(t).$$

This is not a coincidence, as a $\text{Bin}(n, \pi)$ random variable is equal to the sum of n independent $\text{Bernoulli}(\pi)$ random variables (since π is constant, the sum of n independent and identically distributed Bernoulli random variables).

(c) If $X \sim \text{Geo}(\pi)$, then the moment generating function is:

$$\begin{aligned} M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) &= \sum_x \mathrm{e}^{tx} p_X(x) \\ &= \sum_{x=1}^{\infty} \mathrm{e}^{tx} (1-\pi)^{x-1} \pi \\ &= \pi \mathrm{e}^t \sum_{x=1}^{\infty} ((1-\pi)\mathrm{e}^t)^{x-1} \\ &= \frac{\pi \mathrm{e}^t}{1 - (1-\pi)\mathrm{e}^t} \end{aligned}$$

provided $|(1-\pi)\mathrm{e}^t| < 1$. Hence the cumulant generating function is:

$$K_X(t) = \log M_X(t) = \log\left(\frac{\pi \mathrm{e}^t}{1 - (1-\pi)\mathrm{e}^t}\right).$$

(d) If $X \sim \text{Neg. Bin}(r, \pi)$, then the moment generating function is:

$$\begin{aligned} M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) &= \sum_x \mathrm{e}^{tx} p_X(x) \\ &= \sum_{x=r}^{\infty} \mathrm{e}^{tx} \binom{x-1}{r-1} \pi^x (1-\pi)^{x-r} \\ &= (\pi \mathrm{e}^t)^r \sum_{x=r}^{\infty} \frac{(x-1)!}{(r-1)!\,(x-r)!} ((1-\pi)\mathrm{e}^t)^{x-r} \\ &= (\pi \mathrm{e}^t)^r \sum_{y=0}^{\infty} \frac{(y+r-1)!}{(r-1)!\,y!} ((1-\pi)\mathrm{e}^t)^y \qquad (\text{setting } y = x-r) \\ &= \left(\frac{\pi \mathrm{e}^t}{1 - (1-\pi)\mathrm{e}^t}\right)^r \end{aligned}$$

using the negative binomial expansion. Hence the cumulant generating function is:

$$K_X(t) = \log M_X(t) = \log\left(\left(\frac{\pi \mathrm{e}^t}{1 - (1-\pi)\mathrm{e}^t}\right)^r\right).$$

Note the relationship between the moment generating functions of Geo(π) and Neg. Bin(r, π). A Neg. Bin(r, π) random variable is equal to the sum of r independent Geometric(π) random variables (since π is constant, the sum of r independent and identically distributed geometric random variables).

14. If $X \sim \text{Pois}(\lambda)$, then we know:

$$M_X(t) = \exp(\lambda(\mathrm{e}^t - 1)).$$

Hence the cumulant generating function is:

$$K_X(t) = \log M_X(t) = \log\left(\exp(\lambda(\mathrm{e}^t - 1))\right) = \lambda(\mathrm{e}^t - 1) = \lambda\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots\right).$$

By comparing coefficients, the third cumulant is $\kappa_3 = \lambda$. We also have that $\mu_2' = \text{Var}(X) = \lambda$ and $\mu_3' = \kappa_3 = \lambda$. Therefore, the coefficient of skewness is:

$$\gamma_1 = \frac{\mu_3'}{(\mu_2')^{3/2}} = \frac{1}{\sqrt{\lambda}}.$$

15. Let $Y = |X|$. We have:

$$F_Y(y) = P(Y \leq y) = P(|X| \leq y) = P(-y \leq X \leq y) = F_X(y) - F_X(-y).$$

In full, the distribution function of Y is:

$$F_Y(y) = \begin{cases} F_X(y) - F_X(-y) & \text{for } y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The density function is obtained by differentiating, hence:

$$f_Y(y) = \begin{cases} f_X(y) + f_X(-y) & \text{for } y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

16. Let $Y = |X - \mu|$, where $X \sim N(\mu, \sigma^2)$, hence $X - \mu \sim N(0, \sigma^2)$. Therefore:

$$f_X(y) + f_X(-y) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{-y^2/(2\sigma^2)} + \frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{-y^2/(2\sigma^2)} = \sqrt{\frac{2}{\pi\sigma^2}}\,\mathrm{e}^{-y^2/(2\sigma^2)}.$$

In full, the density function of Y is:

$$f_Y(y) = \begin{cases} \sqrt{\dfrac{2}{\pi\sigma^2}}\,\mathrm{e}^{-y^2/(2\sigma^2)} & \text{for } y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

17. Let $Y = 1/X$. As X is a positive random variable, the function $g(x) = 1/x$ is well-behaved and monotonic. Therefore:

$$f_Y(y) = \begin{cases} f_X(1/y)/y^2 & \text{for } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

18. If $X \sim \text{Exp}(\lambda)$, then its density function is:

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

Hence:

$$f_Y(y) = f_X\left(\frac{1}{y}\right)\left|-\frac{1}{y^2}\right| = \lambda \mathrm{e}^{\lambda/y} y^{-2}.$$

In full, the density function of Y is:

$$f_Y(y) = \begin{cases} \lambda y^{-2} \mathrm{e}^{-\lambda/y} & \text{for } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

19. A distribution is uniquely characterised by its cumulant generating function. Therefore, if $K_{X_n}(t) \to K_X(t)$ as $n \to \infty$, then $X_n \xrightarrow{d} x$.

20. This is a special case of the result that convergence in probability implies convergence in distribution. Note that convergence in distribution requires convergence to the distribution function, except at discontinuities. We note that:

$$P(|X_n - a| < \epsilon) = (1 - F_{X_n}(a + \epsilon)) + F_{X_n}(a - \epsilon) + P(X_n = a - \epsilon).$$

Since X_n converges in probability to a, for any $\epsilon > 0$, the left-hand side converges to zero. Each element on the right-hand side is positive, so we must have:

$$F_{X_n}(a + \epsilon) \to 1 \quad \text{and} \quad F_{X_n}(a - \epsilon) \to 0.$$

Therefore, at each point where the distribution function is continuous, $F_{X_n} \to F_X$, where F_X is the distribution function of a degenerate random variable with all mass at a. Hence $X_n \xrightarrow{d} a$.

B.3 Chapter 4 – Multivariate distributions

1. If $F_{X_1,X_2,\ldots,X_n}$ is the joint distribution function of $X_1, X_2, \ldots, X_n$, then for any $i = 1, 2, \ldots, n$ we have:
   - $\lim_{x_i \to -\infty} F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = 0$
   - $\lim_{x_1 \to \infty, x_2 \to \infty, \ldots, x_n \to \infty} F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = 1$
   - $\lim_{h \downarrow 0} F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) = F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$.

2. (a) This cannot work, because:

$$G(\infty, \infty) = F_X(\infty) + F_Y(\infty) = 1 + 1 = 2$$

which is inconsistent with $G(\infty, \infty)$ being a probability.

(b) If we consider the case where X and Y are independent, with joint distribution function G, then:

$$G(x,y) = P(X \le x, Y \le y) = P(X \le x)\,P(Y \le y) = F_X(x)\,F_Y(y).$$

This verifies that $F_X(x)\,F_Y(y)$ is a joint distribution function.

(c) This one does not work because:

$$G(-\infty,\infty) = \max(F_X(-\infty), F_Y(\infty)) = \max(0,1) = 1$$

but this should be 0.

(d) Consider the singular joint distribution for which X and Y always take exactly the same value, and for which the marginal distribution function of X is $F_X(x)$. A typical value for (X,Y) is (x,x). The support of this distribution in the XY-plane is the line $y = x$. If G is the joint distribution function we have

$$G(x,y) = P(X \le x, Y \le y) = P(X \le x, X \le y) = \min(F_X(x), F_X(y)).$$

So we have shown that this is a distribution function, though for a singular distribution.

3. (a) We consider the sample space:

Outcome	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
X	3	2	2	1	2	1	1	0
Y	1	0	1	0	0	0	0	0

A table summarising the joint mass function is:

		$X = x$			
$p_{X,Y}(x,y)$		0	1	2	3
$Y = y$	0	1/8	3/8	2/8	0
	1	0	0	1/8	1/8

(b) The marginal mass functions are added to the joint mass function as follows:

		$X = x$				
$p_{X,Y}(x,y)$		0	1	2	3	$p_Y(y)$
$Y = y$	0	1/8	3/8	2/8	0	3/4
	1	0	0	1/8	1/8	1/4
$p_X(x)$		1/8	3/8	3/8	1/8	1

The marginal mass functions are exactly as expected, since it is clear from how X and Y are defined that $X \sim \text{Bin}(3, 0.5)$ and $Y \sim \text{Bernoulli}(0.25)$.

4. (a) Integrating the joint density function over its support, we have:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^2\int_0^1 kxy\,\mathrm{d}x\,\mathrm{d}y = \int_0^2 \left[\frac{kx^2y}{2}\right]_0^1 \mathrm{d}y = \int_0^2 \frac{ky}{2}\,\mathrm{d}y$$

$$= \left[\frac{ky^2}{4}\right]_0^2$$

$$= k.$$

Since the integral of the joint density over its support is 1, immediately we have that $k = 1$.

(b) The marginal density functions are obtained by integrating out the other variable. Hence:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \int_0^2 xy\,\mathrm{d}y = \left[\frac{xy^2}{2}\right]_0^2 = 2x$$

so, in full:

$$f_X(x) = \begin{cases} 2x & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Similarly:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x = \int_0^1 xy\,\mathrm{d}x = \left[\frac{x^2y}{2}\right]_0^1 = \frac{y}{2}$$

so, in full:

$$f_Y(y) = \begin{cases} y/2 & \text{for } 0 \le y \le 2 \\ 0 & \text{otherwise.} \end{cases}$$

(c) We have (choosing to integrate with respect to y first, although the order of integration makes no difference):

$$\begin{aligned} P(X < Y) = \int_{-\infty}^{\infty}\int_{x}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x &= \int_0^1\int_x^2 xy\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_0^1 \left[\frac{xy^2}{2}\right]_x^2 \mathrm{d}x \\ &= \int_0^1 \left(2x - \frac{x^3}{2}\right)\mathrm{d}x \\ &= \left[x^2 - \frac{x^4}{8}\right]_0^1 \\ &= \frac{7}{8}. \end{aligned}$$

5. (a) Immediately, we note that $f_{X,Y}(x,y)$ is a positive real-valued function. Integrating, we have:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^1\int_0^y 2\,\mathrm{d}x\,\mathrm{d}y = \int_0^1 \Big[2x\Big]_0^y \mathrm{d}y = \int_0^1 2y\,\mathrm{d}y = \Big[y^2\Big]_0^1 = 1.$$

Hence the necessary conditions for a valid joint density are satisfied.

(b) The marginal density functions are obtained by integrating out the other variable. Hence:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \int_x^1 2\,\mathrm{d}y = \Big[2y\Big]_x^1 = 2(1-x)$$

so, in full:

$$f_X(x) = \begin{cases} 2(1-x) & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Similarly:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x = \int_0^y 2\,\mathrm{d}x = \Big[2x\Big]_0^y = 2y$$

so, in full:

$$f_Y(y) = \begin{cases} 2y & \text{for } 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

(c) We have (choosing to integrate with respect to x first, although the order of integration makes no difference):

$$\begin{aligned} P\left(Y < X + \frac{1}{2}\right) = \int_{-\infty}^{\infty}\int_{y-1/2}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y &= \int_0^{1/2}\int_0^y 2\,\mathrm{d}x\,\mathrm{d}y + \int_{1/2}^1\int_{y-1/2}^y 2\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^{1/2}\Big[2x\Big]_0^y\,\mathrm{d}y + \int_{1/2}^1\Big[2x\Big]_{y-1/2}^y\,\mathrm{d}y \\ &= \int_0^{1/2} 2y\,\mathrm{d}y + \int_{1/2}^1 1\,\mathrm{d}y \\ &= \frac{1}{4} + \frac{1}{2} \\ &= \frac{3}{4}. \end{aligned}$$

6. The individual random variables share the common density function:

$$f_X(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}.$$

Due to independence, the joint density function is the product of the individual densities, hence:

$$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = \prod_{i=1}^n f_{X_i}(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x_i^2/2} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right).$$

7. Multiplying out the brackets inside the expectation, we have:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))) \\ &= \mathrm{E}(XY - Y\,\mathrm{E}(X) - X\,\mathrm{E}(Y) + \mathrm{E}(X)\,\mathrm{E}(Y)) \\ &= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y). \end{aligned}$$

8. For independent random variables, the joint moment generating function is the product of the individual moment generating functions, that is:

$$M_{X_1,X_2,\ldots,X_n}(t_1,t_2,\ldots,t_n) = \mathrm{E}\left(\exp\left(\sum_{i=1}^n t_i X_i\right)\right) = \prod_{i=1}^n \mathrm{E}(\exp(t_i X_i)) = \prod_{i=1}^n M_{X_i}(t_i).$$

Taking logarithms, and noting the logarithm property that the 'logarithm of the product' is the 'sum of the logarithms', the joint cumulant generating function is:

$$K_{X_1,X_2,\ldots,X_n}(t_1,t_2,\ldots,t_n) = \log M_{X_1,X_2,\ldots,X_n}(t_1,t_2,\ldots,t_n) = \sum_{i=1}^{n} \log M_{X_i}(t_i) = \sum_{i=1}^{n} K_{X_i}(t_i).$$

9. By the rules of matrix multiplication, the (i,j)th entry in the matrix $\mathrm{Var}(\mathbf{X})$ is:

$$\mathrm{E}((X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))) = \mathrm{Cov}(X_i, X_j).$$

Since $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$, we have hence established the result.

10. (a) The (i,j)th entry in the matrix $\mathrm{Cov}(\mathbf{X}, \mathbf{Y})$, for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$, is:

$$\mathrm{E}((X_i - \mathrm{E}(X_i))(Y_j - \mathrm{E}(Y_j))) = \mathrm{Cov}(X_i, X_j).$$

(b) Similarly, the (j,i)th entry in the matrix $\mathrm{Cov}(\mathbf{Y}, \mathbf{X})$ is $\mathrm{Cov}(Y_j, X_i)$. By symmetry of covariance for the univariate case, we have that the (j,i)th entry in the matrix $\mathrm{Cov}(\mathbf{Y}, \mathbf{X})$ is equal to the (i,j)th entry in the matrix $\mathrm{Cov}(\mathbf{X}, \mathbf{Y})$. This proves the result.

11. The transformation is $x = u/v$ and $y = v$, hence the inverse transformation is $u = xy$ and $v = y$. The Jacobian is $(y \times 1 - 0 \times x) = y$. Therefore, the joint density function of the transformed variables is:

$$f_{X,Y}(x,y) = f_{U,V}(xy, y)\,|y|$$

and the marginal density function of u/v is:

$$f_X(x) = \int_{-\infty}^{\infty} f_{U,V}(xy, y)\,|y|\,\mathrm{d}y.$$

12. Since $U \sim \mathrm{Exp}(\lambda)$ and $V \sim \mathrm{Exp}(\lambda)$ are independent, then:

$$f_{U,V}(u,v) = f_U(u)\,f_V(v) = \begin{cases} \lambda^2 \mathrm{e}^{-\lambda(u+v)} & \text{for } u \geq 0 \text{ and } v \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

We have:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{U,V}(xy, y)\,|y|\,\mathrm{d}y \\ &= \int_0^{\infty} \lambda^2 y\,\mathrm{e}^{-\lambda(xy+y)}\,\mathrm{d}y \\ &= \frac{\lambda^2 \Gamma(2)}{(\lambda(x+1))^2} \int_0^{\infty} \frac{1}{\Gamma(2)} (\lambda(x+1))^2 y\,\mathrm{e}^{-\lambda(x+1)y}\,\mathrm{d}y \\ &= \frac{1}{(1+x^2)} \end{aligned}$$

for $x \geq 0$, and 0 otherwise. Note that the integrand is the density function of $\mathrm{Gamma}(2, \lambda(x+1))$.

13. As $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ is a vector of independent standard normal random variables, its joint density is:

$$f_{\mathbf{X}}(\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\,\mathbf{x}^T\mathbf{x}\right).$$

Defining $\mathbf{Y} = \mathbf{A}\mathbf{X}$, we have:

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \frac{1}{\|\mathbf{A}\|} f_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y}) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \|\mathbf{A}\|^{-1} \exp\left(-\frac{1}{2}\mathbf{y}^T(\mathbf{A}^{-1})^T\mathbf{A}^{-1}\mathbf{y}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \|\mathbf{\Sigma}\|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y}\right) \end{aligned}$$

where $\mathbf{\Sigma} = \mathrm{Var}(\mathbf{Y}) = \mathbf{A}\mathbf{A}^T$. Note this uses the fact that for non-singular matrices $\mathbf{A}$ and $\mathbf{B}$, we have:

$$|\mathbf{A}^T| = |\mathbf{A}|, \quad (\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T \quad \text{and} \quad (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}.$$

14. If $Z = X + Y$, then:

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(u, z-u)\,\mathrm{d}u.$$

The support of $f_{X,Y}$ is $\{(x,y) : 0 \le x \le y \le 1\}$. For the integrand above to take non-zero values, we require $0 \le u \le z - u \le 1$, which implies:

$$\begin{cases} 0 \le u \le z/2 & \text{for } 0 \le z \le 1 \\ z - 1 < u < z/2 & \text{for } 1 < z \le 2. \end{cases}$$

Therefore, the marginal density function of the sum is:

$$f_Z(z) = \begin{cases} \int_0^{z/2} 2\,\mathrm{d}u & \text{for } 0 \le z \le 1 \\ \int_{z-1}^{z/2} 2\,\mathrm{d}u & \text{for } 1 < z \le 2 \\ 0 & \text{otherwise.} \end{cases}$$

15. Since X and Y are positive random variables, for the $f_{X,Y}(u, z-u)$ term in:

$$\int_0^z f_{X,Y}(u, z-u)\,\mathrm{d}u$$

to be non-zero we require $u > 0$ and $z - u > 0 \Rightarrow u < z$. Hence the u integral is over the interval $[0, z]$.

16. Using the definition of the moment generating function and that the X_is are independent and identically distributed, we have:

$$M_S(t) = \mathrm{E}\left(\mathrm{e}^{t\sum_{i=1}^n X_i}\right) = \prod_{i=1}^n \mathrm{E}(\mathrm{e}^{tX_i}) = \prod_{i=1}^n M_{X_i}(t) = (M_X(t))^n$$

and:

$$K_S(t) = \log(M_S(t)) = n\log(M_X(t)) = nK_X(t).$$

17. Suppose that $X_i \sim \text{Bin}(n_i, \pi_i)$ and let $S = \sum_{i=1}^{n} X_i$. If the X_is are independent, then:

$$M_Z(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} (1 - \pi_i + \pi_i \mathrm{e}^t)^{n_i}.$$

If $\pi_i = \pi$ for all $i = 1, 2, \ldots, n$, then $S \sim \text{Bin}\left(\sum_{i=1}^{n} n_i, \pi\right)$.

18. Using the moment generating function of a gamma distribution, we have:

$$M_X(t) = \left(\frac{\beta_1}{\beta_1 - t}\right)^{\alpha_1} \quad \text{and} \quad M_Y(t) = \left(\frac{\beta_2}{\beta_2 - t}\right)^{\alpha_2}.$$

Since $Z = X + Y$, and X and Y are independent random variables, we have:

$$M_Z(t) = M_X(t)\, M_Y(t) = \left(\frac{\beta_1}{\beta_1 - t}\right)^{\alpha_1} \left(\frac{\beta_2}{\beta_2 - t}\right)^{\alpha_2}.$$

If $\beta_1 = \beta_2$, then $Z \sim \text{Gamma}(\alpha_1 + \alpha_2, \beta_1)$.

19. 1. We know that $X \sim N(0, 1)$ since $X = U$ and $U \sim N(0, 1)$. A sum of normal random variables is also normal, hence Y is normal. By definition of Y and independence of U and V, we have:

$$\mathrm{E}(Y) = 0 \quad \text{and} \quad \mathrm{Var}(Y) = \rho^2 \mathrm{Var}(U) + (1 - \rho^2)\,\mathrm{Var}(V) = 1.$$

Hence $Y \sim N(0, 1)$.

2. For the correlation we exploit the independence of U and V, and the fact that all the random variables involved have mean 0 and variance 1. We have:

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathrm{E}(XY) \qquad (\text{since } \mathrm{E}(X) = \mathrm{E}(Y) = 0) \\
&= \mathrm{E}(U(\rho U + \sqrt{1 - \rho^2} V)) \qquad (\text{from definition of } X \text{ and } Y) \\
&= \rho\, \mathrm{E}(U^2) + \sqrt{1 - \rho^2}\, \mathrm{E}(UV) \qquad (\text{linearity of expectation}) \\
&= \rho \qquad (\text{since } \mathrm{E}(U^2) = 1 \text{and } \mathrm{E}(UV) = \mathrm{E}(U)\,\mathrm{E}(V) = 0)
\end{aligned}$$

hence $\mathrm{Corr}(X, Y) = \rho$ since $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$.

3. The transformation here is $x = u$ and $y = \rho u + \sqrt{1 - \rho^2} v$. So the inverse transformation is:

$$u = x \quad \text{and} \quad (1 - \rho^2)^{-1/2}(y - \rho x).$$

We can readily see that the Jacobian of the inverse is $(1 - \rho^2)^{-1/2}$. Applying the change of variables formula, we have:

$$\begin{aligned}
f_{X,Y}(x, y) &= f_{U,V}(x, (1 - \rho^2)^{-1/2}(y - \rho x))(1 - \rho^2)^{-1/2} \\
&= \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left(-\frac{1}{2}\left(x^2 + \frac{(y - \rho x)^2}{1 - \rho^2}\right)\right) \\
&= \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left(-\frac{(x^2 - 2\rho x y + y^2)}{2(1 - \rho^2)}\right).
\end{aligned}$$

4. Working directly from the definition of the joint moment generating function:

$$\begin{aligned}
M_{X,Y}(s,t) &= \mathrm{E}(\exp(sX+tY)) \\
&= \mathrm{E}(\exp(sU+t(\rho U+\sqrt{1-\rho^2}V))) \\
&= \mathrm{E}(\exp((s+t\rho)U)\exp(t\sqrt{1-\rho^2}V)) \\
&= M_U(s+t\rho)\,M_V(t\sqrt{1-\rho^2}) \\
&= \exp\left(\frac{(s+t\rho)^2}{2}+\frac{t^2(1-\rho^2)}{2}\right) \\
&= \exp\left(\frac{s^2+2st\rho+t^2\rho^2+t^2-t^2\rho^2}{2}\right) \\
&= \exp\left(\frac{s^2+2\rho st+t^2}{2}\right).
\end{aligned}$$

20. Since U and V are independent, we have:

$$f_{U,V}(u,v) = f_U(u)\,f_V(v) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-u^2/2}\,\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-v^2/2} = \frac{1}{2\pi}\,\mathrm{e}^{-(u^2+v^2)/2}$$

and:

$$M_{U,V}(s,t) = M_U(s)\,M_V(t) = \mathrm{e}^{s^2/2}\mathrm{e}^{t^2/2} = \mathrm{e}^{(s^2+t^2)/2}.$$

B.4 Chapter 5 – Conditional distributions

1. Define A to be the event that the first throw is a head. Applying the definition of conditional probability, we have:

$$P(X=x\,|\,A) = 2\times P(X=x \text{ and first throw is a head}) = \begin{cases} 1/4 & \text{for } x=1 \\ 1/2 & \text{for } x=2 \\ 1/4 & \text{for } x=3 \\ 0 & \text{otherwise} \end{cases}$$

and:

$$P(Y=y\,|\,A) = \begin{cases} 1/2 & \text{for } y=1 \\ 1/2 & \text{for } y=0. \end{cases}$$

2. We note that $f_{Y|X}$ is a positive, real-valued function and:

$$\int_{-\infty}^{\infty} f_{Y|X}(y\,|\,x)\,\mathrm{d}y = \int_{-\infty}^{\infty}\frac{f_{X,Y}(x,y)}{f_X(x)}\,\mathrm{d}y = \frac{f_X(x)}{f_X(x)} = 1.$$

3. Since $X\sim N(\mu_X,\sigma_X^2)$, $Y\sim N(\mu_Y,\sigma_Y^2)$ and $\mathrm{Corr}(X,Y)=\rho$, we have:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_X\sigma_Y}\exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2}-\frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right.$$

for $x, y \in \mathbb{R}$ and:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right) \quad \text{for } x \in \mathbb{R}.$$

Hence:

$$f_{Y|X}(y \,|\, x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \quad \Rightarrow \quad Y \,|\, X = x \sim N\left(\mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x-\mu_X), \sigma_Y^2(1-\rho^2)\right).$$

4. The marginal density functions are obtained by integration, i.e. we have:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \int_x^1 2\,\mathrm{d}y = 2(1-x) \quad \text{for } 0 \leq x \leq 1$$

and 0 otherwise, also:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x = \int_0^y 2\,\mathrm{d}x = 2y \quad \text{for } 0 \leq y \leq 1$$

and 0 otherwise. Hence:

$$f_{Y|X}(y \,|\, x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \begin{cases} 1/(1-x) & \text{for } x \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and:

$$f_{X|Y}(x \,|\, y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \begin{cases} 1/y & \text{for } 0 \leq x \leq y \\ 0 & \text{otherwise.} \end{cases}$$

Note that $Y \,|\, X = x \sim \text{Uniform}[x, 1]$ and $X \,|\, Y = y \sim \text{Uniform}[0, y]$.

5. Here we condition on a gamma-distributed random variable, which is continuous and whose support is the positive real line. Hence we need to integrate. We have:

$$\begin{aligned} P(Y = y) = p_Y(y) &= \int_{-\infty}^{\infty} p_{Y|X}(y \,|\, x)\, f_X(x)\,\mathrm{d}x \\ &= \int_0^{\infty} \frac{x^y \mathrm{e}^{-x}}{y!} \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} \mathrm{e}^{-\beta x}\,\mathrm{d}x \\ &= \frac{\beta^\alpha}{y!\,\Gamma(\alpha)} \int_0^{\infty} x^{y+\alpha-1} \mathrm{e}^{-(\beta+1)x}\,\mathrm{d}x \\ &= \frac{\beta^\alpha}{y!\,\Gamma(\alpha)} \frac{\Gamma(y+\alpha)}{(1+\beta)^{y+\alpha}} \int_0^{\infty} \frac{1}{\Gamma(y+\alpha)} (1+\beta)^{y+\alpha} x^{y+\alpha-1} \mathrm{e}^{-(1+\beta)x}\,\mathrm{d}x \\ &= \frac{\Gamma(y+\alpha)}{y!\,\Gamma(\alpha)} \frac{\beta^\alpha}{(1+\beta)^{y+\alpha}} \quad \text{for } y = 0, 1, 2, \ldots \end{aligned}$$

and 0 otherwise. Note that the integrand is a Gamma$(y+\alpha, 1+\beta)$ density function.

6. We have:

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{Y|X}(y \mid x)\, f_X(x)\, \mathrm{d}x \\ &= \int_0^{\infty} x\, \mathrm{e}^{-xy}\, \frac{1}{\Gamma(\alpha)}\, \beta^{\alpha} x^{\alpha-1} \mathrm{e}^{-\beta x}\, \mathrm{d}x \\ &= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \int_0^{\infty} x^{\alpha} \mathrm{e}^{-(\beta+y)x}\, \mathrm{d}x \\ &= \frac{\beta^{\alpha}}{\Gamma(\alpha)}\, \frac{\Gamma(\alpha+1)}{(\beta+y)^{\alpha+1}} \int_0^{\infty} \frac{1}{\Gamma(\alpha+1)}\, (\beta+y)^{\alpha+1} x^{\alpha} \mathrm{e}^{-(\beta+y)x}\, \mathrm{d}x \\ &= \frac{\alpha}{\beta(1+y/\beta)^{\alpha+1}} \quad \text{for } y \geq 0 \end{aligned}$$

and 0 otherwise. Note that the integrand is a Gamma$(\alpha+1, \beta+y)$ density function.

7. Since $\mathrm{E}(Y \mid X)$ is a function of X, it is constant for a given X. We have:

$$\begin{aligned} \mathrm{Var}(Y \mid X) &= \mathrm{E}((Y - \mathrm{E}(Y \mid X))^2 \mid X) \\ &= \mathrm{E}((Y^2 - 2Y\, \mathrm{E}(Y \mid X) + (\mathrm{E}(Y \mid X))^2) \mid X) \\ &= \mathrm{E}(Y^2 \mid X) - 2\, \mathrm{E}(Y \mid X)\mathrm{E}(Y \mid X) + (\mathrm{E}(Y \mid X))^2 \\ &= \mathrm{E}(Y^2 \mid X) - (\mathrm{E}(Y \mid X))^2. \end{aligned}$$

8. (a) We have:

$$\mathrm{E}(M_{Y|X}(u \mid X)) = \mathrm{E}(\mathrm{E}(\mathrm{e}^{uY} \mid X)) = \mathrm{E}(\mathrm{e}^{uY}) = M_Y(u).$$

(b) We have:

$$\mathrm{E}(\mathrm{e}^{tX} M_{Y|X}(u \mid X)) = \mathrm{E}(\mathrm{e}^{tX} \mathrm{E}(\mathrm{e}^{uY} \mid X)) = \mathrm{E}(\mathrm{e}^{tX+uY}) = M_{X,Y}(t, u).$$

9. From Activity 5.4 we have:

$$f_X(x) = \begin{cases} 2(1-x) & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and:

$$f_{Y|X}(y \mid x) = \begin{cases} 1/(1-x) & \text{for } x \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, \mathrm{d}x = \int_0^1 (2x - 2x^2)\, \mathrm{d}x = \left[x^2 - \frac{2x^3}{3} \right]_0^1 = \frac{1}{3}$$

and:

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2\, f_X(x)\, \mathrm{d}x = \int_0^1 (2x^2 - 2x^3)\, \mathrm{d}x = \left[\frac{2x^3}{3} - \frac{x^4}{2} \right]_0^1 = \frac{1}{6}.$$

Hence:

$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y \mid X)) = \mathrm{E}\left(\frac{X}{2} + \frac{1}{2}\right) = \frac{1}{6} + \frac{1}{2} = \frac{2}{3}.$$

Also:

$$\mathrm{E}(XY) = \mathrm{E}(\mathrm{E}(XY \mid X)) = \mathrm{E}(X\,\mathrm{E}(Y \mid X)) = \mathrm{E}\left(\frac{X^2}{2} + \frac{X}{2}\right) = \frac{\mathrm{E}(X^2)}{2} + \frac{\mathrm{E}(X)}{2} = \frac{1}{4}.$$

Finally:

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = \frac{1}{4} - \frac{1}{3} \times \frac{2}{3} = \frac{1}{36}.$$

10. Using the moment generating function of a Poisson distribution, we have:

$$M_Y(t) = \mathrm{E}(M_{Y|\Lambda}(t \mid \Lambda)) = \mathrm{E}(\exp(\Lambda(\mathrm{e}^t - 1))) = M_\Lambda(\mathrm{e}^t - 1).$$

11. Using the result in Activity 5.5, since $Y \mid X = x \sim \mathrm{Pois}(x)$ and $X \sim \mathrm{Gamma}(\alpha, \beta)$, we have:

$$M_Y(t) = M_X(\mathrm{e}^t - 1) = \left(\frac{\beta}{\beta - (\mathrm{e}^t - 1)}\right)^\alpha = \left(1 - \frac{1}{\beta} - \frac{\mathrm{e}^t}{\beta}\right)^{-\alpha}.$$

12. Let $V \sim \mathrm{Bernoulli}(\pi)$ be the random variable to denote investment in A ($V = 1$) or B ($V = 0$). Therefore:

$$p_Z(z) = \sum_v f_{Z|V}(z \mid v)\, p_V(v) = \pi\, f_X(z) + (1 - \pi)\, f_Y(z).$$

Hence:

$$\mathrm{E}(Z) = \int_{-\infty}^{\infty} z\, f_Z(z)\, \mathrm{d}z = \pi\,\mathrm{E}(X) + (1 - \pi)\,\mathrm{E}(Y)$$

and:

$$\begin{aligned}\mathrm{Var}(Z) &= \mathrm{E}(Z^2) - (\mathrm{E}(Z))^2 \\ &= \pi\,\mathrm{E}(X^2) + (1 - \pi)\,\mathrm{E}(Y^2) - (\pi\,\mathrm{E}(X) + (1 - \pi)\,\mathrm{E}(Y))^2 \\ &= \pi\,\mathrm{Var}(X) + (1 - \pi)\,\mathrm{Var}(Y) + \pi(1 - \pi)\,\mathrm{E}((X - Y)^2).\end{aligned}$$

13. Let N be the number of basin hurricanes in a year, and let $X_i \sim \mathrm{Bernoulli}(\pi)$ take the value 1 if the ith hurricane makes landfall, and 0 otherwise. Hence the number of hurricanes making landfall is the random sum $\sum_{i=1}^{N} X_i$.

14. (a) We have:

$$\mathrm{E}(S) = \mathrm{E}(N)\,\mathrm{E}(X) = \lambda\,\mathrm{E}(X).$$

(b) We have:

$$\mathrm{Var}(S) = \mathrm{E}(N)\,\mathrm{Var}(X) + (\mathrm{E}(X))^2\mathrm{Var}(N) = \lambda\,\mathrm{Var}(X) + \lambda(\mathrm{E}(X))^2 = \lambda\,\mathrm{E}(X^2).$$

(c) We have:

$$K_Y(t) = \lambda(\mathrm{e}^{K_X(t)} - 1) = \lambda(M_X(t) - 1) = \lambda \sum_{i=1}^{\infty} \frac{t^i \mathrm{E}(X^i)}{i!}.$$

The kth cumulant is the coefficient of $t^i/i!$. Hence the kth cumulant is $\lambda\,\mathrm{E}(X^k)$.

Appendix C
Solutions to Sample examination questions

C.1 Chapter 2 – Probability space

1. (a) We have:

$$P(A \cap (B^c \cup C)) = P(A \,|\, B^c \cup C)\, P(B^c \cup C).$$

However:

$$\begin{aligned} P(B^c \cup C) &= 1 - P(B \cap C^c) \\ &= 1 - P(B \,|\, C^c)\, P(C^c) \\ &= 1 - P(B \,|\, C^c)\,(1 - P(C)) \\ &= 1 - 0.3 \times (1 - 0.1) = 0.73. \end{aligned}$$

So $P(A \cap (B^c \cup C)) = 0.4 \times 0.73 = 0.292$.

(b) We have:

$$\begin{aligned} P(A_2 \cup A_3 \,|\, A_1) &= \frac{P((A_2 \cup A_3) \cap A_1)}{P(A_1)} \\ &= \frac{P((A_2 \cap A_1) \cup (A_3 \cap A_1))}{P(A_1)} \\ &= \frac{P(A_2 \cap A_1) + P(A_3 \cap A_1) - P(A_2 \cap A_1 \cap A_3)}{P(A_1)} \\ &= P(A_2 \,|\, A_1) + P(A_3 \,|\, A_1) - P(A_2 \cap A_3 \,|\, A_1). \end{aligned}$$

(c) We have:

$$P(C \cap A) = 0.2 \times P(A)$$

so:

$$\begin{aligned} P(A \cup C) &= P(A) + P(C) - P(A \cap C) \\ &= P(A) + 0.1 - 0.2 \times P(A) \\ &= 0.8 \times P(A) + 0.1. \end{aligned}$$

Hence $0.5 = 0.8 \times P(A) + 0.1$, implying $P(A) = 0.5$. At the same time:

$$\begin{aligned} P(B \cap C^c \,|\, A) &= 1 - P(B^c \cup C \,|\, A) \\ &= 1 - \frac{P(A \cap (B \cup C^c))}{P(A)} \\ &= 1 - \frac{0.292}{0.5} \\ &= 0.416. \end{aligned}$$

Hence using (b) we have:

$$\begin{aligned} P(B \cup C^c \,|\, A) &= P(B \,|\, A) + P(C^c \,|\, A) - P(B \cap C^c \,|\, A) \\ &= 0.4 + (1 - 0.2) - 0.416 \\ &= 0.784. \end{aligned}$$

2. (a) We have $P(A \cup A^c) = P(A) + P(A^c)$, since A and A^c are mutually exclusive. However, $P(A \cup A^c) = P(\Omega) = 1$, since A and A^c are collectively exhaustive. Hence:

$$P(A) + P(A^c) = 1 \quad \Rightarrow \quad P(A^c) = 1 - P(A).$$

(b) Since $B = (A \cap B) \cup (A^c \cap B)$, which is a partition of B, we have:

$$P(B) = P((A \cap B) \cup (A^c \cap B)) = P(A \cap B) + P(A^c \cap B)$$

hence:

$$P(A^c \cap B) = P(B) - P(A \cap B).$$

(c) Since $A \cup B = A \cup (A^c \cap B)$, and using (b), we have:

$$\begin{aligned} P(A \cup B) &= P(A \cup (A^c \cap B)) \\ &= P(A) + P(A^c \cap B) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$

(d) Using (c), we have:

$$\begin{aligned} P((A \cup B) \cap (A^c \cup B^c)) &= P(A \cup B) + P(A^c \cup B^c) - P(A \cup B \cup A^c \cup B^c) \\ &= P(A \cup B) + P((A \cap B)^c) - P(\Omega) \\ &= P(A) + P(B) - P(A \cap B) + (1 - P(A \cap B)) - 1 \\ &= P(A) + P(B) - 2 \times P(A \cap B). \end{aligned}$$

The event $(A \cup B) \cap (A^c \cup B^c)$ is the event corresponding to either A or B, but not both.

3. We have:

$$\begin{aligned}\pi_n &= P(n\text{th toss is heads and there are two heads in the first } n-1 \text{ tosses})\\ &= 0.5 \times \frac{(n-1)!}{2!\,(n-3)!} \times (0.5)^2 \times (1-0.5)^{n-3}\\ &= (0.5)^n \times \frac{(n-1)(n-2)}{2}\\ &= (n-1)(n-2)2^{-n-1}.\end{aligned}$$

Let N be the toss number when the third head occurs in the repeated tossing of a fair coin. Therefore:

$$1 = P(N < \infty) = \sum_{n=3}^{\infty} P(N = n) = \sum_{n=3}^{\infty} \pi_n = \sum_{n=3}^{\infty} (n-1)(n-2)2^{-n-1}$$

and so:

$$\sum_{n=3}^{\infty} (n-1)(n-2)2^{-n} = 2.$$

C.2 Chapter 3 – Random variables and univariate distributions

1. (a) We must have:

$$1 = \sum_{x=1}^{\infty} k^x = \frac{k}{1-k}.$$

Solving, $k = 1/2$.

(b) For $x = 1, 2, \ldots$, we have:

$$F_X(x) = P(X \le x) = \sum_{i=1}^{x} \left(\frac{1}{2}\right)^i = \frac{(1/2)(1-(1/2)^x)}{1-1/2} = 1 - \left(\frac{1}{2}\right)^x.$$

(c) We have:

$$M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \sum_{x=1}^{\infty} k^x \mathrm{e}^{tx} = \sum_{x=1}^{\infty} (k\mathrm{e}^t)^x = \frac{k\mathrm{e}^t}{1-k\mathrm{e}^t} = \frac{\mathrm{e}^t}{2-\mathrm{e}^t}.$$

For the above to be valid, the sum to infinity has to be valid. That is, $k\mathrm{e}^t < 1$, meaning $t < \log 2$. We then have:

$$M'_X(t) = \frac{2\mathrm{e}^t}{(2-\mathrm{e}^t)^2}$$

so that $\mathrm{E}(X) = M'_X(0) = 2$.

2. (a) Let $I(A)$ be the indicator function equal to 1 under A, and 0 otherwise. For any $a > 0$, we have:

$$P(X \geq a) = \mathrm{E}(I(X \geq a)) \leq \mathrm{E}\left(\frac{I(X \geq a)X}{a}\right) \leq \frac{\mathrm{E}(X)}{a}.$$

(b) Substitute $Y = (X - \mathrm{E}(X))^2$ in (a) and replacing a by a^2, we have:

$$P(|X - \mathrm{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

(c) Let $X \sim \mathrm{Exp}(1)$, with mean and variance both equal to 1. Hence, for $a > 0$, we have:

$$P(X > a) = \int_a^\infty \mathrm{e}^{-x}\,\mathrm{d}x = \mathrm{e}^{-a}.$$

So, by the Markov inequality, $\mathrm{e}^{-a} \leq \mathrm{E}(X)/a = 1/a$, implying $a\mathrm{e}^{-a} \leq 1$. At the same time, for $0 < a < 1$, we have:

$$\begin{aligned} P(|X - 1| > a) &= P(X > 1 + a) + P(X < 1 - a) \\ &= \int_{1+a}^\infty \mathrm{e}^{-x}\,\mathrm{d}x + \int_0^{1-a} \mathrm{e}^{-x}\,\mathrm{d}x \\ &= e^{-(1+a)} + 1 - \mathrm{e}^{-(1-a)}. \end{aligned}$$

Hence, using (b), we have:

$$\mathrm{e}^{-(1+a)} + 1 - \mathrm{e}^{-(1-a)} \leq \frac{\mathrm{Var}(X)}{a^2} = \frac{1}{a^2}$$

implying the other inequality.

3. (a) For $0 < w < 1$, we have:

$$P(W_1 \leq w) = P(-\sqrt{w} < X_1 < \sqrt{w}) = \int_{-\sqrt{w}}^{\sqrt{w}} \frac{1}{2}\,\mathrm{d}x = \sqrt{w}.$$

Hence the cumulative distribution function of W_1 is $F_{W_1}(w) = \sqrt{w}$, for $0 < w < 1$, and so:

$$f_{W_1}(w) = F'_{W_1}(w) = \begin{cases} \dfrac{1}{2\sqrt{w}} & \text{for } 0 < w < 1 \\ 0 & \text{otherwise.} \end{cases}$$

(b) The range of W_2 is $[0, 4]$. For $0 < w < 1$, we have:

$$F_{W_2}(w) = P(W_2 < w) = P(-\sqrt{w} < X_2 < \sqrt{w}) = \int_{-\sqrt{w}}^{\sqrt{w}} \frac{1}{3}\,\mathrm{d}x = \frac{2\sqrt{w}}{3}.$$

For $1 \leq w < 4$, X_2 is in the range $[-2, -1]$. Hence for $1 \leq w < 4$ we have:

$$F_{W_2}(w) - F_{W_2}(1) = P(1 \leq W_2 < w) = P(-\sqrt{w} < X_2 < -1) = \int_{-\sqrt{w}}^{-1} \frac{1}{3}\,\mathrm{d}x = \frac{-1 + \sqrt{w}}{3}.$$

Hence differentiating with respect to w, we get:

$$f_{W_2}(w) = \begin{cases} 1/(3\sqrt{w}) & \text{for } 0 < w < 1 \\ 1/(6\sqrt{w}) & \text{for } 1 \leq w < 4 \\ 0 & \text{otherwise.} \end{cases}$$

(c) For $0 < y < 2$, we have:

$$f_Y(y) = f_{W_2}(y^2)\left|\frac{\mathrm{d}}{\mathrm{d}y}\,y^2\right| = \begin{cases} 2y/3y & \text{for } 0 < y^2 < 1 \\ 2y/6y & \text{for } 1 \leq y^2 < 4 \\ 0 & \text{otherwise.} \end{cases}$$

Hence:

$$f_Y(y) = \begin{cases} 2/3 & \text{for } 0 < y < 1 \\ 1/3 & \text{for } 1 \leq y < 2 \\ 0 & \text{otherwise.} \end{cases}$$

C.3 Chapter 4 – Multivariate distributions

1. (a) We must have:

$$1 = \int_0^2 \int_0^2 a(x+y)\,\mathrm{d}x\,\mathrm{d}y = a\int_0^2 \left[\frac{x^2}{2} + xy\right]_0^2 \mathrm{d}y = a\int_0^2 (2+2y)\,\mathrm{d}y = a[2y+y^2]_0^2 = 8a$$

so that $a = 1/8$.

(b) For $0 < x < 2$, we have:

$$F_X(x) = \frac{1}{8}\int_0^2 \int_0^x (x'+y)\,\mathrm{d}x'\,\mathrm{d}y = \frac{1}{8}\int_0^2 \left(\frac{x^2}{2} + xy\right)\mathrm{d}y = \frac{2x+x^2}{8}.$$

The mean is:

$$\begin{aligned} \mathrm{E}(X) = \frac{1}{8}\int_0^2 \int_0^2 (x^2+xy)\,\mathrm{d}x\,\mathrm{d}y &= \frac{1}{8}\int_0^2 \left[\frac{x^3}{3} + \frac{x^2y}{2}\right]_0^2 \mathrm{d}y \\ &= \frac{1}{8}\int_0^2 \left(\frac{8}{3} + 2y\right)\mathrm{d}y \\ &= \frac{1}{8}\left[\frac{8y}{3} + y^2\right]_0^2 \\ &= \frac{7}{6}. \end{aligned}$$

2. (a) Integrating out x first, we have:

$$1 = \int_0^2 \int_0^y axy^2\,\mathrm{d}x\,\mathrm{d}y = \int_0^2 a\frac{y^4}{2}\,\mathrm{d}y = \frac{2^5}{10} \times a.$$

Hence $a = 5/16$.

(b) We have:

$$Y - X = \frac{U}{V} \quad \text{and} \quad Y + X = U$$

so that:

$$Y = \frac{U}{2}(1 + V^{-1}) \quad \text{and} \quad X = \frac{U}{2}(1 - V^{-1}).$$

With $0 < X < Y < 2$, we have:

$$0 < \frac{U}{2}(1 - V^{-1}) < \frac{U}{2}(1 + V^{-1}) < 2.$$

Solving for the first inequality, we have $U(1 - V^{-1}) > 0$, so that $V > 1$ since $U > 0$. The second inequality is always true. Solving for the third one, we have:

$$U < \frac{4}{1 + V^{-1}} = \frac{4V}{1 + V}.$$

Hence the valid region is $v > 1$ and $0 < u < 4v/(1 + v)$. For the density, we have:

$$\begin{aligned} f_{U,V}(u, v) &= f_{X,Y}(u(1 - v^{-1})/2, u(1 + v^{-1})/2) \left|\begin{matrix} (1 - v^{-1})/2 & uv^{-2}/2 \\ (1 + v^{-1})/2 & -uv^{-2}/2 \end{matrix}\right| \\ &= \frac{5}{16} \times \Big(\frac{u(v - 1)}{2v}\Big)\Big(\frac{u(v + 1)}{2v}\Big)^2 \frac{u}{2v^2} \\ &= \frac{5u^4}{256v^5}(v - 1)(v + 1)^2, \quad 0 < u < \frac{4v}{1 + v}, \quad v > 1. \end{aligned}$$

(c) First note that:

$$\sqrt{\frac{y}{y - x}} = \sqrt{\frac{u(1 + v^{-1})/2}{u(1 + v^{-1})/2 - u(1 - v^{-1})/2}} = \sqrt{\frac{v + 1}{2}}.$$

Hence:

$$\begin{aligned} \mathrm{E}\left(\sqrt{\frac{Y}{Y - X}}\right) = \frac{1}{\sqrt{2}}\,\mathrm{E}(\sqrt{v + 1}) &= \frac{1}{\sqrt{2}} \int_1^\infty \frac{4(v - 1)}{(v + 1)^{5/2}}\,\mathrm{d}v \\ &= \frac{1}{\sqrt{2}} \int_2^\infty 4(w - 2)w^{-5/2}\,\mathrm{d}w \\ &= \frac{4}{\sqrt{2}}\left[-2w^{-1/2} + \frac{4w^{-3/2}}{3}\right]_2^\infty \\ &= \frac{4}{\sqrt{2}}\left(\frac{2}{\sqrt{2}} - \frac{4}{6\sqrt{2}}\right) \\ &= \frac{8}{3}. \end{aligned}$$

3. (a) Solving for X and Y, we have:

$$X = UV \quad \text{and} \quad Y = V(1 - U).$$

Hence:

$$f_{U,V}(u,v) = f_{X,Y}(uv, v(1-u)) \left|\left| \begin{matrix} v & u \\ -v & 1-u \end{matrix} \right|\right| = \frac{auv^2(1-u)}{v^{3/2}}|v| = au(1-u)v^{3/2}.$$

For the support of the density, solving $x < y$ implies $uv < v(1-u)$, so that:

$$u < \frac{1}{2}$$

while solving $y < 2$ implies $v(1-u) < 2$, so that:

$$v < \frac{2}{1-u}.$$

Finally, the definitions of U and V imply that they are positive. Hence the region is $0 < v < 2/(1-u)$, $0 < u < 1/2$.

(b) Substituting $q = 1 - u$ (starting from the fourth line), we have:

$$\begin{aligned} 1 = a\int_0^{1/2}\int_0^{2/(1-u)} u(1-u)v^{3/2}\,\mathrm{d}v\,\mathrm{d}u &= a\int_0^{1/2} \frac{2}{5}u(1-u)2^{5/2}(1-u)^{-5/2}\,\mathrm{d}u \\ &= \frac{8a\sqrt{2}}{5}\int_0^{1/2} u(1-u)^{-3/2}\,\mathrm{d}u \\ &= \frac{8a\sqrt{2}}{5}\int_{1/2}^{1} (1-q)q^{-3/2}\,\mathrm{d}q \\ &= \frac{8a\sqrt{2}}{5}[-2q^{-1/2} - 2q^{1/2}]_{1/2}^{1} \\ &= \frac{8a\sqrt{2}}{5}(-4 + 2\sqrt{2} + \sqrt{2}) \\ &= \frac{a(48 - 32\sqrt{2})}{5}. \end{aligned}$$

Hence:

$$a = \frac{5}{48 - 32\sqrt{2}}.$$

(c) We have:

$$\begin{aligned}
\mathrm{E}(X) = \mathrm{E}(UV) &= a \int_0^{1/2} \int_0^{2/(1-u)} u^2(1-u)v^{5/2} \,\mathrm{d}v \,\mathrm{d}u \\
&= a \int_0^{1/2} \frac{2}{7}\, u^2(1-u)2^{7/2}(1-u)^{-7/2} \,\mathrm{d}u \\
&= \frac{16a\sqrt{2}}{7} \int_0^{1/2} u^2(1-u)^{-5/2} \,\mathrm{d}u \\
&= \frac{16a\sqrt{2}}{7} \int_{1/2}^{1} (1-2q+q^2)q^{-5/2} \,\mathrm{d}q \\
&= \frac{16a\sqrt{2}}{7} \left[-\frac{2q^{-3/2}}{3} + 4q^{-1/2} + 2q^{1/2} \right]_{1/2}^{1} \\
&= \frac{16a\sqrt{2}}{7} \left(-\frac{2}{3} + 6 + \frac{4\sqrt{2}}{3} - 4\sqrt{2} - \sqrt{2} \right) \\
&= \frac{16a\sqrt{2}(16 - 11\sqrt{2})}{21} \\
&= \frac{80(16\sqrt{2} - 22)}{21(48 - 32\sqrt{2})}.
\end{aligned}$$

C.4 Chapter 5 – Conditional distributions

1. (a) We start from the conditional density function and then compute conditional moments and use the law of iterated expectations. The conditional density function is:

$$f_{Y|X}(y \,|\, x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \begin{cases} 2y/x^2 & \text{for } 0 < y < x \\ 0 & \text{otherwise.} \end{cases}$$

Hence:

$$\mathrm{E}(Y \,|\, X = x) = \int_{-\infty}^{\infty} y\, f_{Y|X}(y \,|\, x) \,\mathrm{d}y = \int_0^x \frac{2y^2}{x^2} \,\mathrm{d}y = \left[\frac{2y^3}{3x^2} \right]_0^x = \frac{2}{3}\, x.$$

Therefore:

$$\mathrm{E}(Y \,|\, X) = \frac{2}{3}\, X.$$

The unconditional expectation of Y is then:

$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y \,|\, X)) = \frac{2}{3}\, \mathrm{E}(X) = \frac{8}{15}.$$

We have that:

$$\mathrm{E}(XY) = \mathrm{E}(\mathrm{E}(XY \,|\, X)) = \mathrm{E}(X\, \mathrm{E}(Y \,|\, X)) = \frac{2}{3}\, \mathrm{E}(X^2)$$

where:

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x = \int_0^1 4x^5\,\mathrm{d}x = \left[\frac{2}{3}x^6\right]_0^1 = \frac{2}{3}.$$

Therefore:

$$\mathrm{E}(XY) = \frac{2}{3} \times \frac{2}{3} = \frac{4}{9}.$$

(b) It is easier to work from general to specific conditioning. We have:

$$\mathrm{Var}(Y \mid X = x) = \mathrm{E}(Y^2 \mid X = x) - (\mathrm{E}(Y \mid X = x))^2$$

where:

$$\mathrm{E}(Y^2 \mid X = x) = \int_{-\infty}^{\infty} y^2 f_{Y|X}(y \mid x)\,\mathrm{d}y = \int_0^x \frac{2y^3}{x^2}\,\mathrm{d}y = \left[\frac{y^4}{2x^2}\right]_0^x = \frac{x^2}{2}.$$

Therefore:

$$\mathrm{Var}(Y \mid X = x) = \frac{1}{2}\,x^2 - \frac{4}{9}\,x^2 = \frac{1}{18}\,x^2.$$

Hence:

$$\mathrm{Var}(Y \mid X = 0.5) = \frac{1}{18} \times \frac{1}{4} = \frac{1}{72}.$$

2. (a) If X is fixed, we have a fixed number of coin tosses, so $Y \mid X = x \sim \mathrm{Bin}(x, 0.5)$. Note that $\mathrm{E}(X) = 3.5$. Using properties of the binomial distribution, we have:

$$\mathrm{E}(Y \mid X = x) = 0.5x \quad \Rightarrow \quad \mathrm{E}(Y \mid X) = 0.5X.$$

Applying the law of iterated expectations:

$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y \mid X)) = 0.5\,\mathrm{E}(X) = 0.5 \times 3.5 = 1.75.$$

(b) From the binomial probability mass function:

$$P(Y = 5 \mid X = x) = \binom{x}{5}\left(\frac{1}{2}\right)^6$$

for $x = 5$ or 6. By the total probability formula:

$$P(Y = 5) = \sum_{x=5}^{6} P(Y = 5 \mid X = x)\,P(X = x) = \sum_{x=5}^{6} \binom{x}{5}\left(\frac{1}{2}\right)^6 \frac{1}{6} = 0.0182.$$

3. Setting $X = \sigma^2$ we have that $Y \mid X \sim N(0, X)$. For the mean of Y we note that:

$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y \mid X)) = \mathrm{E}(0) = 0.$$

For the variance of Y we note that:

$$\mathrm{Var}(Y) = \mathrm{E}(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(\mathrm{E}(Y \mid X)) = \mathrm{E}(X) + \mathrm{Var}(0) = \frac{1}{\lambda}.$$

The mgf of a $N(0, \sigma^2)$-distributed random variable is $\exp(\sigma^2 t^2/2)$, hence:

$$M_{Y|X}(t \mid X) = \exp\left(\frac{t^2}{2}X\right).$$

It is also known that $M_Y(t) = \mathrm{E}(M_{Y|X}(t \,|\, X))$. Hence, since $X \sim \mathrm{Exp}(\lambda)$, then:

$$M_Y(t) = \mathrm{E}\left(\exp\left(\frac{t^2}{2}X\right)\right) = M_X\left(\frac{t^2}{2}\right) = \frac{\lambda}{\lambda - t^2/2} = \left(1 - \frac{t^2}{2\lambda}\right)^{-1}$$

provided $|t| < \sqrt{2\lambda}$. Therefore, the cumulant generating function is:

$$K_Y(t) = -\log\left(1 - \frac{t^2}{2\lambda}\right).$$

Appendix D
Sample examination paper

Important note: This Sample examination paper reflects the examination and assessment arrangements for this course in the academic year 2020–21. The format and structure of the examination may have changed since the publication of this subject guide. You can find the most recent examination papers on the VLE where all changes to the format of the examination are posted.

Time allowed: 2 hours.

Candidates should answer all **FOUR** questions: **QUESTION 1** of Section A (40 marks) and all **THREE** questions from Section B (60 marks in total). **Candidates are strongly advised to divide their time accordingly.**

A handheld calculator may be used when answering questions on this paper and it must comply in all respects with the specification given with your Admission Notice. The make and type of machine must be clearly stated on the front cover of the answer book.

SECTION A

Answer all **three** parts of question 1 (40 marks in total).

1. (a) The joint probability density function of (X, Y) is:

$$f_{X,Y}(x,y) = \begin{cases} k(x+y) & \text{for } 0 < x < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

i. Find k.

(5 marks)

ii. Find $f_X(x)$ and $f_Y(y)$, the marginal density function of X and Y, respectively. Are X and Y independent?

(8 marks)

(b) Let $N \sim \text{Poisson}(\lambda)$, and $X \mid N \sim \text{Binomial}(N, p)$, where $\lambda > 0$ and $0 < p < 1$.

i. Show that the joint probability mass function of X and N is:

$$p_{X,N}(x,n) = \frac{\mathrm{e}^{-\lambda}}{x!}\left(\frac{p}{q}\right)^x \frac{(\lambda q)^n}{(n-x)!} \quad \text{for } x = 0, 1, 2, \ldots, n,\ n = 0, 1, 2, \ldots$$

where $q = 1 - p$.

(4 marks)

ii. Using the identity:

$$\mathrm{e}^a = \sum_{i \geq 0} \frac{a^i}{i!} \quad \text{for any } a \in \mathbb{R}$$

show that the marginal probability mass function of X is:

$$p_X(x) = \frac{\mathrm{e}^{-\lambda p}(\mu p)^x}{x!} \quad \text{for } x = 0, 1, 2, \ldots.$$

What is the distribution of X?

(9 marks)

(c) Let Y be a positive random variable with an absolutely continuous probability density $f_Y(y)$ and finite mean μ.

i. Prove the Markov inequality:

$$P(Y > a) \leq \frac{\mu}{a}$$

where $a > 0$.

(4 marks)

ii. Let $X \sim \text{Exp}(1)$. Show that the moment generating function of X is:

$$M_X(t) = \frac{1}{1-t}$$

stating clearly the range of t where this is valid.

(4 marks)

iii. Using i. and ii. by putting $Y = \mathrm{e}^{tX}$, show that:

$$P\left(X > \frac{\log a}{t}\right) \leq \frac{1}{a(1-t)}$$

where the range of t should be stated again.

By putting $a = \mathrm{e}^t$ and calculating $P(X > 1)$, show that:

$$x\mathrm{e}^{-x} \leq 1$$

for any $x > 0$.

(6 marks)

D

SECTION B

Answer all **three** questions in this section (60 marks in total).

2. Let X, Y be two random variables having the joint probability density function:

$$f_{X,Y}(x,y) = \frac{4}{3}\, y\mathrm{e}^{-(x+y)} \quad \text{for } 0 < x < y.$$

(a) Let $U = X + Y$ and $V = X - Y$. Show that the joint density of U, V is:

$$f_{U,V}(u,v) = \frac{1}{3}\,(u - v)\mathrm{e}^{-u} \quad \text{for } -u < v < 0.$$

(Hint: Solve x and y in terms of u and v, and then use $0 < x < y$ to solve for u and v, the region where the joint density is valid.)

(7 marks)

(b) Find the marginal density $f_U(u)$ of U. You are given that the marginal density of V is:

$$f_V(v) = \frac{(1 - 2v)\mathrm{e}^{v}}{3} \quad \text{for } v < 0.$$

Are U and V independent?

(4 marks)

(c) Calculate $\mathrm{E}(U \,|\, V = v)$.

(9 marks)

3. A certain product is produced by five different machines: A, B, C, D and E. The proportions of the total items produced by each machine are 0.1, 0.3, 0.2, 0.25 and 0.15, respectively. The probability that an item is faulty depends on the machine which produced it. The proportions of faulty items are 0.01, 0.2, 0.1, 0.05, and 0.02 for products produced by machines A, B, C, D, and E, respectively.

(a) What is the probability that a randomly chosen item is faulty?

(4 marks)

(b) Given an item is faulty, what is the probability that it is produced by machine B?

(4 marks)

(c) Let the selling price of an item be $S \sim N(\mu_1, \sigma^2)$ when it is produced by machines A, B or C, and $S \sim N(\mu_2, \sigma^2)$ otherwise. Given $S > c$, what is the probability that it is produced by machine A? Leave your answer in terms of $\Phi(\cdot)$, the distribution function of a standard normal random variable.

(6 marks)

(d) Find $\mathrm{E}(S)$ and $\mathrm{Var}(S)$. (Hint: For $\mathrm{Var}(S)$, find $\mathrm{E}(S^2)$ first.)

(6 marks)

D

4. For $i = 1, 2, \ldots, n$, let W_i be independent and identically distributed Bernoulli(p) random variables. Let X_i be independent and identically distributed random variables which are also independent from the W_is, with moment generating function $M_X(t)$ defined on an interval I.

   (a) Let $S_n = \sum_{i=1}^{n} W_i X_i$. By conditioning on the value of X_i, find the moment generating function of $X_i W_i$. Hence show that the moment generating function of S_n is given by:

   $$M_{S_n}(t) = (pM_X(t) + q)^n \quad \text{for } q = 1 - p,\ t \in I.$$

   (7 marks)

   (b) Let N be a random variable with moment generating function $M_N(t)$ defined on an interval J. Show that the sum:

   $$R_N = \sum_{i=1}^{N} X_i$$

   has moment generating function:

   $$M_{R_N}(t) = M_N(K_X(t)) \quad \text{for } t \in I,\ K_X(t) \in J$$

   where $K_X(t)$ is the cumulant generating function of X_i.

   (7 marks)

   (c) Show that when $N \sim \text{Binomial}(n, p)$, the distributions of R_N and S_n are the same.

   (6 marks)

END OF PAPER

Appendix E
Solutions to Sample examination paper

1. (a) i. To find k, note that:

$$\int_0^1 \int_0^y k(x+y)\,\mathrm{d}x\,\mathrm{d}y = 1.$$

Hence:

$$1 = \int_0^1 \int_0^y k(x+y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^1 k\left[\frac{x^2}{2} + yx\right]_0^y \mathrm{d}y = \int_0^1 \frac{3ky^2}{2}\,\mathrm{d}y = \left[\frac{ky^3}{2}\right]_0^1 = \frac{k}{2}$$

meaning that $k = 2$.

ii. We have:

$$f_X(x) = \int_x^1 f_{X,Y}(x,y)\,\mathrm{d}y = \int_x^1 (2x+2y)\,\mathrm{d}y = \Big[2xy + y^2\Big]_x^1 = 2x + 1 - 3x^2$$

for $0 < x < 1$, and 0 otherwise.

We have:

$$f_Y(y) = \int_0^y f_{X,Y}(x,y)\,\mathrm{d}x = \int_0^y (2x+2y)\,\mathrm{d}x = \Big[x^2 + 2xy\Big]_0^y = 3y^2$$

for $0 < y < 1$, and 0 otherwise.

Since $f_{X,Y}(x,y) \neq f_X(x)\,f_Y(y)$, X and Y are not independent.

(b) i. We have:

$$\begin{aligned} p_{X,N}(x,n) &= p_{X|N}(x \mid n)\,p_N(n) \\ &= \binom{n}{x} p^x q^{n-x}\,\frac{\lambda^n \mathrm{e}^{-\lambda}}{n!} \\ &= \frac{1}{x!\,(n-x)!}\left(\frac{p}{q}\right)^x (\lambda q)^n \mathrm{e}^{-\lambda} \\ &= \frac{\mathrm{e}^{-\lambda}}{x!}\left(\frac{p}{q}\right)^x \frac{(\lambda q)^n}{(n-x)!} \end{aligned}$$

for $x = 0, 1, 2, \ldots, n$ and $n = 0, 1, 2, \ldots$.

ii. We have:

$$\begin{aligned} p_X(x) = \sum_{n=x}^{\infty} \frac{\mathrm{e}^{-\lambda}}{x!} \left(\frac{p}{q}\right)^x \frac{(\lambda q)^n}{(n-x)!} &= \frac{\mathrm{e}^{-\lambda}}{x!} \left(\frac{p}{q}\right)^x \sum_{n=x}^{\infty} \frac{(\lambda q)^n}{(n-x)!} \\ &= \frac{\mathrm{e}^{-\lambda}}{x!} \left(\frac{p}{q}\right)^x (\lambda q)^x \sum_{n=x}^{\infty} \frac{(\lambda q)^{n-x}}{(n-x)!} \\ &= \frac{\mathrm{e}^{-\lambda}}{x!} \left(\frac{p}{q}\right)^x (\lambda q)^x \sum_{j=0}^{\infty} \frac{(\lambda q)^j}{j!} \\ &= \frac{\mathrm{e}^{-\lambda}}{x!} \left(\frac{p}{q}\right)^x (\lambda q)^x \mathrm{e}^{\lambda q} \\ &= \frac{\mathrm{e}^{-\lambda}}{x!} (\lambda p)^x \mathrm{e}^{\lambda q} \\ &= \frac{(\lambda p)^x \mathrm{e}^{-(1-q)\lambda}}{x!} \\ &= \frac{(\lambda p)^x \mathrm{e}^{-\lambda p}}{x!} \end{aligned}$$

for $x = 0, 1, 2, \ldots$, and 0 otherwise. Therefore, X has a Poisson distribution with mean λp.

(c) i. We have:

$$P(Y > a) = \int_a^{\infty} f_Y(y)\,\mathrm{d}y \leq \int_a^{\infty} \frac{y}{a} f_Y(y)\,\mathrm{d}y \leq \frac{1}{a} \int_0^{\infty} y\, f_Y(y)\,\mathrm{d}y = \frac{\mu}{a}.$$

ii. We have:

$$\begin{aligned} M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \int_0^{\infty} \mathrm{e}^{tx}\mathrm{e}^{-x}\,\mathrm{d}x &= \int_0^{\infty} \mathrm{e}^{-(1-t)x}\,\mathrm{d}x \\ &= \Big[-(1-t)^{-1}\mathrm{e}^{-(1-t)x} \Big]_0^{\infty} \\ &= (1-t)^{-1} \end{aligned}$$

for $t < 1$.

iii. Putting $Y = \mathrm{e}^{tX}$, for $t < 1$, and using part i., we have:

$$P(\mathrm{e}^{tX} > a) \leq \frac{\mathrm{E}(\mathrm{e}^{tX})}{a} = \frac{1}{a(1-t)} \quad \Rightarrow \quad P\left(X > \frac{\log a}{t}\right) \leq \frac{1}{a(1-t)}.$$

Now:

$$P(X > 1) = \int_1^{\infty} \mathrm{e}^{-x}\,\mathrm{d}x = \mathrm{e}^{-1}.$$

Hence putting $a = \mathrm{e}^t$ we have:

$$\begin{aligned} \mathrm{e}^{-1} = P(X > 1) \leq \frac{1}{\mathrm{e}^t(1-t)} \quad &\text{for } t < 1 \\ \Rightarrow \quad (1-t)\mathrm{e}^{-(1-t)} \leq 1 \quad &\text{for } 1 - t > 0 \\ \Rightarrow \quad x\mathrm{e}^{-x} \leq 1 \quad &\text{for } x > 0. \end{aligned}$$

2. (a) We have $X = (U+V)/2$ and $Y = (U-V)/2$. The Jacobian is:

$$J = \begin{pmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix}.$$

Hence the joint probability density function has the form:

$$f_{U,V}(u,v) = f_{X,Y}(x,y)\,|J| = \frac{4}{3}\,\frac{u-v}{2}\,\mathrm{e}^{-u}\,\frac{1}{2} = \frac{1}{3}\,(u-v)\mathrm{e}^{-u}.$$

To find the region where this is defined, note that:

$$0 < x < y \quad \Rightarrow \quad 0 < \frac{u+v}{2} < \frac{u-v}{2}.$$

Solving all three inequalities, we get:

$$u > -v, \quad u > v \quad \text{and} \quad v < 0$$

which can be expressed as $-u < v < 0$. Hence the joint probability density function is as stated in the question.

(b) We have:

$$f_U(u) = \int_{-u}^{0} \frac{1}{3}\,(u-v)\mathrm{e}^{-u}\,\mathrm{d}v = \frac{1}{2}\,u^2\mathrm{e}^{-u} \quad \text{for } u > 0$$

and 0 otherwise. It is clear that $f_{U,V}(u,v) \neq f_U(u)\,f_V(v)$, and hence U and V are not independent.

(c) We have:

$$\begin{aligned}
\mathrm{E}(U \,|\, V = v) &= \int_{-v}^{\infty} u\, f_{U|v}(u)\,\mathrm{d}u \\
&= \int_{-v}^{\infty} \frac{u(u-v)\mathrm{e}^{-u}/3}{\mathrm{e}^{v}(1-2v)/3}\,\mathrm{d}u \\
&= \frac{1}{\mathrm{e}^{v}(1-2v)} \int_{-v}^{\infty} (u^2 - uv)\mathrm{e}^{-u}\,\mathrm{d}u \\
&= \frac{1}{\mathrm{e}^{v}(1-2v)} \left(\Big[-(u^2-uv)\mathrm{e}^{-u} \Big]_{-v}^{\infty} + \int_{-v}^{\infty} (2u-v)\mathrm{e}^{-u}\,\mathrm{d}u \right) \\
&= \frac{1}{\mathrm{e}^{v}(1-2v)} \left(2v^2\mathrm{e}^{v} + \Big[-(2u-v)\mathrm{e}^{-u} \Big]_{-v}^{\infty} + 2\int_{-v}^{\infty} \mathrm{e}^{-u}\,\mathrm{d}u \right) \\
&= \frac{1}{\mathrm{e}^{v}(1-2v)}\,(2v^2\mathrm{e}^{v} - 3v\mathrm{e}^{v} + 2\mathrm{e}^{v}) \\
&= \frac{2v^2 - 3v + 2}{1-2v}.
\end{aligned}$$

3. (a) We have:

$$\begin{aligned} P(\text{Faulty}) &= P(\text{Faulty} \,|\, A)\, P(A) + P(\text{Faulty} \,|\, B)\, P(B) \\ &\quad + P(\text{Faulty} \,|\, C)\, P(C) + P(\text{Faulty} \,|\, D)\, P(D) + P(\text{Faulty} \,|\, E)\, P(E) \\ &= 0.01 \times 0.1 + 0.2 \times 0.3 + 0.1 \times 0.2 + 0.05 \times 0.25 + 0.02 \times 0.15 \\ &= 0.0965. \end{aligned}$$

(b) We have:

$$P(B \,|\, \text{Faulty}) = \frac{P(B \cap \text{Faulty})}{P(\text{Faulty})} = \frac{0.2 \times 0.3}{P(\text{Faulty})} = \frac{0.06}{0.0965} = 0.6218.$$

(c) We have:

$$\begin{aligned} &P(A \,|\, S > c) \\ &= \frac{P(A \cap S > c)}{P(S > c)} \\ &= \frac{P(S > c \,|\, S \sim N(\mu_1, \sigma^2))\, P(A)}{P(S > c \,|\, A, B \text{ or } C)\, P(A, B \text{ or } C) + P(S > c \,|\, D \text{ or } E)\, P(D \text{ or } E)} \\ &= \frac{0.1 \times (1 - \Phi((c - \mu_1)/\sigma))}{(0.1 + 0.3 + 0.2) \times (1 - \Phi((c - \mu_1)/\sigma)) + (0.25 + 0.15) \times (1 - \Phi((c - \mu_2)/\sigma))} \\ &= \frac{1 - \Phi((c - \mu_1)/\sigma)}{10 - 6 \times \Phi((c - \mu_1)/\sigma) - 4 \times \Phi((c - \mu_2)/\sigma)}. \end{aligned}$$

(d) We have:

$$\begin{aligned} \mathrm{E}(S) &= \mathrm{E}(S \,|\, A, B \text{ or } C)\, P(A, B \text{ or } C) + \mathrm{E}(S \,|\, D \text{ or } E)\, P(D \text{ or } E) \\ &= 0.6\mu_1 + 0.4\mu_2 \end{aligned}$$

and:

$$\begin{aligned} \mathrm{E}(S^2) &= \mathrm{E}(S^2 \,|\, A, B \text{ or } C)\, P(A, B \text{ or } C) + \mathrm{E}(S^2 \,|\, D \text{ or } E)\, P(D \text{ or } E) \\ &= 0.6(\mu_1^2 + \sigma^2) + 0.4(\mu_2^2 + \sigma^2). \end{aligned}$$

Hence:

$$\begin{aligned} \mathrm{Var}(S) &= \mathrm{E}(S^2) - (\mathrm{E}(S))^2 \\ &= 0.6(\mu_1^2 + \sigma^2) + 0.4(\mu_2^2 + \sigma^2) - (0.6\mu_1 + 0.4\mu_2)^2 \\ &= \sigma^2 + 0.24(\mu_1^2 + \mu_2^2) - 0.48\mu_1\mu_2 \\ &= \sigma^2 + 0.24(\mu_1 - \mu_2)^2. \end{aligned}$$

4. (a) We have:

$$M_{S_n}(t) = \mathrm{E}(\mathrm{e}^{tS_n}) = \mathrm{E}\left(\mathrm{e}^{t\sum_{i=1}^{n} W_i X_i}\right) = \prod_{i=1}^{n} \mathrm{E}(\mathrm{e}^{tX_i W_i})$$
$$= \prod_{i=1}^{n} \mathrm{E}(\mathrm{E}(\mathrm{e}^{tX_i W_i} \mid X_i))$$
$$= \prod_{i=1}^{n} \mathrm{E}(p\mathrm{e}^{tX_i} + q)$$
$$= \prod_{i=1}^{n} (pM_X(t) + q)$$
$$= (pM_X(t) + q)^n$$

for $t \in I$.

(b) We have:

$$M_{R_N}(t) = \mathrm{E}(\mathrm{e}^{tR_n}) = \mathrm{E}\left(\mathrm{E}\left(\mathrm{e}^{t\sum_{i=1}^{N} X_i} \,\middle|\, N\right)\right) = \mathrm{E}\left(\prod_{i=1}^{N} \mathrm{E}(\mathrm{e}^{tX_i})\right)$$
$$= \mathrm{E}((M_X(t))^N)$$
$$= \mathrm{E}(\mathrm{e}^{N \log M_X(t)})$$
$$= M_N(\log M_X(t))$$
$$= M_N(K_X(t))$$

for $t \in I$, and $K_X(t) \in J$.

(c) We have:

$$M_N(t) = \mathrm{E}(\mathrm{e}^{tN}) = \sum_{i=0}^{n} (p\mathrm{e}^t)^i (1-p)^{n-i} = (p\mathrm{e}^t + q)^n$$

for $t \in J$. Hence:

$$M_N(K_X(t)) = \left(p\mathrm{e}^{K_X(t)} + q\right)^n = \left(p\mathrm{e}^{\log M_X(t)} + q\right)^n = (pM_X(t) + q)^n$$

for $t \in I$ and $K_X(t) \in J$. We then have $M_{R_N}(t) = M_{S_n}(t)$, so that by the one-to-one correspondence between distribution and moment generating function, R_N and S_n must have the same distribution.