UNIVERSITY OF LONDON | INTERNATIONAL PROGRAMMES

Statistics 2

J.S. Abdey

ST1**04b**

2014

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

This is an extract from a subject guide for an undergraduate course offered as part of the University of London International Programmes in Economics, Management, Finance and the Social Sciences. Materials for these programmes are developed by academics at the London School of Economics and Political Science (LSE).

For more information, see: www.londoninternational.ac.uk

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

This guide was prepared for the University of London International Programmes 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. If you have any comments on this subject guide, favourable or unfavourable, please use the form at the back of this guide.

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

Published by: University of London

Contents

Chapter 1
Introduction

1.1 Route map to the guide

This subject guide provides you with a framework for covering the syllabus of the **ST104b Statistics 2** half course and directs you to additional resources such as readings and the virtual learning environment (VLE).

The following 10 chapters will cover important aspects of elementary statistical theory, upon which many applications in **EC2020 Elements of econometrics** draw heavily. The chapters are not a series of self-contained topics, rather they build on each other sequentially. As such, you are strongly advised to follow the subject guide in chapter order. There is little point in rushing past material which you have only partially understood in order to reach the final chapter. Once you have completed your work on all of the chapters, you will be ready for examination revision. A good place to start is the sample examination paper which you will find at the end of the subject guide.

ST104b Statistics 2 extends the work of **ST104a Statistics 1** and provides a precise and accurate treatment of probability, distribution theory and statistical inference. As such there will be a strong emphasis on mathematical statistics as important discrete and continuous probability distributions are covered and properties of these distributions are investigated.

Point estimation techniques are discussed including method of moments, least squares and maximum likelihood estimation. Confidence interval construction and statistical hypothesis testing follow. Analysis of variance and a treatment of linear regression models, featuring the interpretation of computer-generated regression output and implications for prediction, round off the course.

Collectively, these topics provide a solid training in statistical analysis. As such, **ST104b Statistics 2** is of considerable value to those intending to pursue further study in statistics, econometrics and/or empirical economics. Indeed, the quantitative skills developed in the subject guide are readily applicable to all fields involving real data analysis.

1.2 Introduction to the subject area

Why study statistics?

By successfully completing this half course, you will understand the ideas of randomness and variability, and the way in which they link to probability theory. This will allow the use of a systematic and logical collection of statistical techniques of great

practical importance in many applied areas. The examples in this subject guide will concentrate on the social sciences, but the methods are important for the physical sciences too. This subject aims to provide a grounding in probability theory and some of the most common statistical methods.

The material in **ST104b Statistics 2** is necessary as preparation for other subjects you may study later on in your degree. The full details of the ideas discussed in this subject guide will not always be required in these other subjects, but you will need to have a solid understanding of the main concepts. This can only be achieved by seeing how the ideas emerge in detail.

How to study statistics

For statistics, you need some familiarity with abstract mathematical ideas, as well as the ability and common sense to apply these to real-life problems. The concepts you will encounter in probability and statistical inference are hard to absorb by just reading about them in a book. You need to read, then think a little, then try some problems, and then read and think some more. This procedure should be repeated until the problems are easy to do; *you should not spend a long time reading and forget about solving problems.*

1.3 Syllabus

The syllabus of **ST104b Statistics 2** is as follows:

- **Probability:** Set theory: the basics; Axiomatic definition of probability; Classical probability and counting rules; Conditional probability and Bayes' theorem.
- **Random variables:** Discete random variables; Continuous random variables.
- **Common distributions of random variables:** Common discrete distributions; Common continuous distributions.
- **Multivariate random variables:** Joint probability functions; Conditional distributions; Covariance and correlation; Independent random variables; Sums and products of random variables.
- **Sampling distributions of statistics:** Random samples; Statistics and their sampling distributions; Sampling distribution of a statistic; Sample mean from a normal population; The central limit theorem; Some common sampling distributions; Prelude to statistical inference.
- **Point estimation:** Estimation criteria: bias, variance and mean squared error; Method of moments estimation; Least squares estimation; Maximum likelihood estimation.
- **Interval estimation:** Interval estimation for means of normal distributions; Use of the chi-squared distribution; Confidence intervals for normal variances.

- **Hypothesis testing:** Setting p-value, significance level, test statistic; t tests; General approach to statistical tests; Two types of error; Tests for normal variances; Comparing two normal means with paired observations; Comparing two normal means; Tests for correlation coefficients; Tests for the ratio of two normal variances.
- **Analysis of variance (ANOVA):** One-way analysis of variance; Two-way analysis of variance.
- **Linear regression:** Simple linear regression; Inference for parameters in normal regression models; Regression ANOVA; Confidence intervals for E(y); Prediction intervals for y; Multiple linear regression models.

1.4 Aims of the course

The aim of this half course is to develop students' knowledge of elementary statistical theory. The emphasis is on topics that are of importance in applications to econometrics, finance and the social sciences. Concepts and methods that provide the foundation for more specialised courses in statistics are introduced.

1.5 Learning outcomes for the course

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

- apply and be competent users of standard statistical operators and be able to recall a variety of well-known distributions and their respective moments
- explain the fundamentals of statistical inference and apply these principles to justify the use of an appropriate model and perform hypothesis tests in a number of different settings
- demonstrate understanding that statistical techniques are based on assumptions and the plausibility of such assumptions must be investigated when analysing real problems.

1.6 Overview of learning resources

1.6.1 The subject guide

This course builds on the ideas encountered in **ST104a Statistics 1**. Although this subject guide offers a complete treatment of the course material, students may wish to consider purchasing a textbook. Apart from the textbooks recommended in this subject guide, you may wish to look in bookshops and libraries for alternative textbooks which may help you. A critical part of a good statistics textbook is the collection of problems to solve, and you may want to look at several different textbooks just to see a range of

practice questions, especially for tricky topics. The subject guide is there mainly to describe the syllabus and to show the level of understanding expected.

The subject guide is divided into chapters which should be worked through in the order in which they appear. There is little point in rushing past material you only partly understand to get to later chapters, as the presentation is somewhat sequential and not a series of self-contained topics. You should be familiar with the earlier chapters and have a solid understanding of them before moving on to the later ones.

The following procedure is recommended:

1. Read the introductory comments.
2. Consult the appropriate section of your textbook.
3. Study the chapter content, examples and learning activities.
4. Go through the learning outcomes carefully.
5. Attempt some of the problems from your textbook.
6. Refer back to this subject guide, or to the textbook, or to supplementary texts, to improve your understanding until you are able to work through the problems confidently.

The last two steps are the most important. It is easy to think that you have understood the material after reading it, but *working through problems is the crucial test of understanding*. Problem-solving should take up most of your study time.

Each chapter of the subject guide has suggestions for reading from the main textbook. Usually, you will only need to read the material in the main textbook (see 'Essential reading' below), but it may be helpful from time to time to look at others.

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.

Time management

About one-third of your self-study time should be spent reading and the rest should be spent solving problems. An internal student would expect maybe 15 hours of formal teaching and another 50 hours of private study to be enough to cover the subject. Of the 50 hours of private study, about 17 hours should be spent on the initial study of the textbook and subject guide. The remaining 33 hours should be spent on attempting problems, which may well require more reading.

Calculators

A calculator may be used when answering questions on the examination paper for **ST104b Statistics 2**. 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'.

Make sure you accustom yourself to using your chosen calculator and feel comfortable with it. Specifically, calculators must:

- have no external wires

must be:

- hand held
- compact and portable
- quiet in operation
- non-programmable

and must:

- not be capable of receiving, storing or displaying user-supplied non-numerical data.

The Regulations state: 'The use of a calculator that communicates or displays textual messages, graphical or algebraic information is strictly forbidden. Where a calculator is permitted in the examination, it must be a non-scientific calculator. Where calculators are permitted, only calculators limited to performing just basic arithmetic operations may be used. This is to encourage candidates to show the Examiners the steps taken in arriving at the answer.'

Computers

If you are aiming to carry out serious statistical analysis (which is beyond the level of this course) you will probably want to use some statistical software package such as Minitab, R or SPSS. It is not necessary for this course to have such software available, but if you do have access to it you may benefit from using it in your study of the material.

1.6.2 Essential reading

- Newbold, P., W.L. Carlson and B.M. Thorne, *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060].

Statistical tables

- Lindley, D.V. and W.F. Scott, *New Cambridge Statistical Tables*. (Cambridge: Cambridge University Press, 1995) second edition [ISBN 978-0521484855].

These statistical tables are the same ones that are distributed for use in the examination, so it is advisable that you become familiar with them, rather than those at the end of a textbook.

1.6.3 Further reading

Please note that, as long as you read the Essential reading, you are then free to read around the subject area in any text, paper or online resource. You will need to 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 texts for this course include:

- Johnson, R.A. and G.K. Bhattacharyya, *Statistics: Principles and Methods*. (New York: John Wiley and Sons, 2010) sixth edition [ISBN 9780470505779].
- Larsen, R.J. and M.L. Marx, *Introduction to Mathematical Statistics and Its Applications* (Pearson, 2013) fifth edition [ISBN 9781292023557].

While Newbold et al. is the main textbook for this course, there are many that are just as good. You are encouraged to look at those listed above and at any others you may find. It may be necessary to look at several textbooks for a single topic, as you may find that the approach of one textbook suits you better than that of another.

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

In addition to the subject guide and the Essential reading, it is crucial that you take advantage of the study resources that 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.londoninternational.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:

- Self-testing activities: Doing these allows you to test your own understanding of the subject material.

- Electronic study materials: The printed materials that you receive from the University of London are available to download, including updated reading lists and references.
- Past examination papers and *Examiners' commentaries*: These provide advice on how each examination question might best be answered.
- A student discussion forum: This is an open space for you to discuss interests and experiences, seek support from your peers, work collaboratively to solve problems and discuss subject material.
- Videos: There are recorded academic introductions to the subject, interviews and debates and, for some courses, audio-visual tutorials and conclusions.
- Recorded lectures: For some courses, where appropriate, the sessions from previous years' Study Weekends have been recorded and made available.
- Study skills: Expert advice on preparing for examinations and developing your digital literacy skills.
- Feedback forms.

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.

Making use of the Online Library

The Online Library 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:

http://tinyurl.com/ollathens

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 see the online help pages:

www.external.shl.lon.ac.uk/summon/about.php

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.org

Unless otherwise stated, all websites in this subject guide were accessed in April 2014. We cannot guarantee, however, that they will stay current and you may need to

perform an internet search to find the relevant pages.

1.7 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 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.

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.

The examination is by a two-hour unseen question paper. No books may be taken into the examination, but the use of calculators is permitted, and statistical tables and a formula sheet are provided (the formula sheet can be found in past examination papers available on the VLE).

The examination paper has a variety of questions, some quite short and others longer. All questions must be answered correctly for full marks. You may use your calculator whenever you feel it is appropriate, always remembering that the Examiners can give marks only for what appears on the examination script. Therefore, it is important to always show your working.

In terms of the examination, as always, it is important to manage your time carefully and not to dwell on one question for too long – move on and focus on solving the easier questions, coming back to harder ones later.

Chapter 2
Probability theory

2.1 Synopsis of chapter content

Probability is very important for statistics because it provides the rules that allow us to reason about uncertainty and randomness, which is the basis of statistics. Independence and conditional probability are profound ideas, but they must be fully understood in order to think clearly about any statistical investigation.

2.2 Aims of the chapter

The aims of this chapter are to:

- understand the concept of probability
- work with independent events and determine conditional probabilities
- work with probability problems.

2.3 Learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- explain the fundamental ideas of random experiments, sample spaces and events
- list the axioms of probability and be able to derive all the common probability rules from them
- list the formulae for the number of combinations and permutations of k objects out of n, and be able to routinely use such results in problems
- explain conditional probability and the concept of independent events
- prove the law of total probability and apply it to problems where there is a partition of the sample space
- prove Bayes' theorem and apply it to find conditional probabilities.

2.4 Essential reading

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060] Chapter 3.

In addition there is essential 'watching' of this chapter's accompanying video tutorials accessible via the **ST104b Statistics 2** area at http://my.londoninternational.ac.uk

2.5 Introduction

Consider the following hypothetical example: A country will soon hold a referendum about whether it should join the European Union (EU). An opinion poll of a random sample of people in the country is carried out.

950 respondents say that they plan to vote in the referendum. They answer the question 'Will you vote Yes or No to joining the EU?' as follows:

	Answer		
	Yes	No	Total
Count	513	437	950
%	54%	46%	100%

However, we are not interested in just this sample of 950 respondents, but in the population that they represent, that is all likely voters.

Statistical inference will allow us to say things like the following about the population:

- 'A 95% confidence interval for the population proportion, π, of 'Yes' voters is $(0.508,\ 0.572)$.'
- 'The null hypothesis that $\pi = 0.5$, against the alternative hypothesis that $\pi > 0.5$, is rejected at the 5% significance level.'

In short, the opinion poll gives statistically significant evidence that 'Yes' voters are in the majority among likely voters. Such methods of statistical inference will be discussed later in the course.

The inferential statements about the opinion poll rely on the following assumptions and results:

- Each response X_i is a realisation of a **random variable** from a **Bernoulli distribution** with **probability** parameter π.
- The responses $X_1, X_2, \ldots, X_n$ are **independent** of each other.
- The **sampling distribution** of the sample mean (proportion) $\bar{X}$ has **expected value** π and **variance** $\pi\,(1-\pi)/n$.

- By use of the **central limit theorem**, the sampling distribution is approximately a **normal distribution**.

In the next few chapters, we will learn about the terms in bold, among others.

The need for probability in statistics

In statistical inference, the data that we have observed are regarded as a *sample* from a broader *population*, selected with a **random** process:

- Values in a sample are *variable*: If we collected a different sample we would not observe exactly the same values again.
- Values in a sample are also *random*: We cannot predict the precise values that will be observed before we actually collect the sample.

Probability theory is the branch of mathematics that deals with randomness. So we need to study this first.

A preview of probability

The first basic concepts in probability will be the following:

- **Experiment**: For example, rolling a single die and recording the outcome.
- **Outcome** of the experiment: For example, rolling a 3.
- **Sample space** S: The *set* of all possible outcome; here $\{1, 2, 3, 4, 5, 6\}$.
- **Event**: Any *subset* A of the sample space, for example $A = \{4, 5, 6\}$.[1]

Probability, $P(A)$, will be defined as a function which assigns probabilities (real numbers) to events (sets). This uses the language and concepts of **set theory**. So we need to study the basics of set theory first.

2.6 Set theory: the basics

A **set** is a collection of **elements** (also known as 'members' of the set).

Example 2.1 The following are all examples of sets:

- $A = \{\text{Amy, Bob, Sam}\}$.
- $B = \{1, 2, 3, 4, 5\}$.
- $C = \{x \mid x \text{ is a prime number}\} = \{2, 3, 5, 7, 11, \dots\}$.
- $D = \{x \mid x \geq 0\}$ (that is, the set of all non-negative real numbers).

[1] Strictly speaking not all subsets are events, as discussed later.

Membership of sets and the empty set

$x \in A$ means that object x is an element of set A.

$x \notin A$ means that object x is not an element of set A.

The **empty set**, denoted $\emptyset$, is the set with no elements, i.e. $x \notin \emptyset$ is true for every object x, and $x \in \emptyset$ is not true for any object x.

Example 2.2 If $A = \{1, 2, 3, 4, 5\}$, then:

- $1 \in A$ and $2 \in A$.
- $6 \notin A$ and $1.5 \notin A$.

The familiar **Venn diagrams** help to visualise statements about sets. However, Venn diagrams are *not formal proofs* of results in set theory.

Example 2.3 In Figure 2.1, the darkest area in the middle is $A \cap B$, the total shaded area is $A \cup B$, and the white area is $(A \cup B)^c = A^c \cap B^c$.

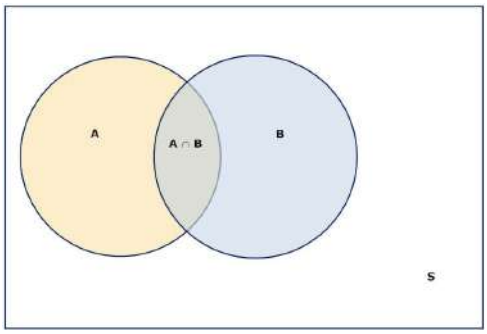


Figure 2.1: Venn diagram depicting $A \cup B$ (the total shaded area).

Subsets and equality of sets

$A \subset B$ means that set A is a **subset** of set B, defined as:

$$A \subset B \quad \text{when} \quad x \in A \quad \Rightarrow \quad x \in B.$$

Hence A is a subset of B if every element of A is also an element of B. An example is shown in Figure 2.2.

Example 2.4 An example of the distinction between subsets and non-subsets is:

- $\{1, 2, 3\} \subset \{1, 2, 3, 4\}$, because all elements appear in the larger set.
- $\{1, 2, 5\} \not\subset \{1, 2, 3, 4\}$, because the element 5 does not appear in the larger set.

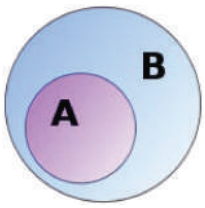


Figure 2.2: Venn diagram depicting a subset, where $A \subset B$.

Two sets A and B are equal ($A = B$) if they have exactly the same elements. This implies that $A \subset B$ *and* $B \subset A$.

Unions of sets ('or')

The **union**, denoted $\cup$, of two sets is:

$$A \cup B = \{x \,|\, x \in A \text{ or } x \in B\}.$$

That is, the set of those elements which belong to A *or* B (or both). An example is shown in Figure 2.3.

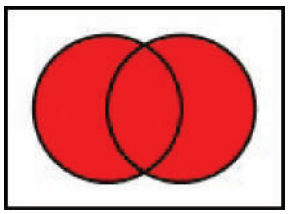


Figure 2.3: Venn diagram depicting the union of two sets.

Example 2.5 If $A = \{1, 2, 3, 4\}$, $B = \{2, 3\}$ and $C = \{4, 5, 6\}$, then:

- $A \cup B = \{1, 2, 3, 4\}$
- $A \cup C = \{1, 2, 3, 4, 5, 6\}$
- $B \cup C = \{2, 3, 4, 5, 6\}$.

Intersections of sets ('and')

The **intersection**, denoted $\cap$, of two sets is:

$$A \cap B = \{x \,|\, x \in A \text{ and } x \in B\}.$$

That is, the set of those elements which belong to both A *and* B. An example is shown in Figure 2.4.

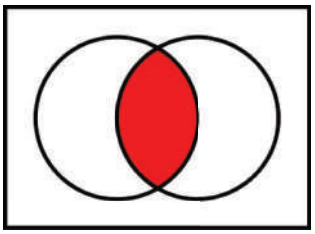


Figure 2.4: Venn diagram depicting the intersection of two sets.

Example 2.6 If $A = \{1, 2, 3, 4\}$, $B = \{2, 3\}$ and $C = \{4, 5, 6\}$, then:

- $A \cap B = \{2, 3\}$
- $A \cap C = \{4\}$
- $B \cap C = \emptyset$.

Unions and intersections of many sets

Both set operators can also be applied to more than two sets, such as $A \cap B \cap C$.

Concise notation for the unions and intersections of sets $A_1, A_2, \ldots, A_n$ is:

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n$$

$$\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \cdots \cap A_n.$$

These can also be used for an infinite number of sets, i.e. when n is replaced by ∞.

Complement ('not')

Suppose S is the set of *all* possible elements which are under consideration. In probability, S will be referred to as the **sample space**.

It follows that $A \subset S$ for every set A we may consider. The **complement** of A with respect to S is:

$$A^c = \{x \,|\, x \in S \text{ and } x \notin A\}.$$

That is, the set of those elements of S that are *not* in A. An example is shown in Figure 2.5.

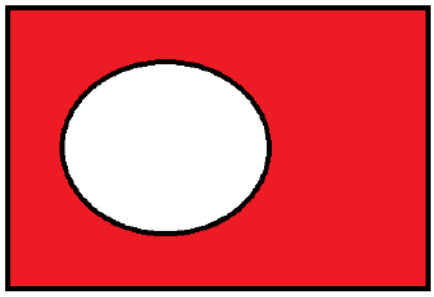


Figure 2.5: Venn diagram depicting the complement of a set.

Properties of set operators

In proofs and derivations about sets, you can use the following results without proof:

- Commutativity:

$$A \cap B = B \cap A \quad \text{and} \quad A \cup B = B \cup A.$$

- Associativity:

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

- Distributive laws:

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

- De Morgan's laws:

$$(A \cap B)^c = A^c \cup B^c \quad \text{and} \quad (A \cup B)^c = A^c \cap B^c.$$

If S is the sample space and A and B are any sets in S, you can also use the following results without proof:

- $\emptyset^c = S$.
- $\emptyset \subset A, \quad A \subset A \quad \text{and} \quad A \subset S$.
- $A \cap A = A \quad \text{and} \quad A \cup A = A$.
- $A \cap A^c = \emptyset \quad \text{and} \quad A \cup A^c = S$.
- If $B \subset A, \quad A \cap B = B \quad \text{and} \quad A \cup B = A$.
- $A \cap \emptyset = \emptyset \quad \text{and} \quad A \cup \emptyset = A$.
- $A \cap S = A \quad \text{and} \quad A \cup S = S$.
- $\emptyset \cap \emptyset = \emptyset \quad \text{and} \quad \emptyset \cup \emptyset = \emptyset$.

Mutually exclusive events

Two sets A and B are **disjoint** or **mutually exclusive** if:

$$A \cap B = \emptyset.$$

Sets $A_1, A_2, \ldots, A_n$ are **pairwise disjoint** if all pairs of sets from them are disjoint, i.e. $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Partition

The sets $A_1, A_2, \ldots, A_n$ form a **partition** of the set A if they are pairwise disjoint and if $\bigcup_{i=1}^{n} A_i = A$, that is $A_1, A_2, \ldots, A_n$ are **collectively exhaustive** of A.

Therefore, a partition divides the entire set A into non-overlapping pieces A_i, as shown in Figure 2.6 for $n = 3$. Similarly, an infinite collection of sets $A_1, A_2, \ldots$ form a partition of A if they are pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = A$.

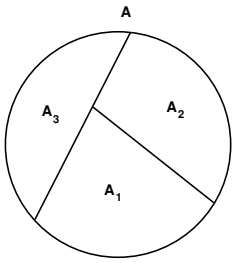
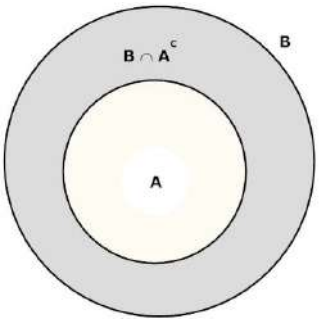


Figure 2.6: The partition of the set A into A_1, A_2 and A_3.

Example 2.7 Suppose that $A \subset B$. Show that A and $B \cap A^c$ form a partition of B.



We have:

$$A \cap (B \cap A^c) = (A \cap A^c) \cap B = \emptyset \cap B = \emptyset$$

and:

$$A \cup (B \cap A^c) = (A \cup B) \cap (A \cup A^c) = B \cap S = B.$$

Hence A and $B \cap A^c$ are mutually exclusive and collectively exhaustive of B, and so they form a partition of B.

2.7 Axiomatic definition of probability

First, we consider four basic concepts in probability.

An **experiment** is a process which produces outcomes and which can have several *different* **outcomes**. The **sample space** S is the set of all possible outcomes of the experiment. An **event** is any subset A of the sample space such that $A \subset S$.

> **Example 2.8** If the experiment is 'select a trading day at random and record the % change in the FTSE 100 index from the previous trading day', then the outcome is the % change in the FTSE 100 index.
>
> $S = [-100, +\infty)$ for the % change in the FTSE 100 index (in principle).
>
> An event of interest might be $A = \{x \mid x > 0\}$ – the event that the daily change is positive, i.e. the FTSE 100 index gains value from the previous trading day.

The sample space and events are represented as sets. For two events A and B, set operations are then interpreted as follows:

- $A \cap B$: both A and B happen.
- $A \cup B$: either A or B happens (or both happen).
- A^c: A does not happen, i.e. something other than A happens.

Once we introduce *probabilities* of events, we can also say that:

- the sample space S is a *certain* event
- the empty set $\emptyset$ is an *impossible* event.

Axioms of probability

'Probability' is formally defined as a function $P(A)$ from subsets (events) of the sample space S onto real numbers.[2] Such a function is a **probability function** if it satisfies the following **axioms** ('self-evident truths'):

Axiom 1: $P(A) \geq 0$ for all events A.

Axiom 2: $P(S) = 1$.

Axiom 3: If events $A_1, A_2, \ldots$ are pairwise disjoint (i.e. $A_i \cap A_j = \emptyset$ for all $i \neq j$), then:

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

[2]The precise definition also requires a careful statement of *which* subsets of S are allowed as events; we can skip that on this course.

The axioms require that a probability function must always satisfy these requirements:

- Axiom 1 requires that probabilities are always non-negative.
- Axiom 2 requires that the outcome is some element from the sample space with certainty (that is, with probability 1). In other words, the experiment must have *some* outcome.
- Axiom 3 states that if events $A_1, A_2, \ldots$ are mutually exclusive, the probability of their union is simply the sum of their individual probabilities.

All other properties of the probability function can be derived from the axioms. We begin by showing that a result like Axiom 3 also holds for *finite* collections of mutually exclusive sets.

2.7.1 Basic properties of probability

Probability property

For the empty set, $\emptyset$, we have:

$$P(\emptyset) = 0. \tag{2.1}$$

Proof: Since $\emptyset \cap \emptyset = \emptyset$ and $\emptyset \cup \emptyset = \emptyset$, Axiom 3 gives:

$$P(\emptyset) = P(\emptyset \cup \emptyset \cup \cdots) = \sum_{i=1}^{\infty} P(\emptyset).$$

But the only real number for $P(\emptyset)$ which satisfies this is $P(\emptyset) = 0$. □

Probability property

(*Finite* additivity:) If $A_1, A_2, \ldots, A_n$ are pairwise disjoint, then:

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

Proof: In Axiom 3, set $A_{n+1} = A_{n+2} = \cdots = \emptyset$, so that:

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

since $P(A_i) = P(\emptyset) = 0$ for $i = n+1, n+2, \ldots$. □

In pictures, the previous result means that in a situation like the one shown in Figure 2.7, the probability of the combined event $A = A_1 \cup A_2 \cup A_3$ is simply the sum of the probabilities of the individual events:

$$P(A) = P(A_1) + P(A_2) + P(A_3).$$

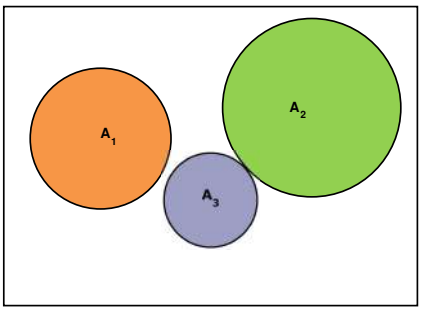


Figure 2.7: Venn diagram depicting three mutually exclusive sets, A_1, A_2 and A_3. Note although A_2 and A_3 have touching boundaries, there is no actual *intersection* and hence they are (pairwise) mutually exclusive.

That is, we can simply sum probabilities of mutually exclusive sets. This is very useful for deriving further results.

Probability property

For any event A, we have:

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

Proof: We have that $A \cup A^c = S$ and $A \cap A^c = \emptyset$. Therefore:

$$1 = P(S) = P(A \cup A^c) = P(A) + P(A^c)$$

using the previous result, with $n = 2$, $A_1 = A$ and $A_2 = A^c$. □

Probability property

For any event A, we have:

$$P(A) \leq 1.$$

Proof (by contradiction): If it was true that $P(A) > 1$ for some A, then we would have:

$$P(A^c) = 1 - P(A) < 0.$$

This violates Axiom 1, so cannot be true. Therefore it must be that $P(A) \leq 1$ for all A. Putting this and Axiom 1 together, we get:

$$0 \leq P(A) \leq 1$$

for all events A. □

Probability property

For any two events A and B, if $A \subset B$, then $P(A) \leq P(B)$.

Proof: We proved in Example 2.7 that we can partition B as $B = A \cup (B \cap A^c)$ where the two sets in the union are disjoint. Therefore:

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

since $P(B \cap A^c) \geq 0$. □

Activity 2.1 For any two events A and B, prove that:

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

In summary, the probability function has the following properties:

- $P(S) = 1$ and $P(\emptyset) = 0$.
- $0 \leq P(A) \leq 1$ for all events A.
- If $A \subset B$, then $P(A) \leq P(B)$.

These show that the probability function has the kinds of values we expect of something called a 'probability'.

- $P(A^c) = 1 - P(A)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

These are useful for deriving probabilities of new events.

Example 2.9 Suppose that, on an average weekday, of all adults in a country:

- 86% spend at least 1 hour watching television (event A, with $P(A) = 0.86$).
- 19% spend at least 1 hour reading newspapers (event B, with $P(B) = 0.19$).
- 15% spend at least 1 hour watching television, *and* at least 1 hour reading newspapers ($P(A \cap B) = 0.15$).

We select a member of the population for an interview at random. Then, for example, we have:

- $P(A^c) = 1 - P(A) = 1 - 0.86 = 0.14$: the probability that the respondent watches *less than* 1 hour of television.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.86 + 0.19 - 0.15 = 0.90$: the probability that the respondent spends at least 1 hour watching television or reading newspapers (or both).

What does 'probability' mean?

Probability theory tells us how to work with the probability function and derive 'probabilities of events' from it. However, it does not tell us what 'probability' really means.

There are several alternative interpretations of the real-world meaning of 'probability' in this sense. One of them is outlined below. The mathematical theory of probability and calculations on probabilities are the same whichever interpretation we assign to 'probability'. So, in this course, we do not need to discuss the matter further.

Frequency interpretation of probability

This states that the probability of an outcome A of an experiment is the proportion (**relative frequency**) of trials in which A would be the outcome if the experiment was repeated a very large number of times under similar conditions.

Example 2.10 How should we interpret the following, as statements about the real world of coins and babies?

- 'The probability that a tossed coin comes up heads is 0.5.' If we tossed a coin a large number of times, and the proportion of heads out of those tosses was 0.5, the 'probability of heads' could be said to be 0.5, for that coin.
- 'The probability is 0.51 that a child born in the UK today is a boy.' If the proportion of boys among a large number of live births was 0.51, the 'probability of a boy' could be said to be 0.51.

How to find probabilities?

A key question is how to determine appropriate numerical values $P(A)$ for the probabilities of particular events.

This is usually done *empirically*, by observing actual realisations of the experiment and using them to **estimate** probabilities. In the simplest cases, this basically applies the frequency definition to observed data.

Example 2.11

- If I toss a coin 10,000 times, and 5,023 of the tosses come up heads, it seems that, approximately, $P(\text{heads}) = 0.5$, for that coin.
- Of the 7,098,667 live births in England and Wales in the period 1999–2009, 51.26% were boys. So we could assign the value of about 0.51 to the probability of a boy in that population.

The estimation of probabilities of events from observed data is an important part of statistics.

2.8 Classical probability and counting rules

Classical probability is a simple special case where values of probabilities can be found by just counting outcomes. This requires that:

- The sample space contains only a *finite* number of outcomes.
- All of the outcomes are *equally probable*.

Standard illustrations of classical probability are devices used in games of chance:

- Tossing a coin (heads or tails) one or more times.
- Rolling one or more dice (each scored 1, 2, 3, 4, 5 or 6).
- Drawing one or more playing cards from a deck of 52 cards.

We will use these often, not because they are particularly important but because they provide simple examples for illustrating various results in probability.

Suppose that the sample space S contains m equally likely outcomes, and that event A consists of $k \leq m$ of these outcomes. Then:

$$P(A) = \frac{k}{m} = \frac{\text{number of outcomes in } A}{\text{total number of outcomes in sample space } S}.$$

That is, the probability of A is the *proportion* of outcomes that belong to A out of all possible outcomes.

In the classical case, the probability of any event can be determined by **counting** the number of outcomes that belong to the event, and the total number of possible outcomes.

Example 2.12 Rolling two dice, what is the probability that the sum of the two scores is 5?

- Sample space: the 36 ordered pairs:

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

- Outcomes in the event: $A = \{(1,4),(2,3),(3,2),(4,1)\}$.
- The probability: $P(A) = 4/36 = 1/9$.

Now that we have a way of obtaining probabilities for events in the classical case, we can use it together with the rules of probability.

The formula $P(A) = 1 - P(A^c)$ is convenient when we want $P(A)$ but the probability of the complementary event A^c, i.e. $P(A^c)$, is easier to find.

> **Example 2.13** When rolling two fair dice, what is the probability that the sum of the dice is greater than 3?
>
> - The complement is that the sum is at most 3, i.e. the complementary event is $A^c = \{(1,1),(1,2),(2,1)\}$.
> - Therefore, $P(A) = 1 - 3/36 = 33/36 = 11/12$.

The formula:

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

says that the probability that A or B happens (or both happen) is the *sum* of the probabilities of A and B, *minus* the probability that both A and B happen.

> **Example 2.14** When rolling two fair dice, what is the probability that the two scores are equal (event A) or that the total score is greater than 10 (event B)?
>
> - $P(A) = 6/36$, $P(B) = 3/36$ and $P(A \cap B) = 1/36$.
> - So $P(A \cup B) = P(A) + P(B) - P(A \cap B) = (6 + 3 - 1)/36 = 8/36 = 2/9$.

2.8.1 Combinatorial counting methods

A powerful set of counting methods answers the following question: How many ways are there to select k objects out of n distinct objects?

The answer will depend on two things:

- Whether the selection is **with replacement** (an object can be selected more than once) or **without replacement** (an object can be selected only once).
- Whether the selected set is treated as **ordered** or **unordered**.

Ordered sets, with replacement

Suppose that the selection of k objects out of n needs to be:

- ordered, so that the selection is an ordered *sequence* where we distinguish between the 1st object, 2nd, 3rd etc.
- with replacement, so that each of the n objects may appear several times in the selection.

Then:

- n objects are available for selection for the 1st object in the sequence
- n objects are available for selection for the 2nd object in the sequence
- ... and so on, until n objects are available for selection for the kth object in the sequence.

The number of possible ordered sequences of k objects selected with replacement from n objects is therefore:

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

Ordered sets, without replacement

Suppose that the selection of k objects out of n is again treated as an ordered sequence, but that selection is now:

- ordered, so that the selection is an ordered *sequence* where we distinguish between the 1st object, 2nd, 3rd etc.
- without replacement: if an object is selected once, it cannot be selected again.

Now:

- n objects are available for selection for the 1st object in the sequence
- $n-1$ objects are available for selection for the 2nd object
- $n-2$ objects are available for selection for the 3rd object
- ... and so on, until $n-k+1$ objects are available for selection for the kth object.

The number of possible ordered sequences of k objects selected without replacement from n objects is therefore:

$$n \times (n-1) \times \cdots \times (n-k+1). \tag{2.2}$$

An important special case is when $k = n$.

Factorials

The number of ordered sets of n objects, selected without replacement from n objects, is:

$$n! = n \times (n-1) \times \cdots \times 2 \times 1.$$

The number $n!$ (read 'n **factorial**') is the total number of different ways in which n objects can be arranged in an ordered sequence. This is known as the number of **permutations** of n objects.

We also define $0! = 1$.

Using factorials, (2.2) can be written as:

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

Unordered sets, without replacement

Suppose now that the *identities* of the objects in the selection matter, but the *order* does not.

- For example, the sequences $(1,2,3)$, $(1,3,2)$, $(2,1,3)$, $(2,3,1)$, $(3,1,2)$, $(3,2,1)$ are now all treated as the same, because they all contain the elements 1, 2 and 3.

The number of such unordered subsets (**combinations**) of k out of n objects is determined as follows:

- The number of ordered sequences is $n!/(n-k)!$.
- Among these, every different combination of k distinct elements appears $k!$ times, in different orders.
- Ignoring the ordering, there are therefore:

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

  different combinations, for each $k = 0, 1, \ldots, n$.

The number $\binom{n}{k}$ is known as the **binomial coefficient**. Note that because $0! = 1$, $\binom{n}{0} = \binom{n}{n} = 1$, so there is only 1 way of selecting 0 or n out of n objects.

Example 2.15 Suppose we have $k = 3$ people (Amy, Bob and Sam). How many different sets of birthdays can they have (day and month, ignoring the year, and pretending February 29th does not exist, so that $n = 365$), in the following cases?

1. It makes a difference who has which birthday (*ordered*), i.e. Amy (January 1st), Bob (May 5th) and Sam (December 5th) is different from Amy (May 5th), Bob (December 5th) and Sam (January 1st), and different people can have the same birthday (*with replacement*). The number of different sets of birthdays is:

$$365^3 = 48{,}627{,}125.$$

2. It makes a difference who has which birthday (*ordered*), and different people must have different birthdays (*without replacement*). The number of different sets of birthdays is:

$$\frac{365!}{(365-3)!} = 365 \times 364 \times 363 = 48{,}228{,}180.$$

3. Only the dates matter, but not who has which one (*unordered*), i.e. Amy (January 1st), Bob (May 5th) and Sam (December 5th) is treated as the same as Amy (May 5th), Bob (December 5th) and Sam (January 1st), and different people must have different birthdays (*without replacement*). The number of different sets of birthdays is:

$$\binom{365}{3} = \frac{365!}{(365-3)!\,3!} = \frac{365 \times 364 \times 363}{3 \times 2 \times 1} = 8{,}038{,}030.$$

Example 2.16 Consider a room with r people in it. What is the probability that *at least two of them have the same birthday* (call this event A)? In particular, what is the smallest r for which $P(A) > 1/2$?

Assume that all days are equally likely.

Label the people 1 to r, so that we can treat them as an ordered list and talk about person 1, person 2 etc. We want to know how many ways there are to assign birthdays to this list of people. We note the following:

1. The number of all possible sequences of birthdays, allowing repeats (i.e. with replacement) is 365^r.
2. The number of sequences where *all birthdays are different* (i.e. without replacement) is $365!/(365-r)!$.

Here '1.' is the size of the sample space, and '2.' is the number of outcomes which satisfy A^c, the complement of the case in which we are interested.

Therefore:

$$P(A^c) = \frac{365!/(365-r)!}{365^r} = \frac{365 \times 364 \times \cdots \times (365-r+1)}{365^r}$$

and:

$$P(A) = 1 - P(A^c) = 1 - \frac{365 \times 364 \times \cdots \times (365-r+1)}{365^r}.$$

Probabilities, for $P(A)$, of at least two people sharing a birthday, for different values of the number of people r are given in the following table:

r	$P(A)$	r	$P(A)$	r	$P(A)$	r	$P(A)$
2	0.003	12	0.167	22	0.476	32	0.753
3	0.008	13	0.194	23	0.507	33	0.775
4	0.016	14	0.223	24	0.538	34	0.795
5	0.027	15	0.253	25	0.569	35	0.814
6	0.040	16	0.284	26	0.598	36	0.832
7	0.056	17	0.315	27	0.627	37	0.849
8	0.074	18	0.347	28	0.654	38	0.864
9	0.095	19	0.379	29	0.681	39	0.878
10	0.117	20	0.411	30	0.706	40	0.891
11	0.141	21	0.444	31	0.730	41	0.903

2.9 Conditional probability and Bayes' theorem

Next we introduce some of the most important concepts in probability:

- Independence
- Conditional probability
- Bayes' theorem.

These give us powerful tools for:

- deriving probabilities of combinations of events
- updating probabilities of events, after we learn that some other events have happened.

Independence

Two events A and B are **(statistically) independent** if:

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

Independence is sometimes denoted $A \perp\!\!\!\perp B$. Intuitively, independence means that:

- if A happens, this does not affect the probability of B happening (and vice versa)
- if you are told that A has happened, this does not give you any new information about the value of $P(B)$ (and vice versa).

For example, independence is often a reasonable assumption when A and B correspond to physically separate experiments.

Example 2.17 Suppose we roll two dice. We assume that all combinations of the values of them are equally likely. Define the events:

- $A =$ 'Score of die 1 is not 6'
- $B =$ 'Score of die 2 is not 6'.

Then:

- $P(A) = 30/36 = 5/6$
- $P(B) = 30/36 = 5/6$
- $P(A \cap B) = 25/36 = 5/6 \times 5/6 = P(A)\,P(B)$, so A and B are independent.

Independence of multiple events

Events $A_1, A_2, \ldots, A_n$ are independent if the probability of the intersection of any subset of these events is the product of the individual probabilities of the events in the subset.

This implies the important result that *if* events $A_1, A_2, \ldots, A_n$ are independent, then:

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)\, P(A_2) \cdots P(A_n).$$

Note that there is a difference between *pairwise independence* and *full independence*. The following example illustrates.

Example 2.18 It can be cold in London. Four impoverished teachers dress to feel warm. Teacher A has a hat and a scarf and gloves, Teacher B has only a hat, Teacher C has only a scarf and Teacher D has only gloves. One teacher out of the four is selected at random. Show that although each *pair* of events $H =$ 'the teacher selected has a hat', $S =$ 'the teacher selected has a scarf', and $G =$ 'the teacher selected has gloves' are independent, all *three* of these events are not independent.

Two teachers have a hat, two teachers have a scarf, and two teachers have gloves, so:

$$P(H) = \frac{2}{4} = \frac{1}{2}, \quad P(S) = \frac{2}{4} = \frac{1}{2} \quad \text{and} \quad P(G) = \frac{2}{4} = \frac{1}{2}.$$

Only one teacher has both a hat and a scarf, so:

$$P(H \cap S) = \frac{1}{4}$$

and similarly:

$$P(H \cap G) = \frac{1}{4} \quad \text{and} \quad P(S \cap G) = \frac{1}{4}.$$

From these results, we can verify that:

$$\begin{aligned} P(H \cap S) &= P(H)\, P(S) \\ P(H \cap G) &= P(H)\, P(G) \\ P(S \cap G) &= P(S)\, P(G) \end{aligned}$$

and so the events are pairwise independent. But one teacher has a hat, a scarf and gloves, so:

$$P(H \cap S \cap G) = \frac{1}{4} \neq P(H)\, P(S)\, P(G).$$

Hence the three events are not independent. If the selected teacher has a hat and a scarf, then we *know* that the teacher has gloves. There is no independence for all three events together.

Independent versus mutually exclusive events

The idea of independent events is quite different from that of *mutually exclusive* (disjoint) events, as shown in Figure 2.8.

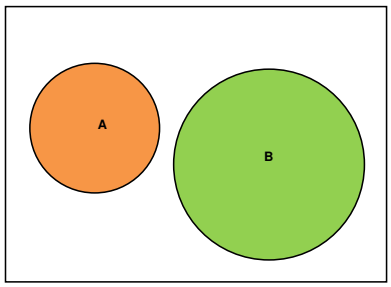


Figure 2.8: Venn diagram depicting mutually exclusive events.

For mutually exclusive events $A \cap B = \emptyset$, and so, from (2.1), $P(A \cap B) = 0$. For independent events, $P(A \cap B) = P(A)\,P(B)$. So since $P(A \cap B) = 0 \neq P(A)\,P(B)$ in general (except in the uninteresting case that $P(A) = 0$ or $P(B) = 0$), then mutually exclusive events and independent events are different.

In fact, mutually exclusive events are extremely *non*-independent (i.e. **dependent**). For example, if you know that A has happened, you know for certain that B has *not* happened. There is no particularly helpful way to represent independent events using a Venn diagram.

Conditional probability

Consider two events A and B. Suppose that you are told that B has occurred. How does this affect the probability of event A?

The answer is given by the conditional probability of A given that B has occurred, or the **conditional probability of A given B** for short, defined as:

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

assuming that $P(B) > 0$. The conditional probability is not defined if $P(B) = 0$.

Example 2.19 Suppose we roll two independent fair dice again. Consider the following events:

- $A =$ 'at least one of the scores is 2'.
- $B =$ 'the sum of the scores is greater than 7'.

There are shown in Figure 2.9. Now $P(A) = 11/36 \approx 0.31$, $P(B) = 15/36$ and $P(A \cap B) = 2/36$. The conditional probability of A given B is therefore:

$$P(A \,|\, B) = \frac{P(A \cap B)}{P(B)} = \frac{2/36}{15/36} = \frac{2}{15} \approx 0.13.$$

Learning that B has happened causes us to *revise* (update) the probability of A downwards, from 0.31 to 0.13.

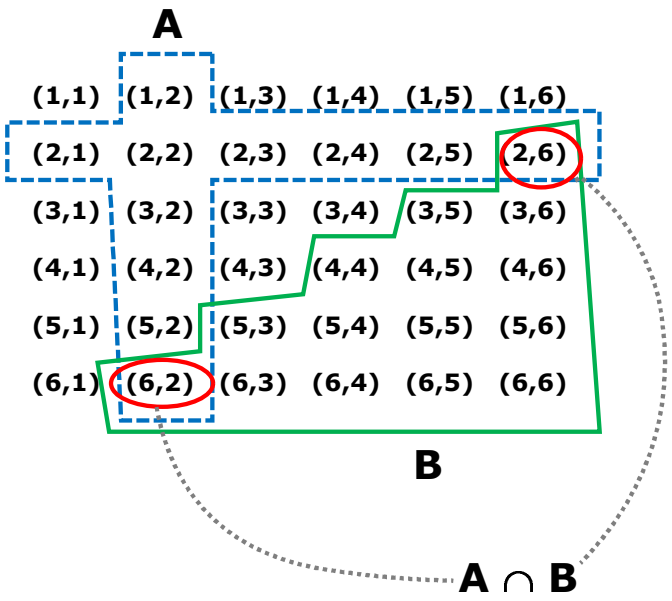


Figure 2.9: Events A, B and $A \cap B$ for Example 2.19.

One way to think about conditional probability is that when we condition on B, we redefine the sample space to be B.

> **Example 2.20** In Example 2.19, when we are told that the conditioning event B has occurred, we know we are within the green line in Figure 2.9. So the 15 outcomes within it become the new sample space. There are 2 outcomes which satisfy A *and* which are inside this new sample space, so:
>
> $$P(A \mid B) = \frac{2}{15} = \frac{\text{cases of } A \text{ within } B}{\text{cases of } B}.$$

Conditional probability of independent events

If $A \perp\!\!\!\perp B$, i.e. $P(A \cap B) = P(A)\,P(B)$, and $P(B) > 0$ and $P(A) > 0$, then:

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

and:

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

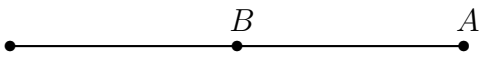
In other words, if A and B are independent, learning that B has occurred does not change the probability of A, and learning that A has occurred does not change the probability of B. This is exactly what we would expect under independence.

Chain rule of conditional probabilities

Since $P(A \mid B) = P(A \cap B)/P(B)$, then:

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

That is, the probability that both A and B occur is the probability that A occurs given that B has occurred multiplied by the probability that B occurs. An intuitive graphical version of this is:



The path to A is to get first to B, and then from B to A.

It is also true that:

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

and you can use whichever is more convenient. Very often some version of this **chain rule** is much easier than calculating $P(A \cap B)$ directly.

The chain rule generalises to multiple events:

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1, A_2) \cdots P(A_n \mid A_1, \ldots, A_{n-1})$$

where, for example, $P(A_3 \mid A_1, A_2)$ is shorthand for $P(A_3 \mid A_1 \cap A_2)$. The events can be taken in any order, as shown in Example 2.21.

Example 2.21 For $n = 3$, we have:

$$\begin{aligned}
P(A_1 \cap A_2 \cap A_3) &= P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1, A_2) \\
&= P(A_1)\, P(A_3 \mid A_1)\, P(A_2 \mid A_1, A_3) \\
&= P(A_2)\, P(A_1 \mid A_2)\, P(A_3 \mid A_1, A_2) \\
&= P(A_2)\, P(A_3 \mid A_2)\, P(A_1 \mid A_2, A_3) \\
&= P(A_3)\, P(A_1 \mid A_3)\, P(A_2 \mid A_1, A_3) \\
&= P(A_3)\, P(A_2 \mid A_3)\, P(A_1 \mid A_2, A_3).
\end{aligned}$$

Example 2.22 Suppose you draw 4 cards from a deck of 52 playing cards. What is the probability of $A =$ 'the cards are the 4 aces (cards of rank 1)'?

We could calculate this using counting rules. There are $\binom{52}{4} = 270{,}725$ possible subsets of 4 different cards, and only 1 of these consists of the 4 aces. Therefore $P(A) = 1/270275$.

Let us try with conditional probabilities. Define A_i as 'the ith card is an ace', so that $A = A_1 \cap A_2 \cap A_3 \cap A_4$. The necessary probabilities are:

- $P(A_1) = 4/52$ since there are initially 4 aces in the deck of 52 playing cards.
- $P(A_2 \mid A_1) = 3/51$. *If* the first card is an ace, 3 aces remain in the deck of 51 playing cards from which the second card will be drawn.
- $P(A_3 \mid A_1, A_2) = 2/50$.
- $P(A_4 \mid A_1, A_2, A_3) = 1/49$.

Putting these together with the chain rule gives:

$$\begin{aligned} P(A) &= P(A_1)\,P(A_2 \mid A_1)\,P(A_3 \mid A_1, A_2)\,P(A_4 \mid A_1, A_2, A_3) \\ &= \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49} = \frac{24}{6497400} = \frac{1}{270725}. \end{aligned}$$

Here we could obtain the result in two ways. However, there are very many situations where classical probability and counting rules are not usable, whereas conditional probabilities and the chain rule are completely general and always applicable.

More methods for summing probabilities

We now return to probabilities of partitions like the situation shown in Figure 2.10.

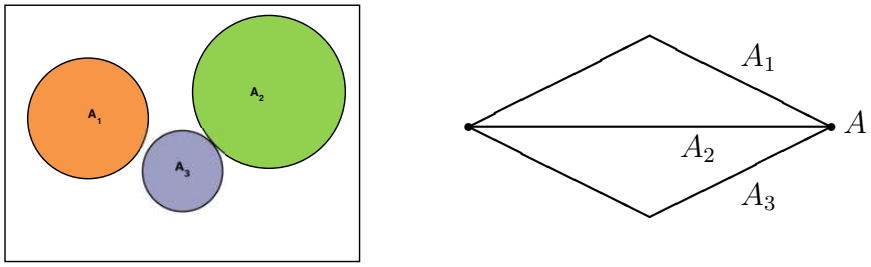


Figure 2.10: On the left, a Venn diagram depicting $A = A_1 \cup A_2 \cup A_3$, and on the right the 'paths' to A.

Both diagrams in Figure 2.10 represent the partition $A = A_1 \cup A_2 \cup A_3$. For the next results, it will be convenient to use diagrams like the one on the right in Figure 2.10, where A_1, A_2 and A_3 are symbolised as different 'paths' to A.

We now develop powerful methods of calculating sums like:

$$P(A) = P(A_1) + P(A_2) + P(A_3).$$

2.9.1 Total probability formula

Suppose $B_1, B_2, \ldots, B_K$ form a partition of the sample space. Then $A \cap B_1, A \cap B_2, \ldots, A \cap B_K$ form a partition of A, as shown in Figure 2.11.

In other words, think of event A as the union of all the $A \cap B_i$s, i.e. of 'all the paths to A via different intervening events B_i'.

To get the probability of A, we now:

1. Apply the chain rule to each of the paths:

$$P(A \cap B_i) = P(A \mid B_i)\,P(B_i).$$

2. Add up the probabilities of the paths:

$$P(A) = \sum_{i=1}^{K} P(A \cap B_i) = \sum_{i=1}^{K} P(A \mid B_i)\,P(B_i).$$

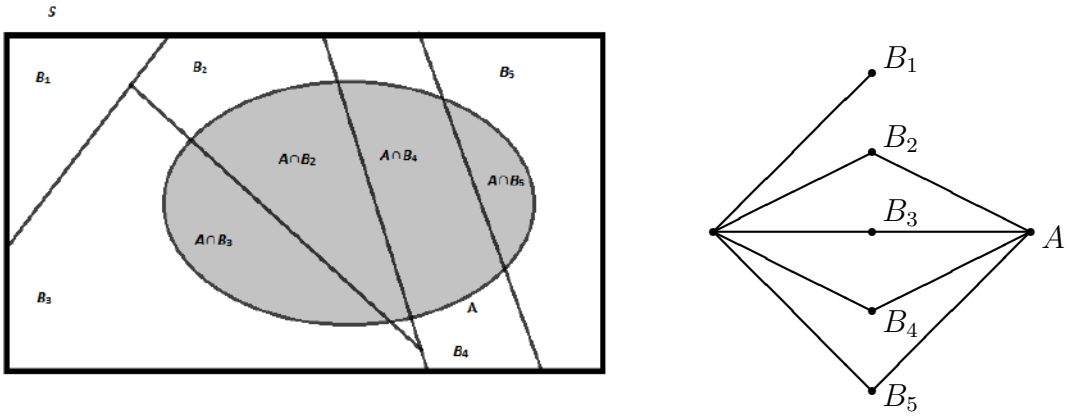
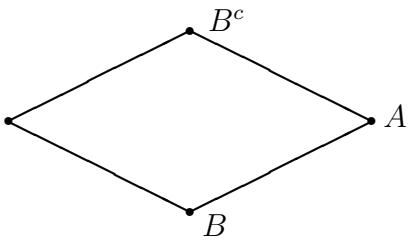


Figure 2.11: On the left, a Venn diagram depicting the set A and the partition of S, and on the right the 'paths' to A.

This is known as the formula of **total probability**. It looks complicated, but it is actually often far easier to use than trying to find $P(A)$ directly.

Example 2.23 Any event B has the property that B and its complement B^c partition the sample space. So if we take $K = 2$, $B_1 = B$ and $B_2 = B^c$ in the formula of total probability, we get:

$$\begin{aligned} P(A) &= P(A \mid B)\, P(B) + P(A \mid B^c)\, P(B^c) \\ &= P(A \mid B)\, P(B) + P(A \mid B^c)\, [1 - P(B)]. \end{aligned}$$



Example 2.24 Suppose that 1 in 10,000 people (0.01%) has a particular disease. A diagnostic test for the disease has 99% *sensitivity*: If a person has the disease, the test will give a positive result with a probability of 0.99. The test has 99% *specificity*: If a person does not have the disease, the test will give a negative result with a probability of 0.99.

Let B denote the presence of the disease, and B^c denote no disease. Let A denote a positive test result. We want to calculate $P(A)$.

The probabilities we need are $P(B) = 0.0001$, $P(B^c) = 0.9999$, $P(A \mid B) = 0.99$ and $P(A \mid B^c) = 0.01$, and therefore:

$$\begin{aligned} P(A) &= P(A \mid B)\, P(B) + P(A \mid B^c)\, P(B^c) \\ &= 0.99 \times 0.0001 + 0.01 \times 0.9999 \\ &= 0.010098. \end{aligned}$$

2.9.2 Bayes' theorem

So far we have considered how to calculate $P(A)$ for an event A which can happen in different ways, 'via' different events $B_1, B_2, \ldots, B_K$.

Now we reverse the question: Suppose we know that A has happened, as shown in Figure 2.12.

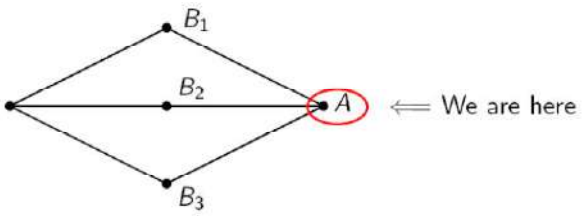


Figure 2.12: Paths to A indicating that A has occurred.

What is the probability that we got there via, say, B_1? In other words, what is the conditional probability $P(B_1 \mid A)$? This situation is depicted in Figure 2.13.

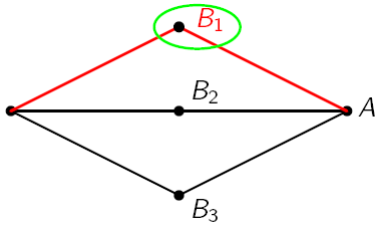


Figure 2.13: A being achieved via B_1.

So we need:

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

and we already know how to get this:

- $P(A \cap B_j) = P(A \mid B_j)\, P(B_j)$ from the chain rule.
- $P(A) = \sum_{i=1}^{K} P(A \mid B_i)\, P(B_i)$ from the total probability formula.

Bayes' theorem

Using the chain rule and the total probability formula, we have:

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

which holds for each B_j, $j = 1, \ldots, K$. This is known as **Bayes' theorem**.

Example 2.25 Continuing with Example 2.24, let B denote the presence of the disease, B^c denote no disease, and A denote a positive test result.

We want to calculate $P(B \mid A)$, i.e. the probability that a person has the disease, given that the person has received a positive test result.

The probabilities we need are:

$$\begin{aligned} P(B) &= 0.0001 & P(B^c) &= 0.9999 \\ P(A \mid B) &= 0.99 \quad \text{and} & P(A \mid B^c) &= 0.01. \end{aligned}$$

Then:

$$P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A \mid B)\, P(B) + P(A \mid B^c)\, P(B^c)} = \frac{0.99 \times 0.0001}{0.010098} \approx 0.0098.$$

Why is this so small? The reason is because most people do not have the disease and the test has a small, but non-zero, false positive rate $P(A \mid B^c)$. Therefore most positive test results are actually *false positives*.

Example 2.26 You are waiting for your bag at the baggage return carousel of an airport. Suppose that you know that there are 200 bags to come from your flight, and you are counting the distinct bags that come out. Suppose that x bags have arrived, and your bag is not among them. What is the probability that your bag will not arrive at all, i.e. that it has been lost (or at least delayed)?

Define $A =$ 'your bag has been lost' and $x =$ 'your bag is not among the first x bags to arrive'. What we want to know is the conditional probability $P(A \mid x)$ for any $x = 0, 1, 2, \ldots, 200$. The conditional probabilities the other way round are:

- $P(x \mid A) = 1$ for all x. If the bag has been lost, it will not arrive!
- $P(x \mid A^c) = (200 - x)/200$ if we assume that bags come out in a completely random order.

Using Bayes' theorem, we get:

$$\begin{aligned} P(A \mid x) &= \frac{P(x \mid A)\, P(A)}{P(x \mid A)\, P(A) + P(x \mid A^c)\, P(A^c)} \\ &= \frac{P(A)}{P(A) + [(200 - x)/200]\,[1 - P(A)]}. \end{aligned}$$

Obviously, $P(A \mid 200) = 1$. If the bag has not arrived when all 200 have come out, it has been lost!

For other values of x we need $P(A)$. This is the general probability that a bag gets lost, before you start observing the arrival of the bags from your particular flight. This kind of probability is known as the **prior probability** of an event A.

Let us assign values to $P(A)$ based on some empirical data. Statistics by the Association of European Airlines (AEA) show how many bags were 'mishandled' per 1,000 *passengers* the airlines carried. This is not exactly what we need (since not all

passengers carry bags, and some have several), but we will use it anyway. In particular, we will compare the results for the best and the worst of the AEA in 2006:

- Air Malta: $P(A) = 0.0044$.
- British Airways: $P(A) = 0.023$.

Figure 2.14 shows a plot of $P(A \mid x)$ as a function of x for these two airlines.

The probabilities are fairly small even for large values of x.

- For Air Malta, $P(A \mid 199) = 0.469$. So even when only 1 bag remains to arrive, the probability is less than 0.5 that your bag has been lost.
- For British Airways, $P(A \mid 199) = 0.825$. Also, we see that $P(A \mid 197) = 0.541$ is the first probability over 0.5.

This is because the baseline probability of lost bags, $P(A)$, is low.

So, the moral of the story is that even when nearly everyone else has collected their bags and left, do not despair!

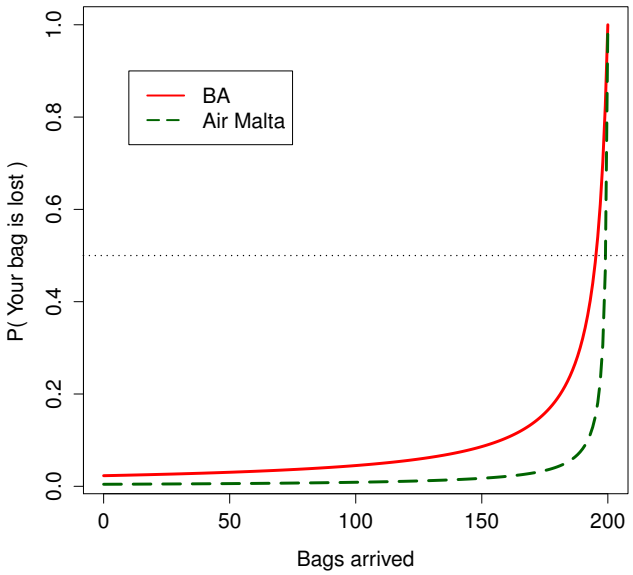


Figure 2.14: Plot of $P(A \mid x)$ as a function of x for the two airlines in Example 2.26, Air Malta and British Airways (BA).

2.10 Overview of chapter

This chapter introduced some formal terminology related to probability. The axioms of probability were introduced, from which various other probability results can be derived. There followed a brief discussion of counting rules (using permutations and combinations). The important concepts of independence and conditional probability were discussed, and Bayes' theorem was derived.

2.11 Key terms and concepts

- Axiom
- Bayes' theorem
- Combination
- Complement
- Conditional probability
- Disjoint
- Elementary outcome
- Empty set
- Experiment
- Event
- Independence
- Intersection
- Mutually exclusive
- Outcome
- Partition
- Permutation
- Probability
- Random experiment
- Sample space
- Set
- Union
- Venn diagram

2.12 Learning activities

1. Why is $S = \{1, 1, 2\}$, not a sensible way to try to define a sample space?

2. Write out all the events for the sample space $S = \{a, b, c\}$. (There are eight of them.)

3. For an event A, work out a simpler way to express the events $A \cap S$, $A \cup S$, $A \cap \emptyset$ and $A \cup \emptyset$.

4. If all elementary outcomes are equally likely, $S = \{a, b, c, d\}$, $A = \{a, b, c\}$ and $B = \{c, d\}$, find $P(A \mid B)$ and $P(B \mid A)$.

5. Suppose that we toss a fair coin twice. The sample space is therefore $S = \{HH, HT, TH, TT\}$, where the elementary outcomes are defined in the obvious way – for instance HT is heads on the first toss and tails on the second toss. Show that if all four elementary outcomes are equally likely, then the events 'heads on the first toss' and 'heads on the second toss' are independent.

6. Show that if A and B are disjoint events, and are also independent, then $P(A) = 0$ or $P(B) = 0$. (Notice that independence and disjointness are not similar ideas.)

7. Write down the condition for the three events A, B and C to be independent.

8. Prove Bayes' theorem from first principles.

9. A statistics teacher knows from past experience that a student who does homework consistently has a probability of 0.95 of passing the examination, whereas a student who does not do homework at all has a probability of 0.30 of passing the examination.

   (a) If 25% of students do their homework consistently, what percentage can expect to pass the examination?

(b) If a student chosen at random from the group gets a pass in the examination, what is the probability that the student had done homework consistently?

10. Plagiarism is a serious problem for assessors of coursework. One check on plagiarism is to compare the coursework with a standard textbook. If the coursework has plagiarised that textbook, then there will be a 95% chance of finding exactly two phrases which are the same in both the coursework and textbook, and a 5% chance of finding three or more phrases which are the same. If the work is not plagiarised, then these probabilities are both 50%.

    Suppose that 5% of coursework is plagiarised. An assessor chooses some coursework at random.

    (a) What is the probability that it has been plagiarised if it has exactly two phrases in the textbook? (Try making a guess before doing the calculation.)

    (b) Repeat (a) for three phrases in the textbook?

    Did you manage to get a roughly correct guess of these results before calculating?

11. A box contains three red balls and two green balls. Two balls are taken from it without replacement.

    (a) What is the probability that none of the balls taken is red?

    (b) Repeat (a) for 1 ball and 2 balls.

    (c) Show that the probability that the first ball taken is red is the same as the probability that the second ball taken is red.

12. Amy, Bob and Claire throw a fair die in that order until a six appears. The person who throws the first six wins. What are their respective chances of winning?

13. In men's singles tennis, matches are played on the best-of-five-sets principle. Therefore the first player to win three sets wins the match, and a match may consist of three, four or five sets. Assuming that two players are perfectly evenly matched, and that sets are independent events, calculate the probabilities that a match lasts three sets, four sets and five sets, respectively.

Solutions to these questions can be found on the VLE in the **ST104b Statistics 2** area at http://my.londoninternational.ac.uk

2.13 Reminder of learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- explain the fundamental ideas of random experiments, sample spaces and events
- list the axioms of probability and be able to derive all the common probability rules from them

- list the formulae for the number of combinations and permutations of k objects out of n, and be able to routinely use such results in problems
- explain conditional probability and the concept of independent events
- prove the law of total probability and apply it to problems where there is a partition of the sample space
- prove Bayes' theorem and apply it to find conditional probabilities.

2.14 Sample examination questions

1. (a) A, B and C are any three events in the sample space S. Prove that:

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

   (b) A and B are events in a sample space S. Show that:

$$P(A \cap B) \leq \frac{P(A) + P(B)}{2} \leq P(A \cup B).$$

2. Suppose A and B are events with $P(A) = p$, $P(B) = 2p$ and $P(A \cup B) = 0.75$.
   (a) Evaluate p and $P(A \mid B)$ if A and B are independent events.
   (b) Evaluate p and $P(A \mid B)$ if A and B are mutually exclusive events.

3. (a) Show that if A and B are independent events in a sample space, then A^c and B^c are also independent.
   (b) Show that if X and Y are mutually exclusive events in a sample space, then X^c and Y^c are not in general mutually exclusive.

4. In a game of tennis, each point is won by one of the two players A and B. The usual rules of scoring for tennis apply. That is, the winner of the game is the player who first scores four points, unless each player has won three points, when deuce is called and play proceeds until one player is two points ahead of the other and hence wins the game.

   A is serving and has a probability of winning any point of 2/3. The result of each point is assumed to be independent of every other point.
   (a) Show that the probability of A winning the game without deuce being called is 496/729.
   (b) Find the probability of deuce being called.
   (c) If deuce is called, show that A's subsequent probability of winning the game is 4/5.
   (d) Hence determine A's overall probability of winning the game.

Solutions to these questions can be found on the VLE in the **ST104b Statistics 2** area at http://my.londoninternational.ac.uk

Chapter 3
Random variables

3.1 Synopsis of chapter content

This chapter introduces the concept of random variables and probability distributions. These distributions are univariate, which means that they are used to model a single numerical quantity. The concepts of expected value and variance are also discussed.

3.2 Aims of the chapter

The aims of this chapter are to:

- be familiar with the concept of random variables
- be able to explain what a probability distribution is
- be able to determine the expected value and variance of a random variable.

3.3 Learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- define a random variable and distinguish it from the values that it takes
- explain the difference between discrete and continuous random variables
- find the mean and the variance of simple random variables whether discrete or continuous
- demonstrate how to proceed and use simple properties of expected values and variances.

3.4 Essential reading

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060] Chapters 4 and 5.

In addition there is essential 'watching' of this chapter's accompanying video tutorials accessible via the **ST104b Statistics 2** area at http://my.londoninternational.ac.uk

3.5 Introduction

In **ST104a Statistics 1**, we considered descriptive statistics for a sample of observations of a variable X. Here we will represent the observations as a sequence of variables, denoted as:

$$X_1, X_2, \ldots, X_n$$

where n is the sample size.

In statistical inference, the observations will be treated as a *sample* drawn at random from a *population*. We will then think of each observation X_i of a variable X as an outcome of an experiment.

- The **experiment** is 'select a unit at random from the population and record its value of X'.
- The **outcome** is the observed value X_i of X.

Because variables X in statistical data are recorded as numbers, we can now focus on experiments where the outcomes are also numbers – *random variables*.

Random variable

A **random variable** is an experiment for which the outcomes are numbers.[1] This means that for a random variable:

- The sample space, S, is the set of real numbers $\mathbb{R}$, or a subset of $\mathbb{R}$.
- The outcomes are numbers in this sample space. Instead of 'outcomes', we often call them the *values* of the random variable.
- Events are sets of numbers (values) in this sample space.

Discrete and continuous random variables

There are two main types of random variables, depending on the nature of S, i.e. the possible values of the random variable.

- A random variable is **continuous** if S is all of $\mathbb{R}$ or some interval(s) of it, for example $[0, 1]$ or $[0, \infty)$.
- A random variable is **discrete** if it is not continuous.[2] More precisely, a discrete random variable takes a finite or countably infinite number of values.

[1] This definition is a bit informal, but it is sufficient for this course.

[2] Strictly speaking, a discrete random variable is not just a random variable which is not continuous as there are many others, such as mixture distributions.

Notation

A random variable is typically denoted by an upper-case letter, for example X (or Y, W, etc.). A specific *value* of a random variable is often denoted by a lower-case letter, for example, x.

Probabilities of values of a random variable are written like this:

- $P(X = x)$ denotes the probability that (the value of) X is x.
- $P(X > 0)$ denotes the probability that X is positive.
- $P(a < X < b)$ denotes the probability that X is between the numbers a and b.

Random variables versus samples

You will notice that many of the quantities we define for random variables are analogous to sample quantities defined in **ST104a Statistics 1**.

Random variable	Sample
Probability distribution	Sample distribution
Mean (expected value)	Sample mean (average)
Variance and standard deviation	Sample variance and standard deviation
Median	Sample median

This is no accident. In statistics, the population is represented as following a probability distribution, and quantities for an observed sample are then used as **estimators** of the analogous quantities for the population.

3.6 Discrete random variables

Example 3.1 The following two examples will be used throughout this chapter.

1. *Number of people living in a randomly selected household in England.*
   - For simplicity, we use the value 8 to represent '8 or more' (because 9 and above are not reported separately in official statistics).
   - This is a discrete random variable, with possible values of 1, 2, 3, 4, 5, 6, 7 and 8.
2. A person throws a basketball repeatedly from the free-throw line, trying to make a basket. Consider the following random variable:

   Number of unsuccessful throws before the first successful throw.
   - The possible values of this are $0, 1, 2, \ldots$.

Probability distribution of a discrete random variable

The **probability distribution** (or just **distribution**) of a discrete random variable X is specified by:

- its possible values x (i.e. its sample space S)
- the probabilities of the possible values, i.e. $P(X = x)$ for all $x \in S$.

So we first need to develop a convenient way of specifying the probabilities.

Example 3.2 Consider the following probability distribution for the household size, X.

Number of people in household (x)	$P(X = x)$
1	0.3002
2	0.3417
3	0.1551
4	0.1336
5	0.0494
6	0.0145
7	0.0034
8	0.0021

Probability function

The **probability function** (pf) of a discrete random variable X, denoted by $p(x)$, is a real-valued function such that for any number x the function is:

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

We can talk of $p(x)$ both as the pf of the random variable X, and as the pf of the probability distribution of X. Both mean the same thing.

Alternative terminology: the pf of a discrete random variable is also often called the **probability mass function** (pmf).

Alternative notation: instead of $p(x)$, the pf is also often denoted by, for example, $p_X(x)$ – especially when it is necessary to indicate clearly to which random variable the function corresponds.

Necessary conditions for a probability function

To be a pf of a discrete random variable X with sample space S, a function $p(x)$ must satisfy the following conditions:

1. $p(x) \geq 0$ for all real numbers x.
2. $\sum_{x_i \in S} p(x_i) = 1$, i.e. the sum of probabilities of all possible values of X is 1.

The pf is defined for *all* real numbers x, but $p(x) = 0$ for any $x \notin S$, i.e. for any value x that is not one of the possible values of X.

Example 3.3 Continuing Example 3.2, here we can simply list all the values:

$$p(x) = \begin{cases} 0.3002, & \text{if } x = 1 \\ 0.3417, & \text{if } x = 2 \\ 0.1551, & \text{if } x = 3 \\ 0.1336, & \text{if } x = 4 \\ 0.0494, & \text{if } x = 5 \\ 0.0145, & \text{if } x = 6 \\ 0.0034, & \text{if } x = 7 \\ 0.0021, & \text{if } x = 8 \\ 0 & \text{otherwise.} \end{cases}$$

These are clearly all non-negative, and their sum is $\sum_{x=1}^{8} p(x) = 1$.

A graphical representation of the pf is shown in Figure 3.1.

For the next example, we need to remember the following results from mathematics, concerning sums of geometric series: If $r \neq 1$, then:

$$\sum_{x=0}^{n-1} a\, r^x = \frac{a(1 - r^n)}{1 - r}$$

and if $|r| < 1$, then:

$$\sum_{x=0}^{\infty} a\, r^x = \frac{a}{1 - r}.$$

Example 3.4 In the basketball example, the number of possible values is infinite, so we cannot simply list the values of the pf. So we try to express it as a formula. Suppose that:

- the probability of a successful throw is π at each throw, and therefore the probability of an unsuccessful throw is $1 - \pi$

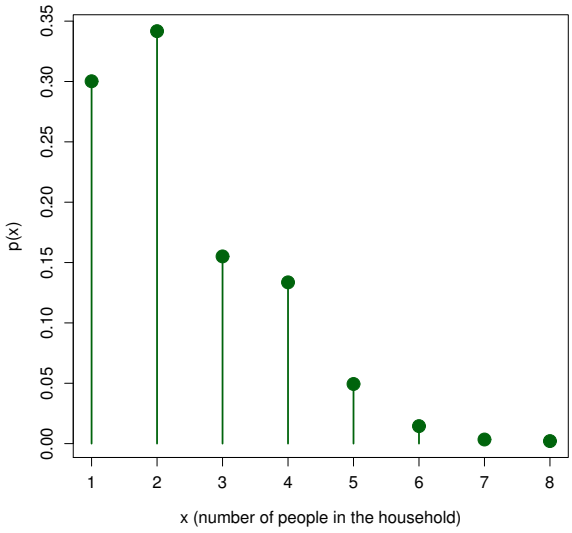


Figure 3.1: Probability function for Example 3.3.

- outcomes of different throws are independent.

Then the probability that the first success occurs after x failures is the probability of a sequence of x failures followed by a success, i.e. the probability is:

$$(1-\pi)^x \pi.$$

So the pf of the random variable X (the number of failures before the first success) is:

$$p(x) = \begin{cases} (1-\pi)^x \pi & \text{for } \; x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where $0 \leq \pi \leq 1$. Let us check that (3.1) satisfies the conditions for a pf.

- Clearly, $p(x) \geq 0$ for all x, since $\pi \geq 0$ and $1 - \pi \geq 0$.
- Using the sum to infinity of a geometric series, we get:

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty} (1-\pi)^x \pi = \pi \sum_{x=0}^{\infty} (1-\pi)^x = \pi \cdot \frac{1}{1-(1-\pi)} = \frac{\pi}{\pi} = 1.$$

The expression of the pf involves a **parameter** π (the probability of a successful throw), a number for which we can choose different values. This defines a whole 'family' of individual distributions, one for each choice of π. For example, Figure 3.2 shows values of $p(x)$ for two values of π reflecting good and poor free-throw shooters.

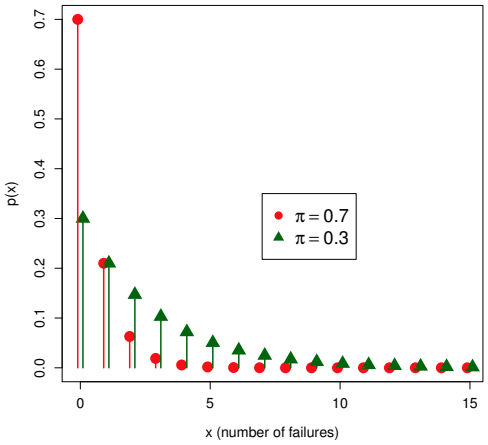


Figure 3.2: Probability function for Example 3.4. $\pi = 0.7$: a fairly good free-throw shooter. $\pi = 0.3$: a pretty poor free-throw shooter.

The cumulative distribution function (cdf)

Another way to specify a probability distribution is to give its **cumulative distribution function (cdf)**, (or just simply **distribution function**).

Cumulative distribution function (cdf)

The cdf is denoted $F(x)$ (or $F_X(x)$) and defined as:

$$F(x) = P(X \leq x) \quad \text{for all real numbers } x.$$

For a discrete random variable it is given by:

$$F(x) = \sum_{x_i \in S,\, x_i \leq x} p(x_i)$$

i.e. the sum of the probabilities of those possible values of X that are less than or equal to x.

Example 3.5 Continuing with the household size example, values of $F(x)$ at all possible values of X are:

Number of people in household (x)	$p(x)$	$F(x)$
1	0.3002	0.3002
2	0.3417	0.6419
3	0.1551	0.7970
4	0.1336	0.9306
5	0.0494	0.9800
6	0.0145	0.9945
7	0.0034	0.9979
8	0.0021	1.0000

These are shown in graphical form in Figure 3.3.

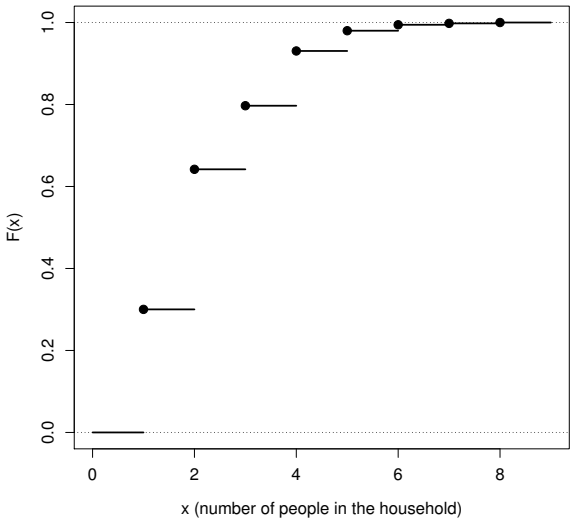


Figure 3.3: Cumulative distribution function for Example 3.5.

Example 3.6 In the basketball example, $p(x) = (1-\pi)^x \pi$ for $x = 0, 1, 2, \ldots$. We can calculate a simple formula for the cdf, using the sum of a geometric series. Since, for any non-negative integer y, we obtain:

$$\sum_{x=0}^{y} p(x) = \sum_{x=0}^{y} (1-\pi)^x \pi = \pi \sum_{x=0}^{y} (1-\pi)^x = \pi \cdot \frac{1-(1-\pi)^{y+1}}{1-(1-\pi)} = 1-(1-\pi)^{y+1}$$

we can write:

$$F(x) = \begin{cases} 0 & \text{when } x < 0 \\ 1-(1-\pi)^{x+1} & \text{when } x = 0, 1, 2, \ldots. \end{cases}$$

The cdf is shown in graphical form in Figure 3.4.

Properties of the cdf for discrete distributions

The cdf $F(x)$ of a discrete random variable X is a **step function** such that:

- $F(x)$ remains constant in all intervals between possible values of X.
- At a possible value x_i of X, $F(x)$ jumps up by the amount $p(x_i) = P(X = x_i)$.
- At such an x_i, the value of $F(x_i)$ is the value at the top of the jump (i.e. $F(x)$ is *right-continuous*).

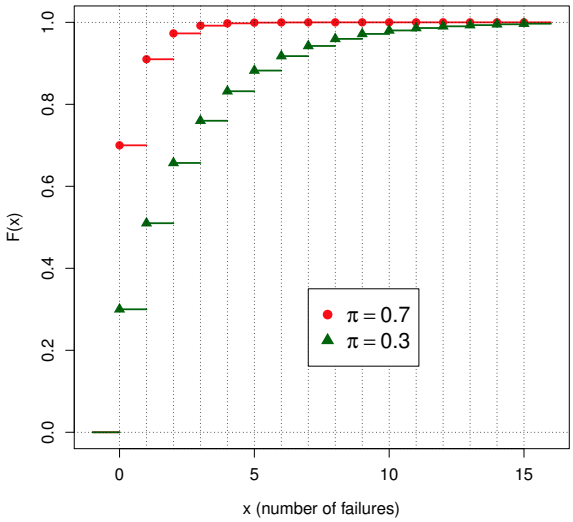


Figure 3.4: Cumulative distribution function for Example 3.6.

General properties of the cdf

These hold for both discrete and continuous random variables:

1. $0 \leq F(x) \leq 1$ for all x (since $F(x)$ is a probability).
2. $F(x) \to 0$ as $x \to -\infty$, and $F(x) \to 1$ as $x \to \infty$.
3. $F(x)$ is a non-decreasing function, i.e. if $x_1 < x_2$, then $F(x_1) \leq F(x_2)$.
4. For any $x_1 < x_2$, $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$.

Either the pf or the cdf can be used to calculate the probabilities of any events for a discrete random variable.

Example 3.7 Continuing with the household size example (for the probabilities, see Example 3.5), then:

- $P(X = 1) = p(1) = F(1) = 0.3002$.
- $P(X = 2) = p(2) = F(2) - F(1) = 0.3417$.
- $P(X \leq 2) = p(1) + p(2) = F(2) = 0.6419$.
- $P(X = 3 \text{ or } 4) = p(3) + p(4) = F(4) - F(2) = 0.2887$.
- $P(X > 5) = p(6) + p(7) + p(8) = 1 - F(5) = 0.0200$.
- $P(X \geq 5) = p(5) + p(6) + p(7) + p(8) = 1 - F(4) = 0.0694$.

Properties of a discrete random variable

Let X be a discrete random variable with sample space S and pf $p(x)$.

Expected value of a discrete random variable

The **expected value** (or **mean**) of X is denoted $\mathrm{E}(X)$, and defined as:

$$\mathrm{E}(X) = \sum_{x_i \in S} x_i \, p(x_i).$$

This can also be written more concisely as $\mathrm{E}(X) = \sum_{x} x \, p(x)$ or $\mathrm{E}(X) = \sum x \, p(x)$.

We can talk of $\mathrm{E}(X)$ as the expected value of both the random variable X, and of the probability distribution of X.

Alternative notation: Instead of $\mathrm{E}(X)$, the symbol μ (the lower-case Greek letter '*mu*'), or μ_X, is often used.

Expected value versus sample mean

The mean (expected value) $\mathrm{E}(X)$ of a probability distribution is analogous to the sample mean (average) $\bar{X}$ of a sample distribution.

This is easiest to see when the sample space is finite. Suppose the random variable X can have K different values $X_1, \ldots, X_K$, and their frequencies in a sample are $f_1, \ldots, f_K$, respectively. Then the sample mean of X is:

$$\bar{X} = \frac{f_1 x_1 + \cdots + f_K x_K}{f_1 + \cdots + f_K} = x_1 \, \widehat{p}(x_1) + \cdots + x_K \, \widehat{p}(x_K) = \sum_{i=1}^{K} x_i \, \widehat{p}(x_i)$$

where:

$$\widehat{p}(x_i) = \frac{f_i}{\sum_{i=1}^{K} f_i}$$

are the **sample proportions** of the values x_i.

The expected value of the random variable X is:

$$\mathrm{E}(X) = x_1 \, p(x_1) + \cdots + x_K \, p(x_K) = \sum_{i=1}^{K} x_i \, p(x_i).$$

So $\bar{X}$ uses the sample proportions $\widehat{p}(x_i)$, whereas $\mathrm{E}(X)$ uses the population probabilities $p(x_i)$.

Example 3.8 Continuing with the household size example:

Number of people in household (x)	$p(x)$	$x\,p(x)$
1	0.3002	0.3002
2	0.3417	0.6834
3	0.1551	0.4653
4	0.1336	0.5344
5	0.0494	0.2470
6	0.0145	0.0870
7	0.0034	0.0238
8	0.0021	0.0168
Sum		2.3579
		$= \mathrm{E}(X)$

The expected number of people in a randomly selected household is 2.36.

Example 3.9 For the basketball example, $p(x) = (1-\pi)^x\,\pi$ for $x = 0, 1, 2, \ldots$, and 0 otherwise.

The expected value of X is then:

$$\begin{aligned}
\mathrm{E}(X) = \sum_{x_i \in S} x_i\,p(x_i) &= \sum_{x=0}^{\infty} x\,(1-\pi)^x\,\pi \\
\text{(starting from } x=1) \qquad &= \sum_{x=1}^{\infty} x\,(1-\pi)^x\,\pi \\
&= (1-\pi)\sum_{x=1}^{\infty} x\,(1-\pi)^{x-1}\,\pi \\
\text{(using } y = x-1) \qquad &= (1-\pi)\sum_{y=0}^{\infty} (y+1)(1-\pi)^y\,\pi \\
&= (1-\pi)\left[\underbrace{\sum_{y=0}^{\infty} y\,(1-\pi)^y\,\pi}_{=\mathrm{E}(X)} + \underbrace{\sum_{y=0}^{\infty} (1-\pi)^y\,\pi}_{=1}\right] \\
&= (1-\pi)\,[\mathrm{E}(X)+1] \\
&= (1-\pi)\,\mathrm{E}(X) + (1-\pi)
\end{aligned}$$

from which we can solve:

$$\mathrm{E}(X) = \frac{1-\pi}{1-(1-\pi)} = \frac{1-\pi}{\pi}.$$

So, for example:

- $\mathrm{E}(X) = 0.3/0.7 = 0.42$ when $\pi = 0.7$.
- $\mathrm{E}(X) = 0.7/0.3 = 2.33$ when $\pi = 0.3$.

So, before scoring a basket, a good free-throw shooter (with $\pi = 0.7$) misses on average about 0.42 shots, and a poor shooter (with $\pi = 0.3$) misses on average about 2.33 shots.

Expected values of functions of a random variable

Let $g(X)$ be a function ('transformation') of a discrete random variable X. This is also a random variable, and its expected value is:

$$\mathrm{E}(g(X)) = \sum g(x)\, p_X(x)$$

where $p_X(x) = p(x)$ is the probability function of X.

Example 3.10 The expected value of the square of X is:

$$\mathrm{E}(X^2) = \sum x^2\, p(x).$$

In general:

$$\mathrm{E}[g(X)] \neq g[\mathrm{E}(X)]$$

when $g(X)$ is a *nonlinear* function of X.

Example 3.11 Note that:

$$\mathrm{E}(X^2) \neq (\mathrm{E}(X))^2 \quad \text{and} \quad \mathrm{E}\left(\frac{1}{X}\right) \neq \frac{1}{\mathrm{E}(X)}.$$

Expected values of linear transformations

Suppose X is a random variable and a and b are **constants**, i.e. known numbers that are not random variables. Then:

$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b.$$

Proof:

$$\begin{aligned} \mathrm{E}(aX+b) &= \sum_x (ax+b)\,p(x) \\ &= \sum_x ax\,p(x) + \sum_x b\,p(x) \\ &= a\sum_x x\,p(x) + b\sum_x p(x) \\ &= a\mathrm{E}(X) + b \end{aligned}$$

where the last step follows from:

i. $\sum_x x\,p(x) = \mathrm{E}(X)$, by definition of $\mathrm{E}(X)$.

ii. $\sum_x p(x) = 1$, by definition of the probability function. □

A special case of the result:

$$\mathrm{E}(aX+b) = a\mathrm{E}(X) + b$$

is obtained when $a = 0$, which gives:

$$\mathrm{E}(b) = b.$$

That is, the expected value of a constant is the constant itself.

Variance and standard deviation of a discrete random variable

The **variance** of a discrete random variable X is defined as:

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] = \sum_x (x - \mathrm{E}(X))^2\,p(x).$$

The **standard deviation** of X is $\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$.

Both $\mathrm{Var}(X)$ and $\mathrm{sd}(X)$ are always ≥ 0. Both are measures of the dispersion (variation) of the distribution of X.

Alternative notation: The variance is often denoted σ^2 ('sigma-squared') and standard deviation by σ ('sigma').

An alternative formula: The variance can also be calculated as:

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

This will be proved later.

Example 3.12 Continuing with the household size example:

x	$p(x)$	$x\,p(x)$	$(x-\mathrm{E}(X))^2$	$(x-\mathrm{E}(X))^2\,p(x)$	x^2	$x^2\,p(x)$
1	0.3002	0.3002	1.844	0.554	1	0.300
2	0.3417	0.6834	0.128	0.044	4	1.367
3	0.1551	0.4653	0.412	0.064	9	1.396
4	0.1336	0.5344	2.696	0.360	16	2.138
5	0.0494	0.2470	6.981	0.345	25	1.235
6	0.0145	0.0870	13.265	0.192	36	0.522
7	0.0034	0.0238	21.549	0.073	49	0.167
8	0.0021	0.0168	31.833	0.067	64	0.134
$\sum$		2.3579		1.699		7.259
		$=\mathrm{E}(X)$		$=\mathrm{Var}(X)$		$=\mathrm{E}(X^2)$

$\mathrm{Var}(X) = \mathrm{E}[(X-\mathrm{E}(X))^2] = 1.699 = 7.259 - (2.358)^2 = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$ and $\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{1.699} = 1.30$.

Example 3.13 For the basketball example, $p(x) = (1-\pi)^x\,\pi$ for $x = 0, 1, 2, \ldots$, and 0 otherwise. It can be shown (although the proof is beyond the scope of the course) that for this distribution:

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

In the two cases we have used as examples:

- $\mathrm{Var}(X) = 0.3/(0.7)^2 = 0.61$ and $\mathrm{sd}(X) = 0.78$ when $\pi = 0.7$.
- $\mathrm{Var}(X) = 0.7/(0.3)^2 = 7.78$ and $\mathrm{sd}(X) = 2.79$ when $\pi = 0.3$.

So the *variation* in how many free throws a poor shooter misses before the first success is much higher than the variation for a good shooter.

Variances of linear transformations

If X is a random variable and a and b are constants, then:

$$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X).$$

Proof:

$$\begin{aligned}
\mathrm{Var}(aX+b) &= \mathrm{E}\left[((aX+b) - \mathrm{E}(aX+b))^2\right] \\
&= \mathrm{E}\left[(aX+b-a\,\mathrm{E}(X)-b)^2\right] \\
&= \mathrm{E}\left[(aX-a\,\mathrm{E}(X))^2\right] \\
&= \mathrm{E}\left[a^2(X-\mathrm{E}(X))^2\right] \\
&= a^2\,\mathrm{E}\left[(X-\mathrm{E}(X))^2\right] \\
&= a^2\,\mathrm{Var}(X).
\end{aligned}$$

Therefore, $\text{sd}(aX + b) = |a| \, \text{sd}(X)$. $\square$

If $a = 0$, this gives:

$$\text{Var}(b) = 0.$$

That is, the variance of a constant is 0. The converse also holds – if a random variable has a variance of 0, it is actually a constant.

Example 3.14 For further practice, let us consider a discrete random variable X which has possible values $0, 1, 2 \ldots, n$, where n is a known positive integer, and X has the following probability function:

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

where $\binom{n}{x} = n!/[x!(n - x)!]$ denotes the binomial coefficient, and π is a probability parameter which can have values $0 \leq \pi \leq 1$.

A random variable like this follows the **binomial distribution**. We will discuss its motivation and uses later in the next chapter.

Here, we consider the following tasks for this distribution:

- Show that $p(x)$ satisfies the conditions for a probability function.
- Calculate probabilities from $p(x)$.
- Write down the cumulative distribution function.
- Derive the expected value, $\text{E}(X)$.

To show that $p(x)$ is a probability function, we need to show the following:

1. $p(x) \geq 0$ for all x. This is clearly true, since $x \geq 0$, $\pi \geq 0$ and $1 - \pi \geq 0$.
2. $\sum_{x=0}^{n} p(x) = 1$. This is easiest to show by using the *binomial theorem*, which states that, for any integer $n \geq 0$ and any real numbers y and z, then:

$$(y + z)^n = \sum_{x=0}^{n} \binom{n}{x} y^x \, z^{n-x}. \tag{3.2}$$

   If we choose $y = \pi$ and $z = 1 - \pi$ in (3.2), we get:

$$1 = 1^n = [\pi + (1 - \pi)]^n = \sum_{x=0}^{n} \binom{n}{x} \pi^x (1 - \pi)^{n-x} = \sum_{x=0}^{n} p(x).$$

   This does not simplify into a simple formula, so we just calculate the values from the definition, by summation.

At the values $x = 0, 1, \ldots, n$, the value of the cdf is:

$$F(x) = P(X \leq x) = \sum_{y=0}^{x} \binom{n}{y} \pi^y (1 - \pi)^{n-y}.$$

Since X is a discrete random variable, $F(x)$ is a step function. For $\mathrm{E}(X)$, we have:

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{x=0}^{n} x \binom{n}{x} \pi^x (1-\pi)^{n-x} \\
&= \sum_{x=1}^{n} x \binom{n}{x} \pi^x (1-\pi)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)![(n-1)-(x-1)]!} \pi \, \pi^{x-1} (1-\pi)^{n-x} \\
&= n\pi \sum_{x=1}^{n} \binom{n-1}{x-1} \pi^{x-1} (1-\pi)^{n-x} \\
&= n\pi \sum_{y=0}^{n-1} \binom{n-1}{y} \pi^{y} (1-\pi)^{(n-1)-y} \\
&= n\pi \cdot 1 \\
&= n\pi
\end{aligned}$$

where $y = x - 1$, and the last summation is over all the values of the pf of another binomial distribution, this time with possible values $0, 1, \ldots, n-1$ and probability parameter π.

3.7 Continuous random variables

A random variable (and its probability distribution) is **continuous** if it can have an uncountably infinite number of possible values.[3]

- In other words, the set of possible values (sample space) is the real numbers $\mathbb{R}$, or one or more intervals in $\mathbb{R}$.

Example 3.15 An example of a continuous random variable, used here as an approximating model, is the size of claim made on an insurance policy (i.e. a claim by the customer to the insurance company), in £000s.

- Suppose the policy has a deductible of £999, so all claims are at least £1,000.
- The possible values of this random variable are therefore $\{x \,|\, x \geq 1\}$.

Most of the concepts introduced for discrete random variables have exact or approximate analogies for continuous random variables, and many results are the same

[3]Strictly speaking, having an uncountably infinite number of possible values does not necessarily imply that it is a continuous random variable. For example, the Cantor distribution (not covered in **ST104b Statistics 2**) is neither a discrete nor an absolutely continuous probability distribution, nor is it a mixture of these. However, we will not consider this matter any further in this course.

for both types. But there are some differences in the details. The most obvious difference is that wherever in the discrete case there are *sums* over the possible values of the random variable, in the continuous case these are *integrals*.

Probability density function (pdf)

For a continuous random variable X, the probability function is replaced by the **probability density function** (pdf), denoted as $f(x)$ [or $f_X(x)$].

Example 3.16 Continuing the insurance example in Example 3.15, we consider a pdf of the following form:

$$f(x) = \begin{cases} 0 & \text{when } x < k \\ \alpha\, k^\alpha / x^{\alpha+1} & \text{when } x \geq k \end{cases}$$

where $\alpha > 0$ is a parameter, and $k > 0$ (the smallest possible value of X) is a known number. In our example, $k = 1$ (due to the deductible). A probability distribution with this pdf is known as the **Pareto distribution**. A graph of this pdf when $\alpha = 2.2$ is shown in Figure 3.5.

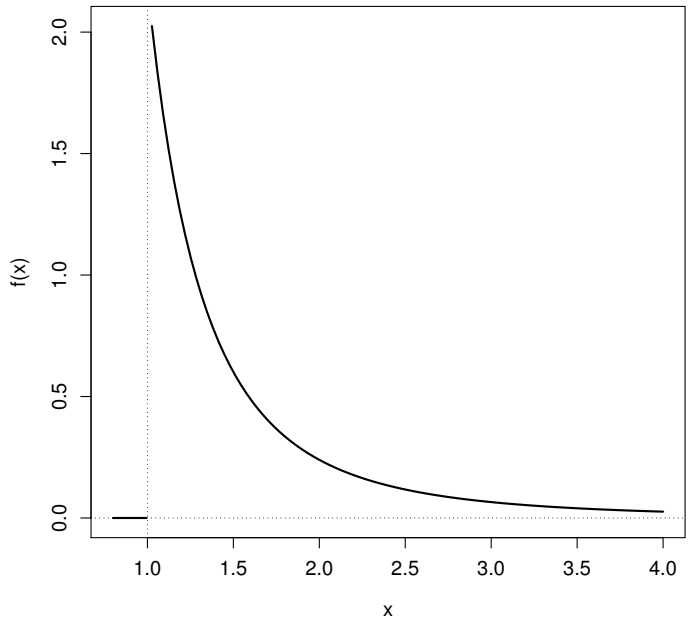


Figure 3.5: Probability density function for Example 3.16.

Unlike for probability functions of discrete random variables, in the continuous case values of the probability density function are not probabilities of individual values, i.e. $f(x) \neq P(X = x)$. In fact, for a continuous distribution:

$$P(X = x) = 0 \quad \text{for all } x. \tag{3.3}$$

That is, the probability that X has any particular value *exactly* is always 0.

Because of (3.3), with a continuous distribution we do not need to be very careful about differences between $<$ and $\leq$, and between $>$ and $\geq$. Therefore, the following probabilities are all equal:

$$P(a < X < b), \quad P(a \leq X \leq b), \quad P(a < X \leq b) \quad \text{and} \quad P(a \leq X < b).$$

Probabilities of intervals for continuous random variables

Integrals of the pdf give probabilities of **intervals** of values:

$$P(a < X \leq b) = \int_a^b f(x)\,\mathrm{d}x$$

for any two numbers $a < b$.

In other words, the probability that the value of X is between a and b is the area under $f(x)$ between a and b. Here a can also be $-\infty$, and/or b can be $+\infty$.

Example 3.17 In Figure 3.6, the shaded area is $P(1.5 < X \leq 3) = \int_{1.5}^{3} f(x)\,\mathrm{d}x$.

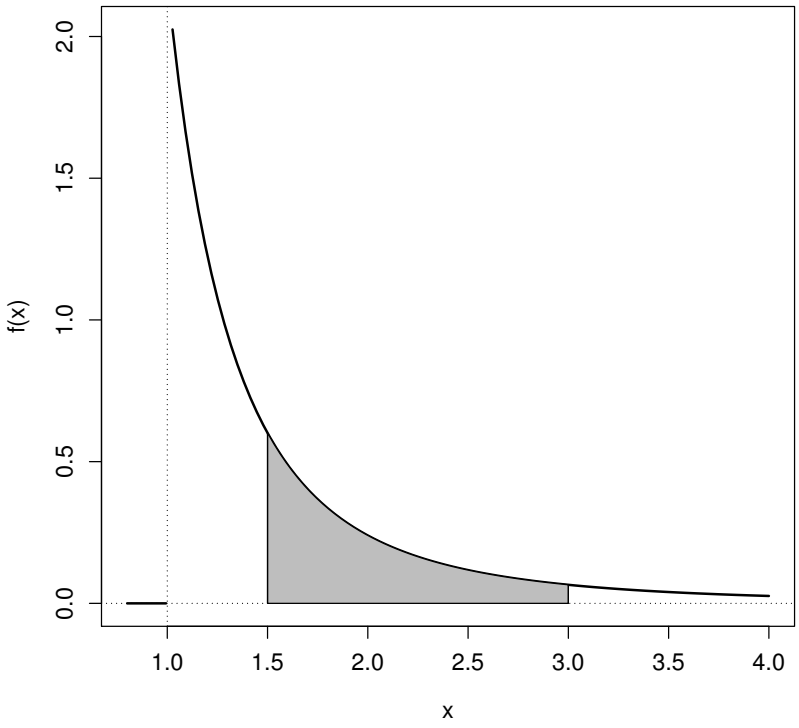


Figure 3.6: Probability density function showing $P(1.5 < X \leq 3)$.

Properties of pdfs

The pdf $f(x)$ of any continuous random variable must satisfy the following conditions:

1.
$$f(x) \geq 0 \quad \text{for all } x.$$

2.
$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$$

These are analogous to the conditions for probability functions of discrete distributions.

Example 3.18 Continuing with the insurance example, we check that the conditions hold for the pdf:

$$f(x) = \begin{cases} 0 & \text{when } x < k \\ \alpha\, k^{\alpha}/x^{\alpha+1} & \text{when } x \geq k \end{cases}$$

where $\alpha > 0$ and $k > 0$.

1. Clearly, $f(x) \geq 0$ for all x, since $\alpha > 0$, $k^{\alpha} > 0$ and $x^{\alpha+1} \geq k^{\alpha+1} > 0$.

2.
$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_{k}^{\infty} \frac{\alpha\, k^{\alpha}}{x^{\alpha+1}}\,\mathrm{d}x &= \alpha\, k^{\alpha} \int_{k}^{\infty} x^{-\alpha-1}\,\mathrm{d}x \\ &= \alpha\, k^{\alpha} \cdot \left(\frac{1}{-\alpha}\right) \left[x^{-\alpha}\right]_{k}^{\infty} \\ &= (-k^{\alpha}) \cdot (0 - k^{-\alpha}) \\ &= 1. \end{aligned}$$

Cumulative distribution function

The **cumulative distribution function** (cdf) of a continuous random variable X is defined exactly as for discrete random variables, i.e. the cdf is:

$$F(x) = P(X \leq x) \quad \text{for all real numbers } x.$$

The general properties of the cdf stated previously also hold for continuous distributions. The cdf of a continuous distribution is not a step function, so results on discrete-specific properties do not hold in the continuous case. A continuous cdf is a smooth, continuous function of x.

Relationship between the cdf and pdf

The cdf is obtained from the pdf through integration:

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t \qquad \text{for all } x.$$

The pdf is obtained from the cdf through differentiation:

$$f(x) = F'(x).$$

Example 3.19 Continuing the insurance example:

$$\begin{aligned}
\int_{-\infty}^{x} f(t)\,\mathrm{d}t &= \int_{k}^{x} \frac{\alpha\, k^{\alpha}}{t^{\alpha+1}}\,\mathrm{d}t \\
&= (-k^{\alpha}) \int_{k}^{x} (-\alpha)\, t^{-\alpha-1}\,\mathrm{d}t \\
&= (-k^{\alpha}) \left[t^{-\alpha}\right]_{k}^{x} \\
&= (-k^{\alpha})(x^{-\alpha} - k^{-\alpha}) \\
&= 1 - k^{\alpha}\, x^{-\alpha} \\
&= 1 - (k/x)^{\alpha}.
\end{aligned}$$

Therefore:

$$F(x) = \begin{cases} 0 & \text{when } \; x < k \\ 1 - (k/x)^{\alpha} & \text{when } \; x \geq k. \end{cases} \tag{3.4}$$

If we were given (3.4), we could obtain the pdf by differentiation, since $F'(x) = 0$ when $x < k$, and:

$$F'(x) = -k^{\alpha}\,(-\alpha)\, x^{-\alpha-1} = \frac{\alpha\, k^{\alpha}}{x^{\alpha+1}} \qquad \text{when } x \geq k.$$

A plot of the cdf is shown in Figure 3.7.

Probabilities from cdfs and pdfs

Since $P(X \leq x) = F(x)$, it follows that $P(X > x) = 1 - F(x)$. In general, for any two numbers $a < b$, we have:

$$P(a < X \leq b) = \int_{a}^{b} f(x)\,\mathrm{d}x = F(b) - F(a).$$

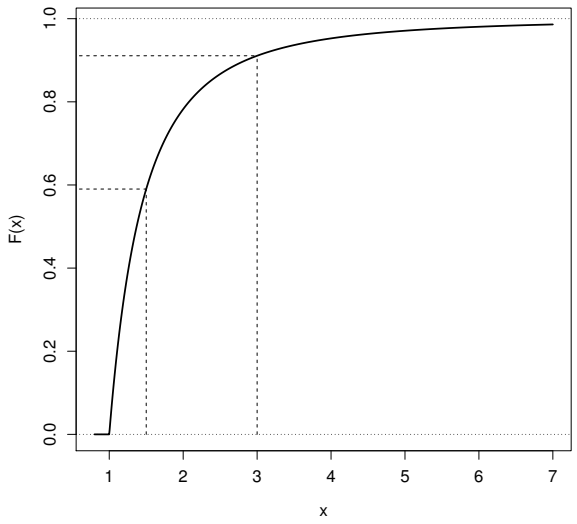


Figure 3.7: Cumulative distribution function for Example 3.19.

Example 3.20 Continuing with the insurance example (with $k = 1$ and $\alpha = 2.2$), then:

$$\begin{aligned} P(X \leq 1.5) &= F(1.5) = 1 - (1/1.5)^{2.2} \approx 0.59 \\ P(X \leq 3) &= F(3) = 1 - (1/3)^{2.2} \approx 0.91 \\ P(X > 3) &= 1 - F(3) \approx 1 - 0.91 = 0.09 \\ P(1.5 \leq X \leq 3) &= F(3) - F(1.5) \approx 0.91 - 0.59 = 0.32. \end{aligned}$$

Example 3.21 Consider now a continuous random variable with the following pdf:

$$f(x) = \begin{cases} \lambda \, \mathrm{e}^{-\lambda x} & \text{for } \; x > 0 \\ 0 & \text{for } \; x \leq 0 \end{cases} \tag{3.5}$$

where $\lambda > 0$ is a parameter. This is the pdf of the **exponential distribution**. The uses of this distribution will be discussed in the next chapter.

Since:

$$\int_0^x \lambda \, \mathrm{e}^{-\lambda t} \, \mathrm{d}t = - \left[\mathrm{e}^{-\lambda t} \right]_0^x = 1 - \mathrm{e}^{-\lambda x}$$

the cdf of the exponential distribution is:

$$F(x) = \begin{cases} 0 & \text{for } \; x \leq 0 \\ 1 - \mathrm{e}^{-\lambda x} & \text{for } \; x > 0. \end{cases}$$

We now show that (3.5) satisfies the conditions for a pdf.

1. Since $\lambda > 0$ and $\mathrm{e}^a > 0$ for any a, $f(x) \geq 0$ for all x.

2. Since we have just done the integration to derive the cdf $F(x)$, we can also use it to show that $f(x)$ integrates to one. This follows from:

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = P(-\infty < X < \infty) = \lim_{x\to\infty} F(x) - \lim_{x\to-\infty} F(x)$$

which here is $\lim_{x\to\infty}(1-\mathrm{e}^{-\lambda x}) - 0 = (1-0) - 0 = 1$.

Expected value and variance of a continuous distribution

Suppose X is a continuous random variable with pdf $f(x)$. Definitions of its expected value, the expected value of any transformation $g(X)$, variance and standard deviation are the same as for discrete distributions, except that summation is replaced by integration:

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} x\,f(x)\,\mathrm{d}x \\
\mathrm{E}[g(X)] &= \int_{-\infty}^{\infty} g(x)\,f(x)\,\mathrm{d}x \\
\mathrm{Var}(X) = \mathrm{E}[(X-\mathrm{E}(X))^2] &= \int_{-\infty}^{\infty} (x-\mathrm{E}(X))^2\,f(x)\,\mathrm{d}x = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 \\
\mathrm{sd}(X) &= \sqrt{\mathrm{Var}(X)}.
\end{aligned}$$

Example 3.22 For the Pareto distribution, introduced in Example 3.16, we have:

$$\begin{aligned}
\mathrm{E}(X) = \int_{-\infty}^{\infty} x\,f(x)\,\mathrm{d}x &= \int_{k}^{\infty} x\,f(x)\,\mathrm{d}x \\
&= \int_{k}^{\infty} x\cdot\frac{\alpha\,k^{\alpha}}{x^{\alpha+1}}\,\mathrm{d}x \\
&= \int_{k}^{\infty} \frac{\alpha\,k^{\alpha}}{x^{\alpha}}\,\mathrm{d}x \\
&= \left(\frac{\alpha\,k}{\alpha-1}\right)\underbrace{\int_{k}^{\infty} \frac{(\alpha-1)\,k^{\alpha-1}}{x^{(\alpha-1)+1}}\,\mathrm{d}x}_{=1} \\
&= \frac{\alpha\,k}{\alpha-1} \quad (\text{if } \alpha > 1).
\end{aligned}$$

Here the last step follows because the last integrand has the form of the Pareto pdf with parameter $\alpha-1$, so its integral from k to ∞ is 1. This integral converges only if $\alpha - 1 > 0$, i.e. if $\alpha > 1$.

Similarly:

$$\begin{aligned}
\mathrm{E}(X^2) = \int_k^\infty x^2\, f(x)\,\mathrm{d}x &= \int_k^\infty x^2 \cdot \frac{\alpha\, k^\alpha}{x^{\alpha+1}}\,\mathrm{d}x \\
&= \int_k^\infty \frac{\alpha\, k^\alpha}{x^{\alpha-1}}\,\mathrm{d}x \\
&= \left(\frac{\alpha\, k^2}{\alpha-2}\right) \underbrace{\int_k^\infty \frac{(\alpha-2)\, k^{\alpha-2}}{x^{(\alpha-2)+1}}\,\mathrm{d}x}_{=1} \\
&= \frac{\alpha\, k^2}{\alpha-2} \quad (\text{if } \alpha > 2)
\end{aligned}$$

and therefore:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{\alpha\, k^2}{\alpha-2} - \frac{\alpha^2\, k^2}{(\alpha-1)^2} = \left(\frac{k}{\alpha-1}\right)^2 \frac{\alpha}{\alpha-2}.$$

In our insurance example, where $k = 1$ and $\alpha = 2.2$, we have:

$$\mathrm{E}(X) = \frac{2.2 \times 1}{2.2-1} \approx 1.8 \quad \text{and} \quad \mathrm{Var}(X) = \left(\frac{1}{2.2-1}\right)^2 \times \frac{2.2}{2.2-2} \approx 7.6.$$

Example 3.23 Consider the exponential distribution introduced in Example 3.21. To find $\mathrm{E}(X)$ we can use integration by parts by considering $x\,\lambda\,\mathrm{e}^{-\lambda x}$ as the product of the functions $f = x$ and $g' = \lambda\,\mathrm{e}^{-\lambda x}$ (so that $g = -\mathrm{e}^{-\lambda x}$). Then:

$$\begin{aligned}
\mathrm{E}(X) = \int_0^\infty x\,\lambda\,\mathrm{e}^{-\lambda x}\,\mathrm{d}x &= \left[-x\,\mathrm{e}^{-\lambda x}\right]_0^\infty - \int_0^\infty -\mathrm{e}^{-\lambda x}\,\mathrm{d}x \\
&= \left[-x\,\mathrm{e}^{-\lambda x}\right]_0^\infty - (1/\lambda)\left[\mathrm{e}^{-\lambda x}\right]_0^\infty \\
&= [0-0] - (1/\lambda)[0-1] \\
&= 1/\lambda.
\end{aligned}$$

To obtain $\mathrm{E}(X^2)$, we choose $f = x^2$ and $g' = \lambda\,\mathrm{e}^{-\lambda x}$, and use integration by parts:

$$\begin{aligned}
\mathrm{E}(X^2) = \int_0^\infty x^2\,\lambda\,\mathrm{e}^{-\lambda x}\,\mathrm{d}x &= \left[-x^2\,\mathrm{e}^{-\lambda x}\right]_0^\infty + 2\int_0^\infty x\,\mathrm{e}^{-\lambda x}\,\mathrm{d}x \\
&= 0 + \frac{2}{\lambda}\int_0^\infty x\,\lambda\,\mathrm{e}^{-\lambda x}\,\mathrm{d}x \\
&= \frac{2}{\lambda^2}
\end{aligned}$$

where the last step follows because the last integral is simply $\mathrm{E}(X) = 1/\lambda$ again. Finally:

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

Means and variances can be 'infinite'

Expected values and variances are said to be infinite when the corresponding integral does not exist (i.e. does not have a finite value).

For the Pareto distribution, the distribution is defined for all $\alpha > 0$, but the mean is infinite if $\alpha < 1$ and the variance is infinite if $\alpha < 2$. This happens because for small values of α the distribution has very heavy tails, i.e. the probabilities of very large values of X are non-negligible.

This is actually useful in some insurance applications, for example liability insurance and medical insurance. There most claims are relatively small, but there is a non-negligible probability of extremely large claims. The Pareto distribution with a small α can be a reasonable representation of such situations. Figure 3.8 shows plots of Pareto cdfs with $\alpha = 2.2$ and $\alpha = 0.8$. When $\alpha = 0.8$, the distribution is so heavy-tailed that $\mathrm{E}(X)$ is infinite.

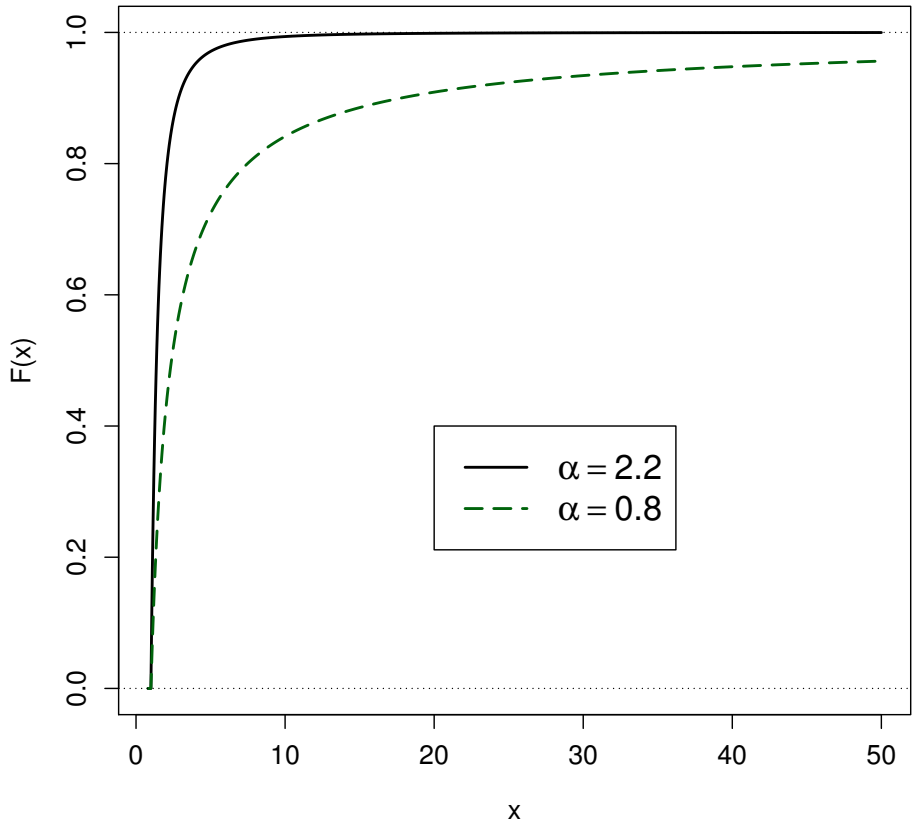


Figure 3.8: Pareto distribution cdfs.

Median of a random variable

Recall from **ST104a Statistics 1** that the *sample median* is essentially the observation 'in the middle' of a set of data, i.e. where half of the observations in the sample are smaller than the median and half of the observations are larger.

The **median** of a random variable (i.e. of its probability distribution) is similar in spirit.

Median of a random variable

The median, m, of a *continuous* random variable X is the value which satisfies:

$$F(m) = 0.5. \tag{3.6}$$

So once we know $F(x)$, we can find the median by solving (3.6).

Example 3.24 For the Pareto distribution we have:

$$F(x) = 1 - (k/x)^\alpha \qquad \text{for } x \geq k.$$

So $F(m) = 1 - (k/m)^\alpha = 1/2$ when:

$$(k/m)^\alpha = 1/2 \quad \Leftrightarrow \quad k/m = 1/\sqrt[\alpha]{2} \quad \Leftrightarrow \quad m = k\sqrt[\alpha]{2}.$$

For example:

- When $k = 1$ and $\alpha = 2.2$, the median is $m = \sqrt[2.2]{2} = 1.37$.
- When $k = 1$ and $\alpha = 0.8$, the median is $m = \sqrt[0.8]{2} = 2.38$.

Example 3.25 For the exponential distribution we have:

$$F(x) = 1 - \mathrm{e}^{-\lambda x} \qquad \text{for } x > 0.$$

So $F(m) = 1 - \mathrm{e}^{-\lambda m} = 1/2$ when:

$$\mathrm{e}^{-\lambda m} = 1/2 \quad \Leftrightarrow \quad -\lambda m = -\log 2 \quad \Leftrightarrow \quad m = \frac{\log 2}{\lambda}.$$

3.8 Overview of chapter

This chapter has formally introduced random variables, making a distinction between discrete and continuous random variables. Properties of probability distributions were discussed, including the determination of expected values and variances.

3.9 Key terms and concepts

- Constant
- Cumulative distribution function
- Estimators
- Experiment
- Median
- Continuous
- Discrete
- Expected value
- Interval
- Outcome

- Parameter
- Probability distribution
- Random variable
- Variance
- Probability density function
- Probability function
- Standard deviation

3.10 Learning activities

1. Suppose that the random variable X takes the values $\{x_1, x_2, \ldots\}$, where $x_1 < x_2 < \cdots$. Prove the following results:

   (a)
   $$\sum_{i=1}^{\infty} p(x_i) = 1.$$

   (b)
   $$p(x_k) = F(x_k) - F(x_{k-1}).$$

   (c)
   $$F(x_k) = \sum_{i=1}^{k} p(x_i).$$

2. At a charity event, the organisers sell 100 tickets to a raffle. At the end of the event, one of the tickets is selected at random and the person with that number wins a prize. Carol buys ticket number 22. Janet buys tickets numbered 1–5. What is the probability that each of them wins the prize?

3. A greengrocer has a very large pile of oranges on his stall. The pile of fruit is a mixture of 50% old fruit with 50% new fruit; one cannot tell which are old fruit and which are new fruit. However, 20% of old oranges are mouldy inside, but only 10% of new oranges are mouldy. Suppose that you choose 5 oranges at random. What is the distribution of the number of mouldy oranges in your sample?

4. What is the expectation of the random variable X if the only possible value it can take is c?

5. Show that $\mathrm{E}(X - \mathrm{E}(X)) = 0$.

6. Show that if $\mathrm{Var}(X) = 0$ then $p(\mu_X) = 1$. (We say in this case that X is *almost surely* equal to its mean.)

7. For a random variable X and constants a and b, prove that:
   $$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X).$$

Solutions to these questions can be found on the VLE in the **ST104b Statistics 2** area at http://my.londoninternational.ac.uk

3.11 Reminder of learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- define a random variable and distinguish it from the values that it takes
- explain the difference between discrete and continuous random variables
- find the mean and the variance of simple random variables whether discrete or continuous
- demonstrate how to proceed and use simple properties of expected values and variances.

3.12 Sample examination questions

1. In an investigation of animal behaviour, rats have to choose between four doors. One of them, behind which is food, is 'correct'. If an incorrect choice is made, the rat is returned to the starting point and chooses again, continuing as long as necessary until the correct choice is made. The random variable X is the serial number of the trial on which the correct choice is made.

   Find the probability function and expectation of X under each of the following hypotheses:

   (a) Each door is equally likely to be chosen on each trial, and all trials are mutually independent.

   (b) At each trial, the rat chooses with equal probability between the doors that it has not so far tried.

   (c) The rat never chooses the same door on two successive trials, but otherwise chooses at random with equal probabilities.

2. Construct suitable examples to show that, for a random variable X, then:

   (a) $\mathrm{E}(X^2) \neq \mathrm{E}(X)^2$, in general.

   (b) $\mathrm{E}(1/X) \neq 1/\mathrm{E}(X)$, in general.

3. (a) Let X be a random variable. Show that:

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

   (b) Let $X_1, X_2, \ldots, X_n$ be independent random variables. Assume that all have a mean of μ and a variance of σ^2. Find expressions for the mean and variance of the random variable $(X_1 + X_2 + \cdots + X_n)/n$.

Solutions to these questions can be found on the VLE in the **ST104b Statistics 2** area at http://my.londoninternational.ac.uk

Chapter 4
Common distributions of random variables

4.1 Synopsis of chapter content

This chapter formally introduces common 'families' of probability distributions which can be used to model various real-world phenomena.

4.2 Aims of the chapter

The aims of this chapter are to:

- be familiar with common probability distributions of both discrete and continuous types
- be familiar with the main properties of each common distribution introduced.

4.3 Learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- summarise basic distributions such as the uniform, Bernoulli, binomial, Poisson, exponential and normal
- calculate probabilities of events for these distributions using the probability function, probability density function or cumulative distribution function
- determine probabilities using statistical tables, where appropriate
- state properties of these distributions such as the expected value and variance.

4.4 Essential reading

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060] Chapters 4 and 5.

In addition there is essential 'watching' of this chapter's accompanying video tutorials accessible via the **ST104b Statistics 2** area at http://my.londoninternational.ac.uk

4.5 Introduction

In statistical inference we will treat observations:

$$X_1, X_2, \ldots, X_n$$

(the sample) as values of a random variable X, which has some probability distribution (population distribution).

How to choose that probability distribution?

- Usually we do not try to invent distributions from scratch.
- Instead, we use one of many existing standard distributions.
- There is a large number of such distributions, such that for most purposes we can find a suitable standard distribution.

This part of the course introduces some of the most common standard distributions for discrete and continuous random variables.

Probability distributions may differ from each other in a broader or narrower sense. In the broader sense, we have different **families** of distributions which may have quite different characteristics, for example:

- continuous versus discrete
- among discrete: finite versus infinite number of possible values
- among continuous: different sets of possible values (for example, all real numbers x, $x > 0$, or $x \in [0, 1]$); symmetric versus skewed distributions.

The 'distributions' discussed in this chapter are really families of distributions in this sense.

In the narrower sense, individual distributions *within* a family differ in having different values of the **parameters** of the distribution. The parameters determine the mean and variance of the distribution, values of probabilities from it, etc.

In the statistical analysis of a random variable X we typically:

- select a *family* of distributions based on the basic characteristics of X
- use observed data to choose (**estimate**) values for the *parameters* of that distribution, and perform statistical inference on them.

> **Example 4.1** An opinion poll on a referendum, where each X_i is an answer to the question 'Will you vote Yes or No to joining the European Union?' has answers

recorded as $X_i = 0$ if 'No' and $X_i = 1$ if 'Yes'. In a poll of 950 people, 513 answered 'Yes'.

How do we choose a distribution to represent X_i?

- Here we need a family of discrete distributions with only two possible values (0 and 1). The *Bernoulli distribution* (discussed in the next section), which has one parameter π (the probability of $X_i = 1$) is appropriate.
- Within the family of Bernoulli distributions, we use the one where the value of π is our best estimate based on the observed data. This is $\widehat{\pi} = 513/950 = 0.54$.

4.6 Common discrete distributions

For discrete random variables, we will consider the following distributions:

- Discrete uniform distribution
- Bernoulli distribution
- Binomial distribution
- Poisson distribution.

4.6.1 Discrete uniform distribution

Suppose a random variable X has k possible values $1, 2, \ldots, k$. X has a **discrete uniform distribution** if all of these values have the same probability, i.e. if:

$$p(x) = P(X = x) = \begin{cases} 1/k & \text{for all } x = 1, 2, \ldots, k \\ 0 & \text{otherwise.} \end{cases}$$

Example 4.2 A simple example of the discrete uniform distribution is the distribution of the score of a fair die, with $k = 6$.

The discrete uniform distribution is not very common in applications, but it is useful as a reference point for more complex distributions.

Mean and variance of a discrete uniform distribution

Calculating directly from the definition,[1] we have:

$$\mathrm{E}(X) = \sum_{x=1}^{k} x\, p(x) = \frac{1 + 2 + \cdots + k}{k} = \frac{k+1}{2} \tag{4.1}$$

[1] (4.1) and (4.2) make use, respectively, of $\sum_{i=1}^{n} i = n(n+1)/2$ and $\sum_{i=1}^{n} i^2 = n(n+1)(2n+1)/6$.

and:

$$\mathrm{E}(X^2) = \frac{1^2 + 2^2 + \cdots + k^2}{k} = \frac{(k+1)(2k+1)}{6}. \tag{4.2}$$

So:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{k^2 - 1}{12}.$$

4.6.2 Bernoulli distribution

A **Bernoulli trial** is an experiment with only *two* possible outcomes. We will number these outcomes 1 and 0, and refer to them as 'success' and 'failure', respectively.

> **Example 4.3** Examples of outcomes of Bernoulli trials are:
>
> - Agree / Disagree
> - Male / Female
> - Employed / Not employed
> - Owns a car / Does not own a car
> - Business goes bankrupt / Continues trading.

The **Bernoulli distribution** is the distribution of the outcome of a single Bernoulli trial. This is the distribution of a random variable X with the following probability function:

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

Therefore $P(X = 1) = \pi$ and $P(X = 0) = 1 - P(X = 1) = 1 - \pi$, and no other values are possible. Such a random variable X has a Bernoulli distribution with (probability) parameter π. This is often written as:

$$X \sim \text{Bernoulli}(\pi).$$

If $X \sim \text{Bernoulli}(\pi)$, then:

$$\mathrm{E}(X) = \sum_{x=0}^{1} x\, p(x) = 0 \times (1-\pi) + 1 \times \pi = \pi \tag{4.3}$$

$$\mathrm{E}(X^2) = \sum_{x=0}^{1} x^2\, p(x) = 0^2 \times (1-\pi) + 1^2 \times \pi = \pi$$

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \pi - \pi^2 = \pi\,(1-\pi). \tag{4.4}$$

4.6.3 Binomial distribution

Suppose we carry out n Bernoulli trials such that:

- at each trial, the probability of success is π
- different trials are statistically independent events.

Let X denote the total number of successes in these n trials. Then X follows a **binomial distribution** with parameters n and π, where $n \geq 1$ is a known integer and $0 \leq \pi \leq 1$. This is often written as:

$$X \sim \text{Bin}(n, \pi).$$

The binomial distribution was first encountered in Example 3.14.

Example 4.4 A multiple choice test has 4 questions, each with 4 possible answers. Bob is taking the test, but has no idea at all about the correct answers. So he guesses every answer and therefore has the probability of 1/4 of getting any individual question correct.

Let X denote the number of correct answers in Bob's test. X follows the binomial distribution with $n = 4$ and $\pi = 0.25$, i.e. we have:

$$X \sim \text{Bin}(4, 0.25).$$

For example, what is the probability that Bob gets 3 of the 4 questions correct?

Here it is assumed that the guesses are independent, and each has the probability $\pi = 0.25$ of being correct. The probability of any particular sequence of 3 correct and 1 incorrect answers, for example 1110, is $\pi^3 (1-\pi)^1$, where '1' denotes a correct answer and '0' denotes an incorrect answer.

However, we do not care about the order of the 0s and 1s, only about the number of 1s. So 1101 and 1011, for example, also count as 3 correct answers. Each of these also has the probability $\pi^3 (1-\pi)^1$.

The total number of sequences with three 1s (and therefore one 0) is the number of locations for the three 1s that can be selected in the sequence of 4 answers. This is $\binom{4}{3} = 4$. Therefore the probability of obtaining three 1s is:

$$\binom{4}{3} \pi^3 (1-\pi)^1 = 4 \times 0.25^3 \times 0.75^1 \approx 0.0469.$$

Binomial distribution probability function

In general, the probability function of $X \sim \text{Bin}(n, \pi)$ is:

$$p(x) = \begin{cases} \binom{n}{x} \pi^x (1-\pi)^{n-x} & \text{for } x = 0, 1, \ldots, n \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

We have already shown that (4.5) satisfies the conditions for being a probability function in the previous chapter (see Example 3.14).

Example 4.5 Continuing Example 4.4, where $X \sim \text{Bin}(4, 0.25)$, we have:

$$p(0) = \binom{4}{0} \times (0.25)^0 \times (0.75)^4 = 0.3164, \quad p(1) = \binom{4}{1} \times (0.25)^1 \times (0.75)^3 = 0.4219,$$
$$p(2) = \binom{4}{2} \times (0.25)^2 \times (0.75)^2 = 0.2109, \quad p(3) = \binom{4}{3} \times (0.25)^3 \times (0.75)^1 = 0.0469,$$
$$p(4) = \binom{4}{4} \times (0.25)^4 \times (0.75)^0 = 0.0039.$$

If $X \sim \text{Bin}(n, \pi)$, then:

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

The expected value $\text{E}(X)$ was derived in the previous chapter. The variance will be derived later.

Example 4.6 Suppose a multiple choice examination has 20 questions, each with 4 possible answers. Consider again a student who guesses each one of the answers. Let X denote the number of correct answers by such a student, so that we have $X \sim \text{Bin}(20, 0.25)$. For such a student, the expected number of correct answers is $\text{E}(X) = 20 \times 0.25 = 5$.

The teacher wants to set the pass mark of the examination so that, for such a student, the probability of passing is less than 0.05. What should the pass mark be?

In other words, what is the smallest x such that $P(X \geq x) < 0.05$, i.e. such that $P(X < x) \geq 0.95$?

Calculating the probabilities of $x = 0, 1, \ldots, 20$ we get (rounded to 2 decimal places):

x	0	1	2	3	4	5	6	7	8	9	10
$p(x)$	0.00	0.02	0.07	0.13	0.19	0.20	0.17	0.11	0.06	0.03	0.01
x		11	12	13	14	15	16	17	18	19	20
$p(x)$		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Calculating the cumulative probabilities, we find that $F(7) = P(X < 8) = 0.898$ and $F(8) = P(X < 9) = 0.959$. Therefore $P(X \geq 8) = 0.102 > 0.05$ and also $P(X \geq 9) = 0.041 < 0.05$. The pass mark should be set at 9.

More generally, consider a student who has the same probability π of the correct answer for every question, so that $X \sim \text{Bin}(20, \pi)$. Figure 4.1 shows plots of the probabilities for $\pi = 0.25$, 0.5, 0.7 and 0.9.

4.6.4 Poisson distribution

The possible values of the **Poisson distribution** are the non-negative integers $0, 1, 2, \ldots$.

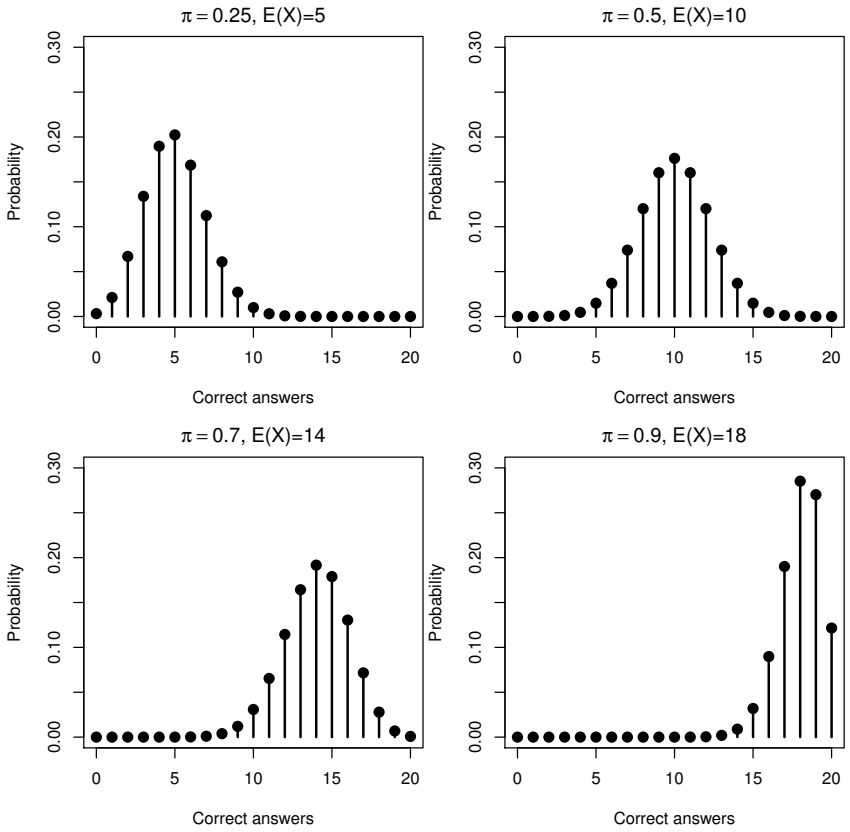


Figure 4.1: Probability plots for Example 4.6.

Poisson distribution probability function

The probability function of the Poisson distribution is:

$$p(x) = \begin{cases} \mathrm{e}^{-\lambda}\,\lambda^x/x! & \text{for } x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

where $\lambda > 0$ is a parameter.

Activity 4.1 Show that (4.6) satisfies the conditions to be a probability function.

Hint: You can use the following result from standard calculus. For any number a,

$$\mathrm{e}^a = \sum_{x=0}^{\infty} \frac{a^x}{x!}.$$

If a random variable X has a Poisson distribution with parameter λ, this is often denoted by:

$$X \sim \text{Poisson}(\lambda) \quad \text{or} \quad X \sim \text{Pois}(\lambda).$$

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

$$\begin{aligned} \mathrm{E}(X) &= \lambda \\ \mathrm{Var}(X) &= \lambda. \end{aligned}$$

Activity 4.2 Prove that the mean and variance of a Poisson-distributed random variable are both equal to λ.

Poisson distributions are used for **counts** of occurrences of various kinds. To give a formal motivation, suppose that we consider the number of occurrences of some phenomenon in time, and that the process that generates the occurrences satisfies the following conditions:

1. The numbers of occurrences in any two *disjoint* intervals of time are independent of each other.
2. The probability of two or more occurrences at the *same* time is negligibly small.
3. The probability of one occurrence in any short time interval of length t is λt for some constant $\lambda > 0$.

In essence, these state that individual occurrences should be independent, sufficiently rare, and happen at a constant rate λ per unit of time. A process like this is a **Poisson process**.

If occurrences are generated by a Poisson process, then the number of occurrences in a randomly selected time interval of length $t = 1$, X, follows a Poisson distribution with mean λ, i.e. $X \sim \text{Poisson}(\lambda)$.

The single parameter λ of the Poisson distribution is therefore the **rate** of occurrences per unit of time.

Example 4.7 Examples of variables for which we might use a Poisson distribution:

- Number of telephone calls received at a call centre per minute.
- Number of accidents on a stretch of motorway per week.
- Number of customers arriving at a checkout per minute.
- Number of misprints per page of newsprint.

Because λ is the rate per unit of time, its value also depends on the unit of time (that is, the length of interval) we consider.

Example 4.8 If X is the number of arrivals per hour and $X \sim \text{Poisson}(1.5)$, then if Y is the number of arrivals per *two* hours, $Y \sim \text{Poisson}(2 \times 1.5) = \text{Poisson}(3)$.

λ is also the mean of the distribution, i.e. $\text{E}(X) = \lambda$.

Both motivations suggest that distributions with higher values of λ have higher probabilities of large values of X.

Example 4.9 Figure 4.2 shows the probabilities $p(x)$ for $x = 0, 1, 2, \ldots, 10$ for $X \sim \text{Poisson}(2)$ and $X \sim \text{Poisson}(4)$.

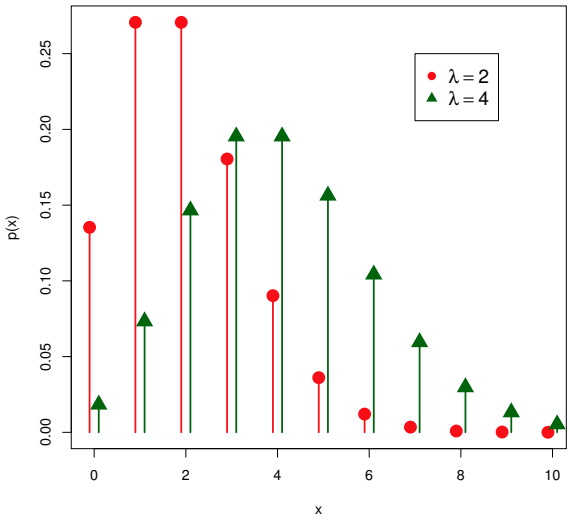


Figure 4.2: Probability plots for Example 4.9.

Example 4.10 Customers arrive at a bank on weekday afternoons randomly at an average rate of 1.6 customers per minute. Let X denote the number of arrivals per minute and Y denote the number of arrivals per 5 minutes.

We assume a Poisson distribution for both, such that:

$$\begin{aligned} X &\sim \text{Poisson}(1.6) \\ \text{and} \quad Y &\sim \text{Poisson}(5 \times 1.6) = \text{Poisson}(8). \end{aligned}$$

1. What is the probability that no customer arrives in a one-minute interval?

   For $X \sim \text{Poisson}(1.6)$, the probability $P(X = 0)$ is:

   $$p_X(0) = \frac{\mathrm{e}^{-\lambda}\,\lambda^0}{0!} = \frac{\mathrm{e}^{-1.6}\,(1.6)^0}{0!} = \mathrm{e}^{-1.6} = 0.2019.$$

2. What is the probability that more than two customers arrive in a one-minute interval?

   $P(X > 2) = 1 - P(X \leq 2) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$ which is:

   $$\begin{aligned} 1 - p_X(0) - p_X(1) - p_X(2) &= 1 - \frac{\mathrm{e}^{-1.6}\,(1.6)^0}{0!} - \frac{\mathrm{e}^{-1.6}\,(1.6)^1}{1!} - \frac{\mathrm{e}^{-1.6}\,(1.6)^2}{2!} \\ &= 1 - \mathrm{e}^{-1.6} - 1.6\mathrm{e}^{-1.6} - 1.28\mathrm{e}^{-1.6} \\ &= 1 - 3.88\mathrm{e}^{-1.6} \\ &= 0.2167. \end{aligned}$$

3. What is the probability that no more than 1 customer arrives in a five-minute interval?

   For $Y \sim \text{Poisson}(8)$, the probability $P(Y \leq 1)$ is:

   $$p_Y(0) + p_Y(1) = \frac{\mathrm{e}^{-8}\,(8)^0}{0!} + \frac{\mathrm{e}^{-8}\,(8)^1}{1!} = \mathrm{e}^{-8} + 8\mathrm{e}^{-8} = 9\mathrm{e}^{-8} = 0.0030.$$

A word on calculators

In the examination you will be allowed a basic calculator only. To calculate binomial and Poisson probabilities directly requires access to a 'factorial' key (for the binomial) and 'e' key (for the Poisson), which will not appear on a basic calculator. Note that any probability calculations which are required in the examination will be possible on a basic calculator. For example, for the Poisson probabilities in Example 4.10, it would be acceptable to give your answers in terms of e (in the simplest form).

4.6.5 Connections between probability distributions

There are close connections between some probability distributions, even across different families of them. Some connections are exact, i.e. one distribution is exactly equal to another, for particular values of the parameters. For example, Bernoulli(π) is the same distribution as Bin(1, π).

Some connections are approximate (or asymptotic), i.e. one distribution is closely approximated by another under some limiting conditions. We next discuss one of these, the Poisson approximation of the binomial distribution.

4.6.6 Poisson approximation of the binomial distribution

Suppose that:

- $X \sim \text{Bin}(n, \pi)$.
- n is large and π is small.

Under such circumstances, the distribution of X is well-approximated by a Poisson(λ) distribution with $\lambda = n\,\pi$.

The connection is exact at the limit, i.e. $\text{Bin}(n, \pi) \to \text{Poisson}(\lambda)$ if $n \to \infty$ and $\pi \to 0$ in such a way that $n\,\pi = \lambda$ remains constant.

> **Activity 4.3** Suppose that $X \sim \text{Bin}(n, \pi)$ and $Y \sim \text{Poisson}(\lambda)$. Show that, if $n \to \infty$ and $\pi \to 0$ in such a way that $n\,\pi = \lambda$ remains constant, then, for any x, we have:
>
> $$P(X = x) \to P(Y = x) \quad \text{as } n \to \infty.$$
>
> Hint 1: Because $n\,\pi = \lambda$ remains constant, substitute λ/n for π from the beginning.
>
> Hint 2: One step of the proof uses the limit definition of the exponential function, which states that, for any number y, we have:
>
> $$\lim_{n\to\infty} \left(1 + \frac{y}{n}\right)^n = \mathrm{e}^y.$$

This 'law of small numbers' provides another motivation for the Poisson distribution.

Example 4.11 A classic example (from Bortkiewicz (1898) *Das Gesetz der kleinen Zahlen*) helps to remember the key elements of the 'law of small numbers'.

Figure 4.3 shows the numbers of soldiers killed by horsekick in each of 14 Army Corps of the Prussian army in each of the years spanning 1875–94.

Suppose that the number of men killed by horsekicks in one corps in one year is $X \sim \text{Bin}(n, \pi)$, where:

- n is large – the number of men in a corps (perhaps 50,000).
- π is small – the probability that a man is killed by a horsekick.

Then X should be well-approximated by a Poisson distribution with some mean λ. The sample frequencies and proportions of different counts are as follows:

Number killed	0	1	2	3	4	More
Count	144	91	32	11	2	0
%	51.4	32.5	11.4	3.9	0.7	0

The sample mean of the counts is $\bar{x} = 0.7$, which we use as λ for the Poisson distribution. $X \sim \text{Poisson}(0.7)$ is indeed a good fit to the data, as shown in Figure 4.4.

	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
G	—	2	2	1	—	—	1	1	—	3	—	2	1	—	—	1	—	1	—	1
I	—	—	—	2	—	3	—	2	—	—	—	1	1	1	—	2	—	3	1	—
II	—	—	—	2	—	2	—	—	1	1	—	—	2	1	1	—	—	2	—	—
III	—	—	—	1	1	1	2	—	2	—	—	—	1	—	1	2	1	—	—	—
IV	—	1	—	1	1	1	1	—	—	—	—	1	—	—	—	—	1	1	—	—
V	—	—	—	—	2	1	—	—	1	—	—	1	—	1	1	1	1	1	1	—
VI	—	—	1	—	2	—	—	1	2	—	1	1	3	1	1	1	—	3	—	—
VII	1	—	1	—	—	—	1	—	1	1	—	—	2	—	—	2	1	—	2	—
VIII	1	—	—	—	1	—	—	1	—	—	—	—	1	—	—	—	1	1	—	1
IX	—	—	—	—	—	2	1	1	1	—	2	1	1	—	1	2	—	1	—	—
X	—	—	1	1	—	1	—	2	—	2	—	—	—	—	2	1	3	—	1	1
XI	—	—	—	—	2	4	—	1	3	—	1	1	1	1	2	1	3	1	3	1
XIV	1	1	2	1	1	3	—	4	—	1	—	3	2	1	—	2	1	1	—	—
XV	—	1	—	—	—	—	—	1	—	1	1	—	—	—	2	2	—	—	—	—

Figure 4.3: Numbers of soldiers killed by horsekick in each of 14 army corps of the Prussian army in each of the years 1875–94. Source: Bortkiewicz (1898) *Das Gesetz der kleinen Zahlen*, Leipzig: Teubner.

Example 4.12 An airline is selling tickets to a flight with 198 seats. It knows that, on average, about 1% of customers who have bought tickets fail to arrive for the flight. Because of this, the airline overbooks the flight by selling 200 tickets. What is the probability that everyone who arrives for the flight will get a seat?

Let X denote the number of people who fail to turn up. Using the binomial distribution, $X \sim \text{Bin}(200, 0.01)$. We have:

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - 0.1340 - 0.2707 = 0.5953.$$

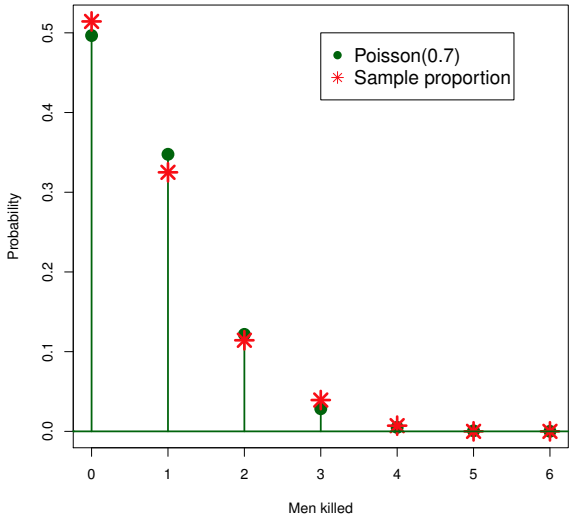


Figure 4.4: Fit of Poisson distribution to the data in Example 4.11.

Using the Poisson approximation, $X \sim \text{Poisson}(200 \times 0.01) = \text{Poisson}(2)$.

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - \mathrm{e}^{-2} - 2\mathrm{e}^{-2} = 1 - 3\mathrm{e}^{-2} = 0.5940.$$

4.6.7 Some other discrete distributions

Just their names and short comments are given here, so that you have an idea of what else there is.

- Geometric(π) distribution:
  - Distribution of the number of failures in Bernoulli trials before the first success.
  - π is the probability of success at each trial.
  - Sample space is $0, 1, 2, \ldots$.
  - See the basketball example in Chapter 3.
- Negative binomial(r, π) distribution:
  - Distribution of the number of failures in Bernoulli trials before r successes occur.
  - π is the probability of success at each trial.
  - Sample space is $0, 1, 2, \ldots$.
  - Negative Binomial(1, π) is the same as Geometric(π).
- Hypergeometric(n, A, B) distribution:
  - Experiment where initially $A + B$ objects are available for selection, and A of them represent 'success'.
  - n objects are selected at random, *without replacement*.
  - Hypergeometric is then the distribution of the number of successes.

- Sample space is the integers x where $\max\{0,\, n - B\} \leq x \leq \min\{n,\, A\}$.
- If the selection was *with* replacement, the distribution of the number of successes would be $\text{Bin}(n,\, A/(A + B))$.

- Multinomial$(n,\, \pi_1,\, \pi_2,\, \ldots,\, \pi_k)$ distribution:
  - Here $\pi_1 + \pi_2 + \cdots + \pi_k = 1$, and the π_is are the probabilities of the values $1, 2, \ldots, k$.
  - If $n = 1$, the sample space is $1, 2, \ldots, k$. This is essentially a generalisation of the discrete uniform distribution, but with non-equal probabilities π_i.
  - If $n > 1$, the sample space is the vectors $(n_1, n_2, \ldots, n_k)$ where $n_i \geq 0$ for all i, and $n_1 + n_2 + \cdots + n_k = n$. This is essentially a generalisation of the binomial to the case where each trial has $K \geq 2$ possible outcomes, and the random variable records the numbers of each outcome in n trials. Note that with $K = 2$, Multinomial$(n,\, \pi_1,\, \pi_2)$ is essentially the same as $\text{Bin}(n,\, \pi)$ with $\pi = \pi_2$ (or with $\pi = \pi_1$).
  - When $n > 1$, the multinomial is the distribution of a *multivariate* random variable, as discussed later in the course.

4.7 Common continuous distributions

For continuous random variables, we will consider the following distributions:

- Uniform distribution
- Exponential distribution
- Normal distribution.

4.7.1 The (continuous) uniform distribution

The (continuous) uniform distribution has non-zero probabilities only on an interval $[a,\, b]$, where $a < b$ are given numbers. The probability that its value is in an interval within $[a,\, b]$ is proportional to the length of that interval. In other words, all intervals (within $[a,\, b]$) which have the same length have the same probability.

Uniform distribution pdf

The pdf of the (continuous) uniform distribution is:

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

A random variable X with this pdf may be written as $X \sim \text{Uniform}[a,\, b]$.

The pdf is 'flat', as shown in Figure 4.5 (along with the cdf). Clearly, $f(x) \geq 0$ for all x, and:

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_a^b \frac{1}{b-a}\,\mathrm{d}x = \frac{1}{b-a}[x]_a^b = \frac{1}{b-a} \cdot [b-a] = 1.$$

The cdf is:

$$F(x) = P(X \leq x) = \int_a^x f(t)\,\mathrm{d}t = \begin{cases} 0 & \text{for } x < a \\ (x-a)/(b-a) & \text{for } a \leq x \leq b \\ 1 & \text{for } x > b. \end{cases}$$

Activity 4.4 Derive the cdf for the continuous uniform distribution.

The probability of an interval $[x_1, x_2]$, where $a \leq x_1 < x_2 \leq b$, is therefore:

$$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1) = \frac{x_2 - x_1}{b-a}.$$

So the probability depends only on the length of the interval, $x_2 - x_1$.

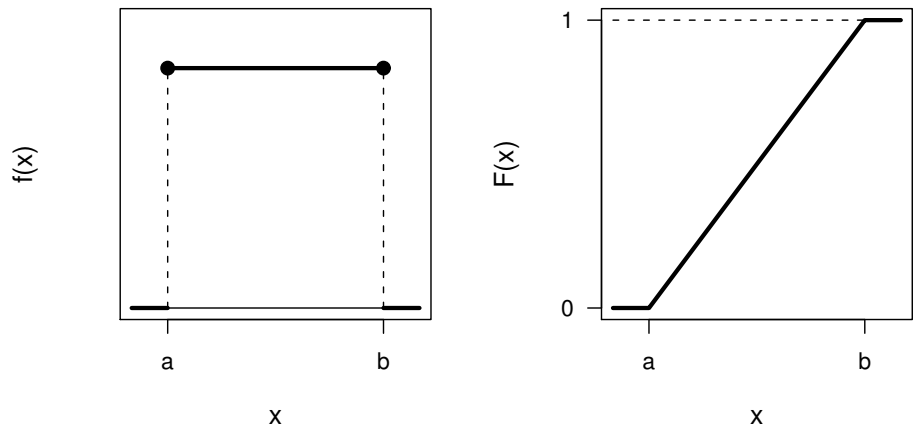


Figure 4.5: Continuous uniform distribution pdf (left) and cdf (right).

If $X \sim \text{Uniform}[a, b]$, we have:

$$\begin{aligned} \mathrm{E}(X) &= \frac{b+a}{2} = \text{median of } X \\ \mathrm{Var}(X) &= \frac{(b-a)^2}{12}. \end{aligned}$$

The mean and median also follow from the fact that the distribution is symmetric about $(b+a)/2$, i.e. the midpoint of the interval $[a, b]$.

Activity 4.5 Derive the mean and variance of the continuous uniform distribution.

4.7.2 Exponential distribution

Exponential distribution pdf

A random variable X has the **exponential distribution** with the parameter λ (where $\lambda > 0$) if its probability density function is:

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

This is often denoted $X \sim \text{Exponential}(\lambda)$ or $X \sim \text{Exp}(\lambda)$.

It was shown in the previous chapter that this satisfies the conditions for a pdf (see Example 3.21). The general shape of the pdf is that of 'exponential decay', as shown in Figure 4.6 (hence the name).

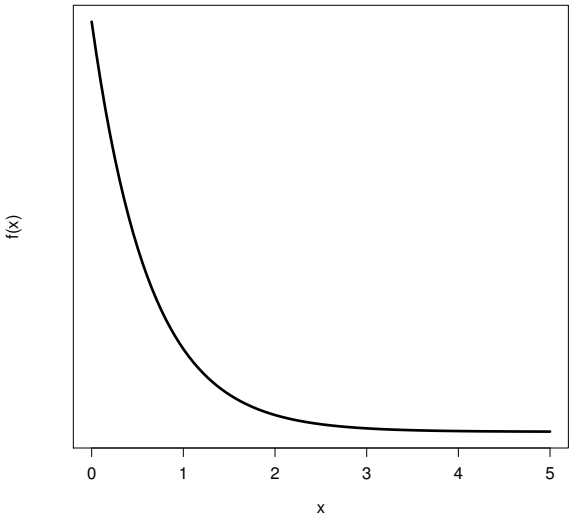


Figure 4.6: Exponential distribution pdf.

The cdf of the Exponential(λ) distribution is:

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 - \mathrm{e}^{-\lambda x} & \text{for } x > 0. \end{cases}$$

The cdf is shown in Figure 4.7 for $\lambda = 1.6$.

For $X \sim \text{Exponential}(\lambda)$, we have:

$$\begin{aligned} \mathrm{E}(X) &= 1/\lambda \\ \mathrm{Var}(X) &= 1/\lambda^2. \end{aligned}$$

These have been derived in the previous chapter (see Example 3.23). The median of the distribution, also previously derived (see Example 3.25), is:

$$m = \frac{\log 2}{\lambda} = (\log 2) \times \frac{1}{\lambda} = (\log 2)\,\mathrm{E}(X) \approx 0.69\,\mathrm{E}(X).$$

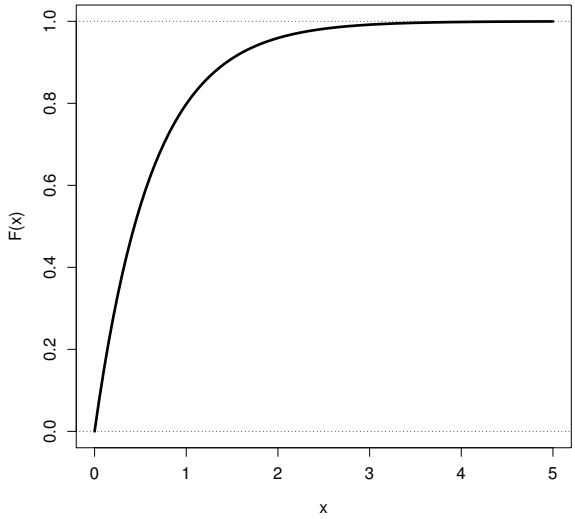


Figure 4.7: Exponential distribution cdf for $\lambda = 1.6$.

Note that the median is always smaller than the mean, because the distribution is skewed to the right.

Uses of the exponential distribution

The exponential is, among other things, a basic distribution of **waiting times** of various kinds. This arises from a connection between the Poisson distribution – the simplest distribution for *counts* – and the exponential.

- If the number of events per unit of time has a Poisson distribution with parameter λ, the time interval (measured in the same units of time) between two successive events has an exponential distribution with the same parameter λ.

Note that the expected values of these behave as we would expect.

- $\mathrm{E}(X) = \lambda$ for Poisson(λ), i.e. a large λ means many events per unit of time, on average.
- $\mathrm{E}(X) = 1/\lambda$ for Exponential(λ), i.e. a large λ means short waiting times between successive events, on average.

Example 4.13 Consider Example 4.10.

- The number of customers arriving at a bank per minute has a Poisson distribution with parameter $\lambda = 1.6$.
- Then the time X, in minutes, between the arrivals of two successive customers follows an exponential distribution with parameter $\lambda = 1.6$.

From this exponential distribution, the expected waiting time between arrivals of customers is $\mathrm{E}(X) = 1/1.6 = 0.625$ (minutes) and the median is calculated to be $(\log 2) \times 0.625 = 0.433$.

We can also calculate probabilities of waiting times between arrivals, using the cumulative distribution function:

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 - \mathrm{e}^{-1.6x} & \text{for } x > 0. \end{cases}$$

For example:

- $P(X \leq 1) = F(1) = 1 - \mathrm{e}^{-1.6\times 1} = 1 - \mathrm{e}^{-1.6} = 0.7981$.
  The probability is about 0.8 that two arrivals are at most a minute apart.
- $P(X > 3) = 1 - F(3) = \mathrm{e}^{-1.6\times 3} = \mathrm{e}^{-4.8} = 0.0082$.
  The probability of a gap of 3 minutes or more between arrivals is very small.

4.7.3 Two other distributions

These are generalisations of the uniform and exponential distributions. Only their names and short comments are given here, just so that you know they exist.

- Beta(α, β) distribution, shown in Figure 4.8.
  - Generalising the uniform, these are distributions for a closed interval, which is taken to be [0, 1].
  - Sample space is therefore $\{x \,|\, 0 \leq x \leq 1\}$.
  - Unlike for the uniform distribution, the pdf is generally not flat.
  - Beta(1, 1) is the same as Uniform[0, 1].
- Gamma(α, β) distribution, shown in Figure 4.9.
  - Generalising the exponential distribution, this is a two-parameter family of skewed distributions for positive values.
  - Sample space is $\{x \,|\, x > 0\}$.
  - Gamma(1, β) is the same as Exponential(β).

4.7.4 Normal (Gaussian) distribution

The **normal distribution** is by far the most important probability distribution in statistics. This is for three broad reasons:

- Many variables have distributions that are approximately normal, for example heights of humans, animals and weights of various products.
- The normal distribution has extremely convenient mathematical properties, which make it a useful default choice of distribution in many contexts.
- Even when a variable is not itself even approximately normally distributed, functions of several observations of the variable ('sampling distributions') are often approximately normal, due to the **central limit theorem**. Because of this, the

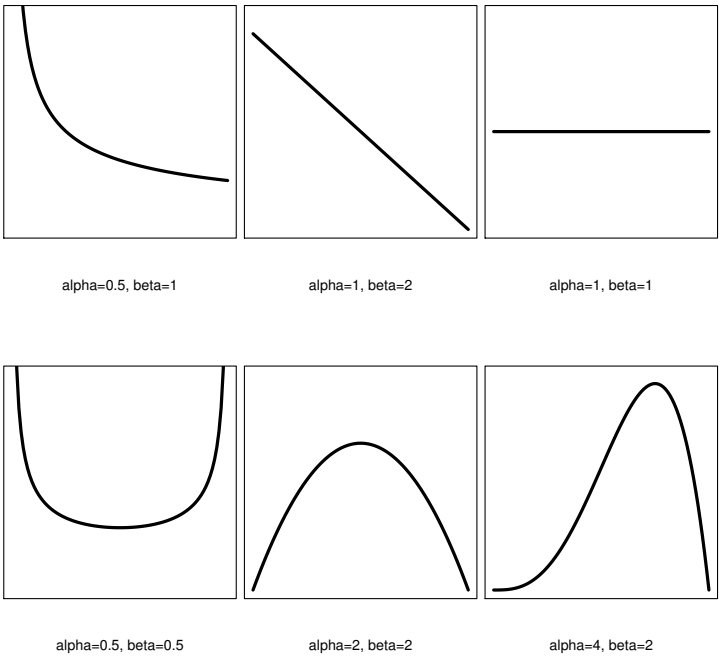


Figure 4.8: Beta distribution density functions.

normal distribution has a crucial role in statistical inference. This will be discussed later in the course.

Normal distribution pdf

The pdf of the normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \qquad \text{for } -\infty < x < \infty$$

where π is the mathematical constant (i.e. $\pi = 3.14159\ldots$), and μ and σ^2 are parameters, with $-\infty < \mu < \infty$ and $\sigma^2 > 0$.

A random variable X with this pdf is said to have a normal distribution with mean μ and variance σ^2, denoted $X \sim N(\mu, \sigma^2)$.

Clearly, $f(x) \geq 0$ for all x. Also, it can be shown that $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$ (do not attempt to show this), so $f(x)$ really is a pdf.

If $X \sim N(\mu, \sigma^2)$, then:

$$\begin{aligned} \mathrm{E}(X) &= \mu \\ \mathrm{Var}(X) &= \sigma^2 \end{aligned}$$

and the standard deviation is therefore $\mathrm{sd}(X) = \sigma$.

The mean can also be inferred from the observation that the normal pdf is **symmetric** about μ. This also implies that the median of the normal distribution is μ.

The normal density is the so-called 'bell curve'. The two parameters affect it as follows:

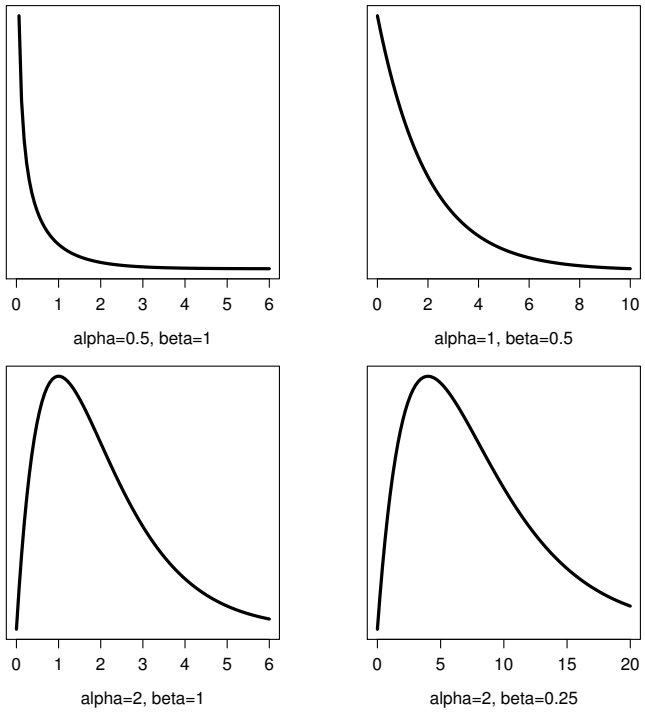


Figure 4.9: Gamma distribution density functions.

- The mean μ determines the location of the curve.
- The variance σ^2 determines the dispersion (spread) of the curve.

Example 4.14 Figure 4.10 shows that:

- $N(0, 1)$ and $N(5, 1)$ have the same dispersion but different location: the $N(5, 1)$ curve is identical to the $N(0, 1)$ curve, but shifted 5 units to the right.
- $N(0, 1)$ and $N(0, 9)$ have the same location but different dispersion: the $N(0, 9)$ curve is centered at the same value, 0, as the $N(0, 1)$ curve, but spread out more widely.

Linear transformations of the normal distribution

We now consider one of the convenient properties of the normal distribution. Suppose X is a random variable, and we consider the linear transformation $Y = aX + b$, where a and b are constants.

Whatever the distribution of X, it is true that $\mathrm{E}(Y) = a\mathrm{E}(X) + b$ and also that $\mathrm{Var}(Y) = a^2\mathrm{Var}(X)$.

Furthermore, if X is *normally* distributed, then so is Y. In other words, if $X \sim N(\mu, \sigma^2)$, then:

$$Y = aX + b \sim N(a\mu + b, a^2\sigma^2). \tag{4.7}$$

This type of result is *not* true in general. For other families of distributions, the distribution of $Y = aX + b$ is not always in the same family as X.

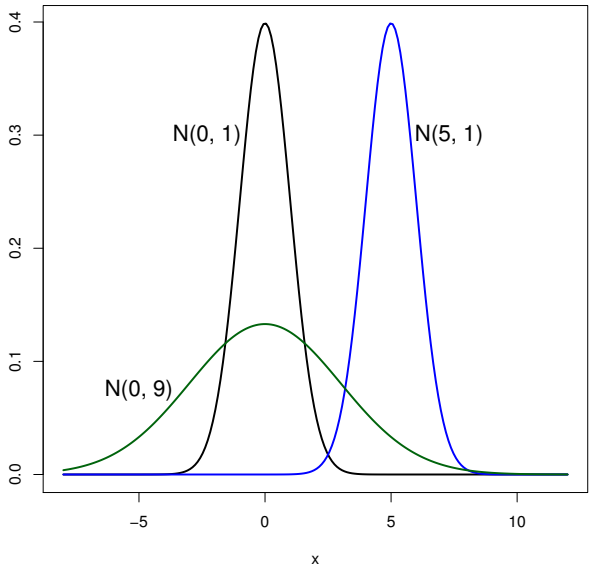


Figure 4.10: Various normal distributions.

Let us apply (4.7) with $a = 1/\sigma$ and $b = -\mu/\sigma$, to get:

$$Z = \frac{1}{\sigma} X - \frac{\mu}{\sigma} = \frac{X - \mu}{\sigma} \sim N\left(\frac{1}{\sigma} \cdot \mu - \frac{\mu}{\sigma}, \left(\frac{1}{\sigma}\right)^2 \cdot \sigma^2\right) = N(0, 1).$$

The transformed variable $Z = (X - \mu)/\sigma$ is known as a **standardised variable** or a **z-score**.

The distribution of the z-score is $N(0, 1)$, i.e. the normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ (and therefore a standard deviation of $\sigma = 1$). This is known as the **standard normal distribution**. Its density function is:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right] \qquad \text{for } -\infty < x < \infty.$$

The cumulative distribution function of the normal distribution is:

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] \mathrm{d}t.$$

In the special case of the standard normal distribution, the cdf is:

$$F(x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{t^2}{2}\right] \mathrm{d}t.$$

Note, this is often denoted $\Phi(x)$.

Such integrals cannot be evaluated in a closed form, so we use statistical tables of them, specifically a table of $\Phi(x)$ (or we could use a computer, but not in the examination).

In the examination, you will have a table of some values of $\Phi(x)$, the cdf of $Z \sim N(0, 1)$. Specifically, Table 4 of the *New Cambridge Statistical Tables* shows values of $\Phi(x) = P(Z \leq x)$ for $x \geq 0$. This table can be used to calculate probabilities of any intervals for any normal distribution. But how? The table seems to be incomplete:

1. It is only for $N(0, 1)$, not for $N(\mu, \sigma^2)$ for any other μ and σ^2.
2. Even for $N(0, 1)$, it only shows probabilities for $x \geq 0$.

The key to using the tables is that the standard normal distribution is symmetric about 0. This means that for an interval in one tail, its 'mirror image' in the other tail has the same probability. Another way to justify these results is that if $Z \sim N(0, 1)$, then $-Z \sim N(0, 1)$ also. See **ST104a Statistics 1** for a discussion of how to use Table 4 of the *New Cambridge Statistical Tables*.

Probabilities for any normal distribution

How about a normal distribution $X \sim N(\mu, \sigma^2)$, for any other μ and σ^2?

What if we want to calculate, for any $a < b$, $P(a < X \leq b) = F(b) - F(a)$?

Remember that $(X - \mu)/\sigma = Z \sim N(0, 1)$. If we apply this transformation to all parts of the inequalities, we get:

$$\begin{aligned} P(a < X \leq b) &= P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} < Z \leq \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

which can be calculated using Table 4 of the *New Cambridge Statistical Tables*. (Note that this also covers the cases of the one-sided inequalities $P(X \leq b)$, with $a = -\infty$, and $P(X > a)$, with $b = \infty$.)

Example 4.15 Let X denote the diastolic blood pressure of a randomly selected person in England. This is approximately distributed as $X \sim N(74.2, 127.87)$.

Suppose we want to know the probabilities of the following intervals:

- $X > 90$ (high blood pressure)
- $X < 60$ (low blood pressure)
- $60 \leq X \leq 90$ (normal blood pressure).

These are calculated using standardisation with $\mu = 74.2$ and $\sigma^2 = 127.87$, and therefore $\sigma = 11.31$. So here:

$$\frac{X - 74.2}{11.31} = Z \sim N(0, 1)$$

and we can refer values of this standardised variable to Table 4 of the *New*

Cambridge Statistical Tables.

$$\begin{aligned} P(X > 90) &= P\left(\frac{X - 74.2}{11.31} > \frac{90 - 74.2}{11.31}\right) \\ &= P(Z > 1.40) \\ &= 1 - \Phi(1.40) \\ &= 1 - 0.9192 \\ &= 0.0808. \end{aligned}$$

$$\begin{aligned} P(X < 60) &= P\left(\frac{X - 74.2}{11.31} < \frac{60 - 74.2}{11.31}\right) \\ &= P(Z < -1.26) \\ &= P(Z > 1.26) \\ &= 1 - \Phi(1.26) \\ &= 1 - 0.8962 \\ &= 0.1038. \end{aligned}$$

Finally:

$$P(60 \leq X \leq 90) = P(X \leq 90) - P(X < 60) = 0.8152.$$

These probabilities are shown in Figure 4.11.

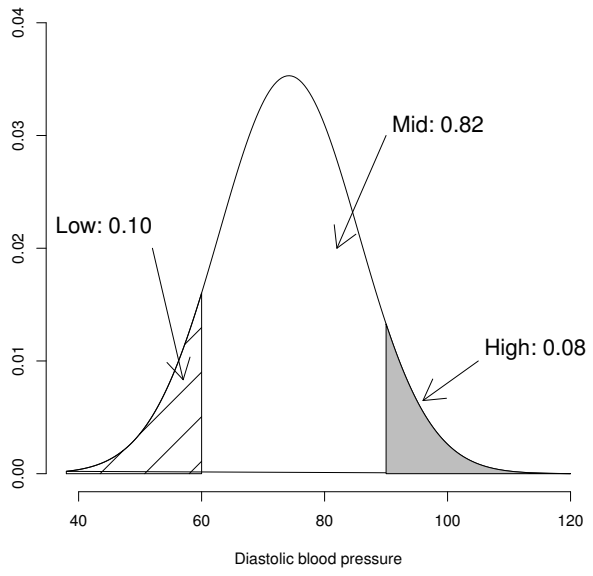


Figure 4.11: Distribution of blood pressure for Example 4.15.

Some probabilities around the mean

The following results hold for all normal distributions:

- $P(\mu - \sigma < X < \mu + \sigma) = 0.683$. In other words, about 68.3% of the total probability is within 1 standard deviation of the mean.
- $P(\mu - 1.96\sigma < X < \mu + 1.96\sigma) = 0.950$.
- $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.954$.
- $P(\mu - 2.58\sigma < X < \mu + 2.58\sigma) = 0.99$.
- $P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997$.

The first two of these are illustrated graphically in Figure 4.12.

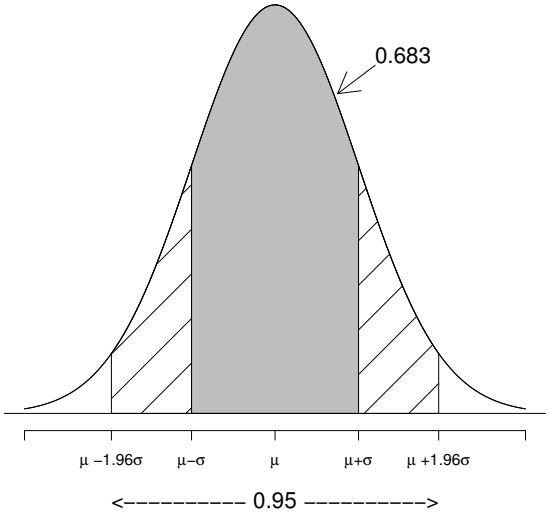


Figure 4.12: Some probabilities around the mean for the normal distribution.

4.7.5 Normal approximation of the binomial distribution

For $0 < \pi < 1$, the binomial distribution $\text{Bin}(n, \pi)$ tends to the normal distribution $N(n\,\pi, n\,\pi\,(1-\pi))$ as $n \to \infty$.

Less formally: The binomial is well-approximated by the normal when the number of trials n is reasonably large.

For a given n, the approximation is best when π is not very close to 0 or 1. One rule-of-thumb is that the approximation is good enough when $n\,\pi > 5$ and $n\,(1-\pi) > 5$. Illustrations of the approximation are shown in Figure 4.13 for different values of n and π. Each plot shows values of the pf of $\text{Bin}(n, \pi)$, and the pdf of the normal approximation, $N(n\,\pi, n\,\pi\,(1-\pi))$.

When the normal approximation is appropriate, we can calculate probabilities for $X \sim \text{Bin}(n, \pi)$ using $Y \sim N(n\,\pi, n\,\pi\,(1-\pi))$ and Table 4 of the *New Cambridge Statistical Tables*.

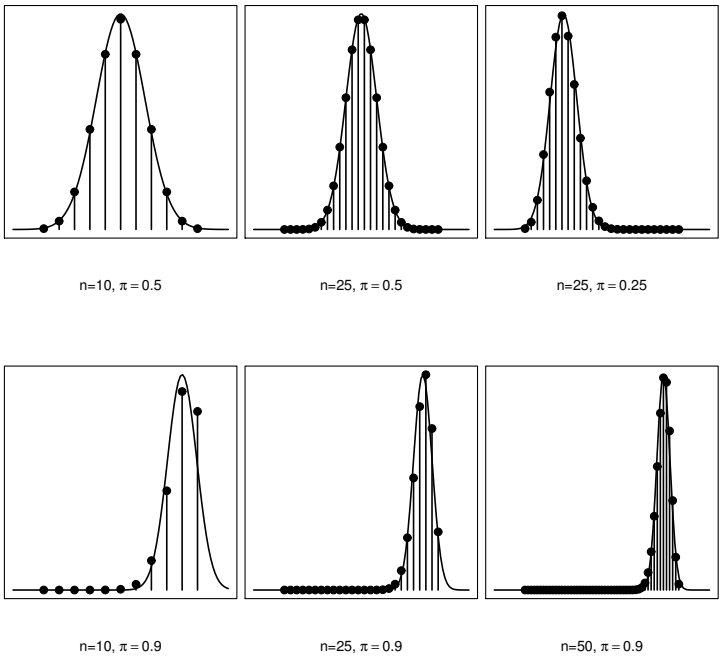


Figure 4.13: Examples of the normal approximation of the binomial distribution.

Unfortunately, there is one small caveat. The binomial distribution is discrete, but the normal distribution is continuous. To see why this is problematic, consider the following. Suppose $X \sim \text{Bin}(40,\, 0.4)$. Since X is discrete, such that $x = 0, 1, \ldots, 40$, then:

$$P(X \leq 4) = P(X \leq 4.5) = P(X < 5)$$

since $P(4 < X \leq 4.5) = 0$ and $P(4.5 < X < 5) = 0$ due to the 'gaps' in the probability mass for this distribution. In contrast if $Y \sim N(16,\, 9.6)$, then:

$$P(Y \leq 4) < P(Y \leq 4.5) < P(Y < 5)$$

since $P(4 < Y < 4.5) > 0$ and $P(4.5 < Y < 5) > 0$ because this is a continuous distribution.

The accepted way to circumvent this problem is to use a **continuity correction** which corrects for the effects of the transition from a discrete $\text{Bin}(n,\, \pi)$ distribution to a continuous $N(n\,\pi,\, n\,\pi\,(1-\pi))$ distribution.

Continuity correction

This technique involves representing each discrete binomial value x, for $0 \leq x \leq n$, by the continuous interval $(x-0.5,\, x+0.5)$. Great care is needed to determine which x values are included in the required probability. Suppose we are approximating $X \sim \text{Bin}(n,\, \pi)$ with $Y \sim N(n\,\pi,\, n\,\pi\,(1-\pi))$, then:

$$\begin{aligned}
P(X < 4) = P(X \leq 3) &\Rightarrow P(Y < 3.5) && \text{(since 4 is excluded)} \\
P(X \leq 4) = P(X < 5) &\Rightarrow P(Y < 4.5) && \text{(since 4 is included)} \\
P(1 \leq X < 6) = P(1 \leq X \leq 5) &\Rightarrow P(0.5 < Y < 5.5) && \text{(since 1 to 5 are included).}
\end{aligned}$$

Example 4.16 In the UK general election in May 2010, the Conservative Party received 36.1% of the votes. We carry out an opinion poll in November 2014, where we survey 1,000 people who say they voted in 2010, and ask who they would vote for if a general election was held now. Let X denote the number of people who say they would now vote for the Conservative Party.

Suppose we assume that $X \sim \text{Bin}(1000, 0.361)$.

1. What is the probability that $X \geq 400$?

   Using the normal approximation, noting $n = 1000$ and $\pi = 0.361$, with $Y \sim N(1000 \times 0.361, 1000 \times 0.361 \times 0.639) = N(361, 230.68)$, we get:

$$\begin{aligned} P(X \geq 400) &\approx P(Y \geq 399.5) \\ &= P\left(\frac{Y - 361}{\sqrt{230.68}} \geq \frac{399.5 - 361}{\sqrt{230.68}}\right) \\ &= P(Z \geq 2.53) \\ &= 1 - \Phi(2.53) \\ &= 0.0057. \end{aligned}$$

   The exact probability from the binomial distribution is $P(X \geq 400) = 0.0059$. Without the continuity correction, the normal approximation would give 0.0051.

2. What is the largest number x for which $P(X \leq x) < 0.01$?

   We need the largest x which satisfies:

$$P(X \leq x) \approx P(Y \leq x + 0.5) = P\left(Z \leq \frac{x + 0.5 - 361}{\sqrt{230.68}}\right) < 0.01.$$

   According to Table 4 of the *New Cambridge Statistical Tables*, the smallest z which satisfies $P(Z \geq z) < 0.01$ is $z = 2.33$, so the largest z which satisfies $P(Z \leq z) < 0.01$ is $z = -2.33$. We then need to solve:

$$\frac{x + 0.5 - 361}{\sqrt{230.68}} \leq -2.33$$

   which gives $x \leq 325.1$. The smallest integer value which satisfies this is $x = 325$. Therefore $P(X \leq x) < 0.01$ for all $x \leq 325$.

   The sum of the exact binomial probabilities from 0 to x is 0.0093 for $x = 325$, and 0.011 for $x = 326$. The normal approximation gives exactly the correct answer in this instance.

3. Suppose that 300 respondents in the actual survey say they would vote for the Conservative Party now. What do you conclude from this?

   From the answer to Question 2, we know that $P(X \leq 300) < 0.01$, *if* $\pi = 0.361$. In other words, *if* the Conservatives' support remains 36.1%, we would be very unlikely to get a random sample where only 300 (or fewer) respondents would say they would vote for the Conservative Party.

   Now $X = 300$ *is* actually observed. We can then conclude one of two things (if we exclude other possibilities, such as a biased sample or lying by the respondents):

(a) The Conservatives' true level of support is still 36.1% (or even higher), but by chance we ended up with an unusual sample with only 300 of their supporters.

(b) The Conservatives' true level of support is currently less than 36.1% (in which case getting 300 in the sample would be more probable).

Here (b) seems a more plausible conclusion than (a). This kind of reasoning is the basis of statistical significance tests.

4.8 Overview of chapter

This chapter has introduced some common discrete and continuous probability distributions. Their properties, uses and applications have been discussed. The relationships between some of these distributions have also been discussed.

4.9 Key terms and concepts

- Bernoulli distribution
- Binomial distribution
- Central limit theorem
- Continuity correction
- Exponential distribution
- Normal distribution
- Parameter
- Poisson distribution
- Population distribution
- Standardised variable
- Uniform distribution
- z-score

4.10 Learning activities

1. London Underground trains on the Northern Line have a probability of 0.05 of failure between Golders Green and King's Cross. Supposing that the failures are all independent, what is the probability that out of 10 journeys between Golders Green and King's Cross more than 8 do not have a breakdown?

2. Suppose that the normal rate of infection for a certain disease in cattle is 25%. To test a new serum which may prevent infection, three experiments are carried out. The test for infection is not always valid for some particular cattle, so the experimental results are incomplete – we cannot always tell whether a cow is infected or not. The results of the three experiments are:

   i. 10 animals are injected; all 10 remain free from infection.

   ii. 17 animals are injected; more than 15 remain free from infection and there are 2 doubtful cases.

   iii. 23 animals are infected; more than 20 remain free from infection and there are 3 doubtful cases.

   Which experiment provides the strongest evidence in favour of the serum?

3. In a large industrial plant there is an accident on average every two days.
   (a) What is the chance that there will be exactly two accidents in a given week?
   (b) Repeat (a) for the chance of two or more accidents in a given week.
   (c) If Karen goes to work there for a four-week period what is the probability that no accident occurs while she is there?

4. The chance that a lottery ticket has a winning number is 0.0000001. Suppose 10,000,000 people buy tickets that are independently numbered.
   (a) What is the probability there is no winner?
   (b) What is the probability there is exactly 1 winner?
   (c) What is the probability there are exactly 2 winners?

5. Suppose that $X \sim \text{Uniform}[0, 1]$. Compute $P(X > 0.2)$, $P(X \geq 0.2)$ and $P(X^2 > 0.04)$.

6. Suppose that the service time for a customer at a fast food outlet has an exponential distribution with parameter 1/3 (customers per minute). What is the probability that a customer waits more than 4 minutes?

7. Suppose that the distribution of men's heights in London, measured in cm, is $N(175, 6^2)$. Find the proportion of men whose height is:
   (a) under 169 cm
   (b) over 190 cm
   (c) between 169 cm and 190 cm.

8. Two statisticians disagree about the distribution of IQ scores for a population under study. Both agree that the distribution is normal, and that $\sigma = 15$, but A says that 5% of the population have IQ scores greater than 134.6735, whereas B says that 10% of the population have IQ scores greater than 109.224. What is the difference between the mean IQ score as assessed by A and that as assessed by B?

9. Helmut goes fishing every Saturday. The number of fish he catches follows a Poisson distribution. On a proportion p of the days he goes fishing, he does not catch anything. He makes it a rule to take home the first fish and then every other fish that he catches (i.e. the first, third, fifth fish and so on).
   (a) Using a Poisson distribution, find the mean number of fish he catches.
   (b) Show that the probability that he takes home the last fish he catches is $(1 - p^2)/2$.

Solutions to these questions can be found on the VLE in the **ST104b Statistics 2** area at http://my.londoninternational.ac.uk

4.11 Reminder of learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- summarise basic distributions such as the uniform, Bernoulli, binomial, Poisson, exponential and normal
- calculate probabilities of events for these distributions using the probability function, probability density function or cumulative distribution function
- determine probabilities using statistical tables, where appropriate
- state properties of these distributions such as the expected value and variance.

4.12 Sample examination questions

1. A doctor wishes to procure subjects possessing a certain chromosome abnormality which is present in 4% of the population. How many randomly chosen independent subjects should be procured if the doctor wishes to be 95% confident that at least one subject has the abnormality?

2. At one stage in the manufacture of an article a piston of circular cross-section has to fit into a similarly-shaped cylinder. The distributions of diameters of pistons and cylinders are known to be normal with the following parameters:
   - Piston diameters: mean 10.42 cm, standard deviation 0.03 cm.
   - Cylinder diameters: mean 10.52 cm, standard deviation 0.04 cm.

   (a) If pairs of pistons and cylinders are selected at random for assembly, for what proportion will the piston not fit into the cylinder (i.e. for which the piston diameter exceeds the cylinder diameter)?

   (b) What is the chance that in 100 pairs, selected at random:
   i. every piston will fit
   ii. not more than two of the pistons will fail to fit?

   (c) Calculate the probabilities in (b) using a Poisson approximation. Discuss the appropriateness of using this approximation.

3. Show that, for a binomial random variable such that $X \sim \text{Bin}(n, \pi)$, we have:

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

   Hence find $\text{E}(X)$ and $\text{Var}(X)$.

   [The wording of the question implies that you use the result that you have just proved. Other methods of derivation will not be accepted!]

4. Cars independently pass a point on a busy road at an average rate of 150 per hour.
   (a) Assuming a Poisson distribution, find the probability that none passes in a given minute.
   (b) What is the expected number passing in two minutes?
   (c) Find the probability that the expected number actually passes in a given two-minute period.

   Other motor vehicles (vans, motorcycles etc.) pass the same point independently at the rate of 75 per hour. Assume a Poisson distribution for these vehicles too.

   (d) What is the probability of one car and one other motor vehicle in a two-minute period?

Solutions to these questions can be found on the VLE in the **ST104b Statistics 2** area at http://my.londoninternational.ac.uk