# ST102/ST109
# Elementary Statistical Theory

## Course pack

## 2023/24 (Autumn term)

Dr James Abdey

lse.ac.uk/statistics

# ST102/ST109

# Elementary Statistical Theory

# Course pack

# Contents

**Preliminaries** **ix**

0.1 Organisation . . . . . ix

0.2 Course materials . . . . . ix

0.3 Supplementary reading . . . . . ix

0.4 Assessment . . . . . x

0.5 Syllabus . . . . . xi

0.6 And finally, words of wisdom from a wise (youngish) man . . . . . xi

**Preface** **xiii**

0.7 The role of statistics in the research process . . . . . xiii

**1 Data visualisation and descriptive statistics** **1**

1.1 Synopsis of chapter . . . . . 1

1.2 Learning outcomes . . . . . 1

1.3 Introduction . . . . . 1

1.4 Continuous and discrete variables . . . . . 4

1.5 The sample distribution . . . . . 5

1.5.1 Bar charts . . . . . 6

1.5.2 Sample distributions of variables with many values . . . . . 7

1.5.3 Skewness of distributions . . . . . 8

1.6 Measures of central tendency . . . . . 8

1.6.1 Notation for variables . . . . . 9

1.6.2 Summation notation . . . . . 9

1.6.3 The sample mean . . . . . 10

1.6.4 The (sample) median . . . . . 12

1.6.5 Sensitivity to outliers . . . . . 13

1.6.6 Skewness, means and medians . . . . . 13

1.6.7 Mode . . . . . 14

1.7 Measures of dispersion . . . . . 14

1.7.1 Variance and standard deviation . . . . . 15

1.7.2 Sample quantiles . . . . . 16

1.7.3 Quantile-based measures of dispersion . . . . . . . . . . . . . . . . 17
1.7.4 Boxplots . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17
1.8 Associations between two variables . . . . . . . . . . . . . . . . . . . 18
1.8.1 Scatterplots . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18
1.8.2 Line plots (time series plots) . . . . . . . . . . . . . . . . . . . . . 19
1.8.3 Side-by-side boxplots for comparisons . . . . . . . . . . . . . . . . 20
1.8.4 Two-way contingency tables . . . . . . . . . . . . . . . . . . . . . 20
1.9 Overview of chapter . . . . . . . . . . . . . . . . . . . . . . . . . . . 22
1.10 Key terms and concepts . . . . . . . . . . . . . . . . . . . . . . . . . 22

**2 Probability theory 23**
2.1 Synopsis of chapter . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
2.2 Learning outcomes . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
2.3 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
2.4 Set theory: the basics . . . . . . . . . . . . . . . . . . . . . . . . . . 25
2.5 Axiomatic definition of probability . . . . . . . . . . . . . . . . . . . 31
2.5.1 Basic properties of probability . . . . . . . . . . . . . . . . . . . . 32
2.6 Classical probability and counting rules . . . . . . . . . . . . . . . . . 36
2.6.1 Brute force: listing and counting . . . . . . . . . . . . . . . . . . . 38
2.6.2 Combinatorial counting methods . . . . . . . . . . . . . . . . . . . 38
2.6.3 Combining counts: rules of sum and product . . . . . . . . . . . . 43
2.7 Conditional probability and Bayes' theorem . . . . . . . . . . . . . . . 44
2.7.1 Independence of multiple events . . . . . . . . . . . . . . . . . . . 45
2.7.2 Independent versus mutually exclusive events . . . . . . . . . . . . 46
2.7.3 Conditional probability of independent events . . . . . . . . . . . 48
2.7.4 Chain rule of conditional probabilities . . . . . . . . . . . . . . . . 48
2.7.5 Total probability formula . . . . . . . . . . . . . . . . . . . . . . . 50
2.7.6 Bayes' theorem . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52
2.8 Overview of chapter . . . . . . . . . . . . . . . . . . . . . . . . . . . 56
2.9 Key terms and concepts . . . . . . . . . . . . . . . . . . . . . . . . . 56

**3 Random variables 59**
3.1 Synopsis of chapter . . . . . . . . . . . . . . . . . . . . . . . . . . . 59
3.2 Learning outcomes . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59
3.3 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59
3.4 Discrete random variables . . . . . . . . . . . . . . . . . . . . . . . . 61

3.4.1 Probability distribution of a discrete random variable . . . . . . . 61
3.4.2 The cumulative distribution function (cdf) . . . . . . . . . . . . . 65
3.4.3 Properties of the cdf for discrete distributions . . . . . . . . . . . 66
3.4.4 General properties of the cdf . . . . . . . . . . . . . . . . . . . . . 66
3.4.5 Properties of a discrete random variable . . . . . . . . . . . . . . 67
3.4.6 Expected value versus sample mean . . . . . . . . . . . . . . . . . 68
3.4.7 Moments of a random variable . . . . . . . . . . . . . . . . . . . . 73
3.4.8 The moment generating function . . . . . . . . . . . . . . . . . . 75
3.5 Continuous random variables . . . . . . . . . . . . . . . . . . . . . . . . 77
3.5.1 Moment generating functions . . . . . . . . . . . . . . . . . . . . 86
3.5.2 Median of a random variable . . . . . . . . . . . . . . . . . . . . . 86
3.6 Overview of chapter . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 88
3.7 Key terms and concepts . . . . . . . . . . . . . . . . . . . . . . . . . . . 88

**4 Common distributions of random variables 89**
4.1 Synopsis of chapter content . . . . . . . . . . . . . . . . . . . . . . . . . 89
4.2 Learning outcomes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89
4.3 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89
4.4 Common discrete distributions . . . . . . . . . . . . . . . . . . . . . . . 91
4.4.1 Discrete uniform distribution . . . . . . . . . . . . . . . . . . . . . 91
4.4.2 Bernoulli distribution . . . . . . . . . . . . . . . . . . . . . . . . . 92
4.4.3 Binomial distribution . . . . . . . . . . . . . . . . . . . . . . . . . 92
4.4.4 Poisson distribution . . . . . . . . . . . . . . . . . . . . . . . . . . 95
4.4.5 Connections between probability distributions . . . . . . . . . . . 98
4.4.6 Poisson approximation of the binomial distribution . . . . . . . . 98
4.4.7 Some other discrete distributions . . . . . . . . . . . . . . . . . . 100
4.5 Common continuous distributions . . . . . . . . . . . . . . . . . . . . . . 101
4.5.1 The (continuous) uniform distribution . . . . . . . . . . . . . . . . 101
4.5.2 Exponential distribution . . . . . . . . . . . . . . . . . . . . . . . . 102
4.5.3 Two other distributions . . . . . . . . . . . . . . . . . . . . . . . . 104
4.5.4 Normal (Gaussian) distribution . . . . . . . . . . . . . . . . . . . . 105
4.5.5 Normal approximation of the binomial distribution . . . . . . . . 112
4.6 Overview of chapter . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 115
4.7 Key terms and concepts . . . . . . . . . . . . . . . . . . . . . . . . . . . 115

**5 Multivariate random variables 117**

5.1 Synopsis of chapter . . . 117

5.2 Learning outcomes . . . 117

5.3 Introduction . . . 117

5.4 Joint probability functions . . . 118

5.5 Marginal distributions . . . 120

5.6 Continuous multivariate distributions . . . 121

5.7 Conditional distributions . . . 122

5.7.1 Properties of conditional distributions . . . 124

5.7.2 Conditional mean and variance . . . 125

5.7.3 Continuous conditional distributions . . . 125

5.8 Covariance and correlation . . . 126

5.8.1 Covariance . . . 127

5.8.2 Correlation . . . 128

5.8.3 Sample covariance and correlation . . . 129

5.9 Independent random variables . . . 131

5.9.1 Joint distribution of independent random variables . . . 132

5.10 Sums and products of random variables . . . 133

5.10.1 Distributions of sums and products . . . 133

5.10.2 Expected values and variances of sums of random variables . . . 134

5.10.3 Expected values of products of independent random variables . . . 135

5.10.4 Some proofs of previous results . . . 135

5.10.5 Distributions of sums of random variables . . . 136

5.11 Overview of chapter . . . 138

5.12 Key terms and concepts . . . 138

**A Data visualisation and descriptive statistics 139**

A.1 (Re)vision of fundamentals . . . 139

A.2 Worked example . . . 144

A.3 Practice questions . . . 145

**B Probability theory 147**

B.1 Worked examples . . . 147

B.2 Practice questions . . . 158

**C Random variables 161**

C.1 Worked examples . . . 161

C.2 Practice questions . . . . . 172

D Common distributions of random variables 175

D.1 Worked examples . . . . . 175

D.2 Practice questions . . . . . 188

E Multivariate random variables 191

E.1 Worked examples . . . . . 191

E.2 Practice questions . . . . . 201

F Solutions to Practice questions 203

F.1 Chapter 1 – Data visualisation and descriptive statistics . . . . . 203

F.2 Chapter 2 – Probability theory . . . . . 205

F.3 Chapter 3 – Random variables . . . . . 208

F.4 Chapter 4 – Common distributions of random variables . . . . . 210

F.5 Chapter 5 – Multivariate random variables . . . . . 215

# Preliminaries

## 0.1 Organisation

- **Course lecturer:** Dr. James Abdey.
  - Email: J.S.Abdey@lse.ac.uk
  - Office hours: see the Student Hub.
- **Lectures:**
  - Tuesdays 09:00–11:00, Peacock Theatre, Autumn term weeks 1–10.

  Lecture recordings will be made available via Moodle, but note these should *not* be a substitute for attending lectures!
- **Example workshops (optional):**
  - Tuesdays 15:00–16:00, Peacock Theatre, Autumn term weeks 1–11.

  Recordings of these sessions will also be available.
- **Classes (90-minute duration):**
  - Autumn term weeks 2–11.

## 0.2 Course materials

All course materials will be available on Moodle. Your principal resource should be this course pack featuring (i) lecture material (which covers the entire syllabus for the Autumn term in detail[1]), and (ii) an examples manual. *You are strongly advised to read the relevant material in these notes* ***before*** *the corresponding lecture – this initial exposure to the material will enhance your understanding of the lecture itself!*

The optional 'Example workshops' will, unsurprisingly, be workshops of extra examples. No new material will be covered in these sessions – instead I will go through examination-style questions providing further practice of the course material. References to the examples manual will be made in these sessions.

## 0.3 Supplementary reading

- *In principle*, the course materials should be sufficient to fully understand all the exciting topics which you will encounter on this statistical journey of discovery.

[1] Conditional on your artistic tendencies, feel free to run wild with highlighters, a spectrum of Post-it notes etc. to annotate this tome. With luck, you may be referring back to this at a later stage in your studies!

However, you will all be aware of the considerable, dare I say *significant*, heterogeneity within the student body. Therefore, some may seek additional reassurance in the form of recommended reading.[2]

- I should stress that purchase of a textbook is at your sole discretion. I suspect this decision will be based on the extent to which your student finances have been studiously managed to date. For the shopaholics among you, the recommended text is:

  - Larsen, R.J. and M.J. Marx (2017). *An Introduction to Mathematical Statistics and Its Applications*, Pearson, sixth edition.[3]

- Of course, numerous titles are available covering the topics frequently found in first-year undergraduate service-level courses in statistics. Again, due to the doubtless heterogeneous preferences among you all, some may find one author's style readily accessible, while others may despair at the baffling presentation of material.[4] Consequently, my best advice would be to *sample*[5] a range of textbooks, and choose your preferred one. Any textbook would act as a supplement to the course materials. In particular, textbooks are filled with additional exercises to check understanding – and if you're lucky, they'll give you (some) solutions too!

## 0.4 Assessment

- Classes will involve going through the solutions to exercises, and full solutions will be made available on Moodle upon completion of all that week's classes. Further unseen problems will also be covered. As if the sheer joy of studying the discipline was not incentive enough to engage with the exercises, they are the *best* preparation for the…

  - **two-hour written examination in week 0 of Winter term** – this has 50% weighting for ST102, and 100% weighting for ST109. For details, please see the 'Past exam papers (January)' section of Moodle.

- **Statistical tables** are provided in the examination. Specifically, these will be from:

  Murdoch, J. and J.A. Barnes *Statistical Tables*, fourth edition (probably), Palgrave Macmillan.

  You do **not** need to purchase this. Relevant abstracts (Tables 3, 7, 8 and 9) are provided in electronic form on Moodle.

- You will need a scientific calculator for both the classes and the examination. The only permitted calculators for **in-person examinations** are the **Casio fx-83** or **fx-85** range, available from many retailers.

[2] Of course, if I've done my job properly, these notes should transcend all other publications!

[3] Second-hand earlier editions will be just as valid.

[4] One clearly hopes the former applies to this humble author.

[5] A sacred word in statistics.

## 0.5 Syllabus

- The full syllabus for ST102 (Autumn term) and ST109 Elementary Statistical Theory can be found in the table of contents.

## 0.6 And finally, words of wisdom from a wise (youngish) man

- (Not so) many years ago, I too was an undergraduate. Fresh-faced and full of enthusiasm (some things never change), I discovered the strategy for success in statistics.[6] As you embark on this statistical voyage, I feel compelled to share the following with you.

- Statistics is fundamentally a *cumulative* discipline – that is, the following chapters are not a series of self-contained units, rather they build on each other sequentially. As such, you are strongly advised to thoroughly study and understand all topics, since accruing this knowledge will make it easier to make sense of later topics.

- **Repetition** is the key to success. **Repetition** is the key to success. Familiarity breeds recognition of how to solve problems (and hence 'examination questions'), so repeatedly attempting exercise sets and related questions in the examples manual is highly recommended. To illustrate this point, consider the following:

  > Cdnuolt blveiee taht I cluod aulaclty uesdnatnrd waht I was rdanieg. The phaonmneal pweor of the hmuan mnid, aoccdrnig to rarseech at Cmabrigde Uinervtisy. It dn'seot mttaer in waht oredr the ltteers in a wrod are, the olny iprmoatnt tihng is taht the frist and lsat ltteer be in the rghit pclae. The rset can be a taotl mses and you can sitll raed it wouthit a porbelm.

  Contrast this with:

  > Miittluvraae asilyans sattes an idtenossiy ctuoonr epilsle is the itternoiecson of a panle pleralal to the $xl$-$y$apne and the sruacfe of a btiiarave nmarol dbttiisruein.

  I suspect most of you could pretty much follow the first passage due to your familiarity with most of the 'true' words. The second passage is constructed in exactly the same way, but is probably harder to comprehend due to your *relative* unfamiliarity with these words and the difficulty of the concepts involved.[7] Therefore, repeated exposure to something eases comprehension and your capacity to perform well in statistics follows this exact idea.

[6] I expect this approach is not restricted to statistics, so feel free to apply it in an interdisciplinary setting.

[7] For the interested reader, the correct 'translation' is 'Multivariate analysis states an isodensity contour ellipse is the intersection of a plane parallel to the $xy$-plane and the surface of a bivariate normal distribution'. And before you panic, no we will not be covering isodensity contour ellipses!

- So, to conclude, perseverance with problem solving is your passport to a strong examination performance. Attempting (ideally succesfully!) the weekly exercise sets is of paramount importance. Here endeth the first lesson.

*Practise, practise, practise.*
(James Abdey)

# Preface

*Torture numbers, and they'll confess to anything.*
(Gregg Easterbrook)

## 0.7 The role of statistics in the research process

Before we get into details, let us begin with the 'big picture'. First, some definitions.

- **Research**: trying to answer questions about the world in a systematic (scientific) way.
- **Empirical** research: doing research by first collecting relevant information (*data*) about the world.

Research may be about almost any topic: physics, biology, medicine, economics, history, literature etc. Most of our examples will be from the social sciences: economics, management, finance, sociology, political science, psychology etc. Research in this sense is not just what universities do. Governments, businesses, and all of us as individuals do it too. Statistics is used in essentially the same way for all of these.

**Example 0.1** It all starts with a question.

- Can labour regulation hinder economic performance?
- Understanding the gender pay gap: what has competition got to do with it?
- Children and online risk: powerless victims or resourceful participants?
- Refugee protection as a collective action problem: is the European Union (EU) shirking its responsibilities?
- Do directors perform for pay?
- Heeding the push from below: how do social movements persuade the rich to listen to the poor?
- Does devolution lead to regional inequalities in welfare activity?
- The childhood origins of adult socio-economic disadvantage: do cohort and gender matter?
- Parental care as unpaid family labour: how do spouses share?

### Key stages of the empirical research process

We can think of the empirical research process as having five key stages.

1. Formulating the *research question*.
2. *Research design*: deciding what kinds of data to collect, how and from where.
3. *Collecting* the data.
4. *Analysis* of the data to answer the research question.
5. *Reporting* the answer and how it was obtained.

The main job of statistics is the *analysis* of data, although it also informs other stages of the research process. Statistics are used when the data are **quantitative**, i.e. in the form of numbers.

*Statistical* analysis of quantitative data has the following features.

- It can cope with large volumes of data, in which case the first task is to provide an understandable summary of the data. This is the job of **descriptive statistics**.
- It can deal with situations where the observed data are regarded as only a part (a *sample*) from all the data which could have been obtained (the *population*). There is then *uncertainty* in the conclusions. Measuring this uncertainty is the job of **statistical inference**.

We conclude this preface with an example of how statistics can be used to help answer a research question.

**Example 0.2** **CCTV, crime and fear of crime.**

Our research question is what is the effect of closed-circuit television (CCTV) surveillance on:

- the number of recorded crimes?
- the fear of crime felt by individuals?

We illustrate this using part of the following study.

- Gill and Spriggs (2005): Assessing the impact of CCTV. *Home Office Research Study 292.*

The research design of the study comprised the following.

- **Target area**: a housing estate in northern England.
- **Control area**: a second, comparable housing estate.

- **Intervention**: CCTV cameras installed in the target area but not in the control area.
- **Compare** measures of crime and the fear of crime in the target and control areas in the 12 months before and 12 months after the intervention.

The data and data collection were as follows.

- **Level of crime**: the number of crimes recorded by the police, in the 12 months before and 12 months after the intervention.
- **Fear of crime**: a *survey* of residents of the areas.
  - Respondents: random samples of residents in each of the areas.
  - In each area, one sample before the intervention date and one about 12 months after.
  - Sample sizes:

| | Before | After |
|---|---|---|
| Target area | 172 | 168 |
| Control area | 215 | 242 |

  - Question considered here: 'In general, how much, if at all, do you worry that you or other people in your household will be victims of crime?' (from 1 = 'all the time' to 5 = 'never').

Statistical analysis of the data.

% of respondents who worry 'sometimes', 'often' or 'all the time':

| Target | | | Control | | | | |
|---|---|---|---|---|---|---|---|
| [*a*] | [*b*] | | [*c*] | [*d*] | | | Confidence |
| Before | After | Change | Before | After | Change | RES | interval |
| 26 | 23 | −3 | 53 | 46 | −7 | 0.98 | 0.55–1.74 |

- It is possible to calculate various statistics, for example the *Relative Effect Size* RES = $([d]/[c])/([b]/[a]) = 0.98$ is a summary measure which compares the changes in the two areas.
- RES $< 1$, which means that the observed change in the reported fear of crime has been a bit *less* good in the target area.
- However, there is *uncertainty because of sampling*: only 168 and 242 individuals were actually interviewed at each time in each area, respectively.
- The *confidence interval* for RES includes 1, which means that changes in the self-reported fear of crime in the two areas are 'not statistically significantly different' from each other.

The number of (any kind of) recorded crimes:

| Target area | | | Control area | | | | |
|---|---|---|---|---|---|---|---|
| [*a*] | [*b*] | | [*c*] | [*d*] | | | Confidence |
| Before | After | Change | Before | After | Change | RES | interval |
| 112 | 101 | −11 | 73 | 88 | 15 | 1.34 | 0.79–1.89 |

- Now the RES $>$ 1, which means that the observed change in the number of crimes has been worse in the control area than in the target area.
- However, the numbers of crimes in each area are fairly small, which means that these estimates of the changes in crime rates are fairly uncertain.
- The confidence interval for RES again includes 1, which means that the changes in crime rates in the two areas are not statistically significantly different from each other.

In summary, this study did *not* support the claim that the introduction of CCTV reduces crime or the fear of crime.

- If you want to read more about research of this question, see Welsh and Farrington (2008). Effects of closed circuit television surveillance on crime. *Campbell Systematic Reviews 2008:17.*

Many of the statistical terms and concepts mentioned above have not been explained yet – that is what the rest of the course is for! However, it serves as an interesting example of how statistics can be employed in the social sciences to investigate research questions.

# Chapter 1

# Data visualisation and descriptive statistics

## 1.1 Synopsis of chapter

Graphical representations of data provide us with a useful view of the **distribution** of variables. In this chapter, we shall cover a selection of approaches for displaying data visually – each being appropriate in certain situations. We then consider **descriptive statistics**, whose main objective is to interpret key features of a dataset numerically. Graphs and charts have little intrinsic value *per se*, however their main function is to bring out interesting features of a dataset. For this reason, simple descriptions should be preferred to complicated graphics.

Although data visualisation is useful as a preliminary form of data analysis to get a 'feel' for the data, in practice we also need to be able to summarise data numerically. We introduce descriptive statistics and distinguish between measures of location, measures of dispersion and skewness. All these statistics provide useful summaries of raw datasets.

## 1.2 Learning outcomes

At the end of this chapter, you should be able to:

- interpret and summarise raw data on social science variables graphically
- interpret and summarise raw data on social science variables numerically
- calculate basic measures of location and dispersion
- describe the skewness of a distribution and interpret boxplots
- discuss the key terms and concepts introduced in the chapter.

## 1.3 Introduction

Starting point: a collection of numerical data (a *sample*) has been collected in order to answer some questions. Statistical analysis may have two broad aims.

1. **Descriptive statistics**: summarise the data which were collected, in order to make them more understandable.

2. **Statistical inference**: use the observed data to draw conclusions about some broader population.

Sometimes '1.' is the only aim. Even when '2.' is the main aim, '1.' is still an essential first step.

Data do *not* just speak for themselves. There are usually simply too many numbers to make sense of just by staring at them. Descriptive statistics attempt to **summarise some key features** of the data to make them understandable and **easy to communicate**. These summaries may be **graphical** or **numerical** (tables or individual summary statistics).

**Example 1.1** We consider data for 155 countries on three variables from around 2002. The data can be found in the file '`Countries.csv`'. The variables are the following.

- **Region** of the country.
  - This is a *nominal* variable coded (in alphabetical order) as follows: 1 = Africa, 2 = Asia, 3 = Europe, 4 = Latin America, 5 = Northern America, 6 = Oceania.
- The **level of democracy**, i.e. a democracy index, in the country.
  - This is an 11-point *ordinal* scale from 0 (lowest level of democracy) to 10 (highest level of democracy).
- Gross domestic product per capita (**GDP per capita**) (i.e. per person, in $000s) which is a *ratio* scale.

The statistical data in a sample are typically stored in a **data matrix**, as shown in Figure 1.1.

*Rows* of the data matrix correspond to different **units** (subjects/observations).

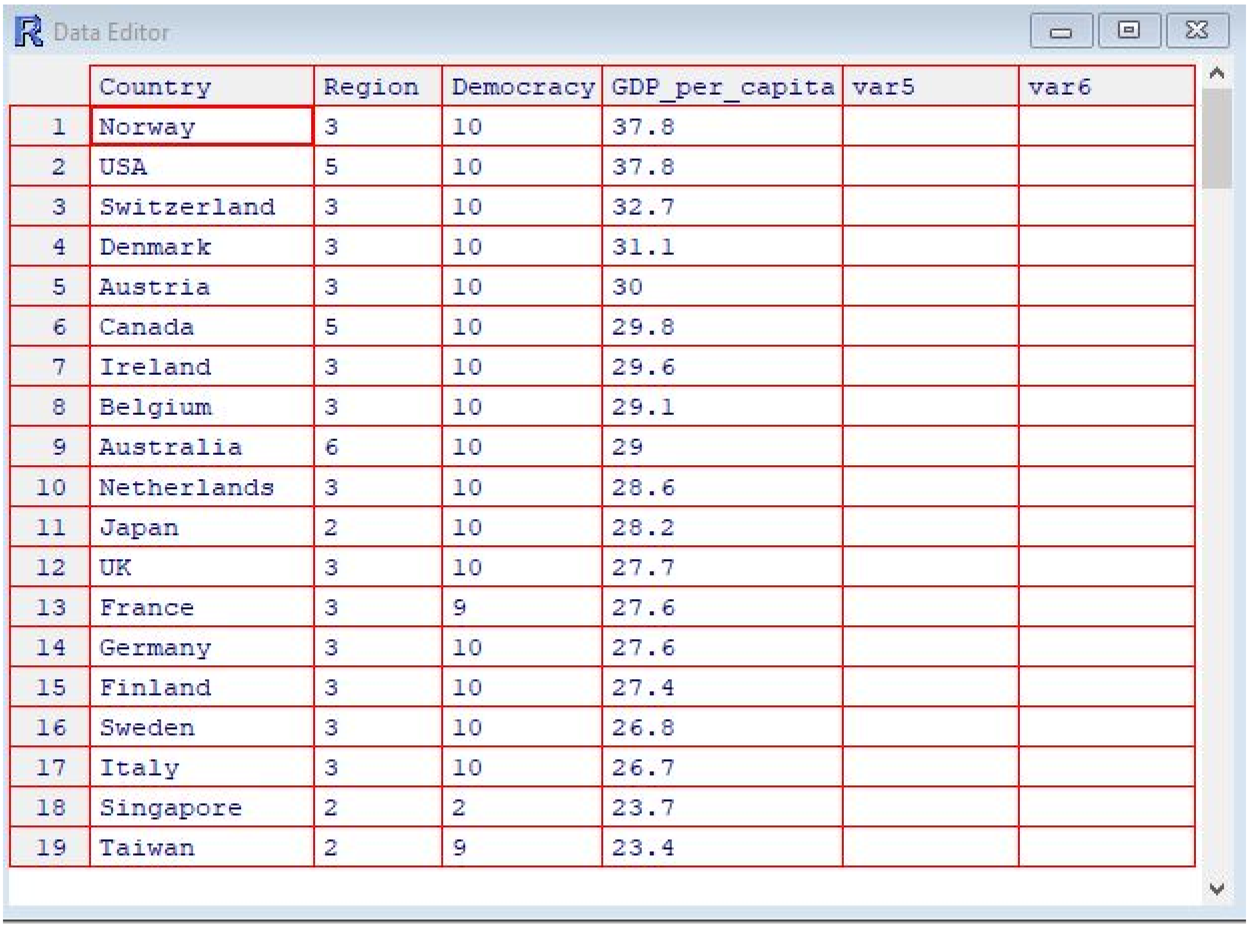
- Here, each unit is a country.

The number of units in a dataset is the **sample size**, typically denoted by $n$.

- Here, $n = 155$ countries.

*Columns* of the data matrix correspond to **variables**, i.e. different characteristics of the units.

- Here, region, the level of democracy, and GDP per capita are the variables.



| | Country | Region | Democracy | GDP_per_capita | var5 | var6 |
|---|---|---|---|---|---|---|
| 1 | Norway | 3 | 10 | 37.8 | | |
| 2 | USA | 5 | 10 | 37.8 | | |
| 3 | Switzerland | 3 | 10 | 32.7 | | |
| 4 | Denmark | 3 | 10 | 31.1 | | |
| 5 | Austria | 3 | 10 | 30 | | |
| 6 | Canada | 5 | 10 | 29.8 | | |
| 7 | Ireland | 3 | 10 | 29.6 | | |
| 8 | Belgium | 3 | 10 | 29.1 | | |
| 9 | Australia | 6 | 10 | 29 | | |
| 10 | Netherlands | 3 | 10 | 28.6 | | |
| 11 | Japan | 2 | 10 | 28.2 | | |
| 12 | UK | 3 | 10 | 27.7 | | |
| 13 | France | 3 | 9 | 27.6 | | |
| 14 | Germany | 3 | 10 | 27.6 | | |
| 15 | Finland | 3 | 10 | 27.4 | | |
| 16 | Sweden | 3 | 10 | 26.8 | | |
| 17 | Italy | 3 | 10 | 26.7 | | |
| 18 | Singapore | 2 | 2 | 23.7 | | |
| 19 | Taiwan | 2 | 9 | 23.4 | | |

**Figure 1.1:** Example of a data matrix.

## 1.4 Continuous and discrete variables

Different variables may have different properties. These determine which kinds of statistical methods are suitable for the variables.

### Continuous and discrete variables

A **continuous** variable can, in principle, take any real values within some interval.

- In Example 1.1, GDP per capita is continuous, taking any non-negative value.

A variable is **discrete** if it is not continuous, i.e. if it can only take certain values, but not any others.

- In Example 1.1, region and the level of democracy are discrete, with possible values of $1, 2, \ldots, 6$, and $0, 1, 2, \ldots, 10$, respectively.

Many discrete variables have only a *finite* number of possible values. In Example 1.1, the region variable has 6 possible values, and the level of democracy has 11 possible values. The simplest possibility is a **binary**, or **dichotomous**, variable, with just *two* possible values. For example, a person's sex could be recorded as 1 = female and 2 = male.[1]

A discrete variable can also have an unlimited number of possible values.

- For example, the number of visitors to a website in a day: $0, 1, 2, \ldots$.[2]

> **Example 1.2** In Example 1.1, the levels of democracy have a meaningful ordering, from less democratic to more democratic countries. The numbers assigned to the different levels must also be in this order, i.e. a larger number = more democratic.
>
> In contrast, different regions (Africa, Asia, Europe, Latin America, Northern America and Oceania) do not have such an ordering. The numbers used for the region variable are just labels for different regions. A different numbering (such as 6 = Africa, 5 = Asia, 1 = Europe, 3 = Latin America, 2 = Northern America and 4 = Oceania) would be just as acceptable as the one we originally used. Some statistical methods are appropriate for variables with both ordered and unordered values, some only in the ordered case. Unordered categories are **nominal** data; ordered categories are **ordinal** data.

[1] Note that because sex is a nominal variable, the coding is arbitrary. We could also have, for example, 0 = male and 1 = female, or 0 = female and 1 = male. However, it is important to remember which coding has been used!

[2] In practice, of course, there is a finite number of internet users in the world. However, it is reasonable to treat this variable as taking an unlimited number of possible values.

## 1.5 The sample distribution

The **sample distribution** of a variable consists of:

- a list of the values of the variable which are observed in the sample
- the number of times each value occurs (the **counts** or **frequencies** of the observed values).

When the number of different observed values is small, we can show the whole sample distribution as a **frequency table** of all the values and their frequencies.

**Example 1.3** Continuing with Example 1.1, the observations of the region variable in the sample are:

```
3  5  3  3  3  5  3  3  6  3  2  3  3  3  3

3  3  2  2  2  3  6  2  3  2  2  2  3  3  2

2  3  3  3  2  4  3  2  3  1  4  3  1  3  3

4  4  4  1  2  4  3  4  3  2  1  2  3  1  3

2  1  4  2  4  3  1  4  6  2  1  3  4  2  1

4  4  4  2  3  2  4  1  4  1  4  2  2  2  4

2  2  1  4  2  1  4  2  2  4  4  1  6  3  1

2  1  2  2  1  1  2  1  1  3  2  2  1  2  4

2  1  2  1  1  2  1  2  1  2  1  1  1  1  1

1  1  1  2  1  1  1  1  1  2  1  1  1  1  1

1  1  1  2  1
```

We may construct a frequency table for the region variable as follows:

| Region | Frequency (count) | Relative frequency (%) |
|---|---|---|
| (1) Africa | 48 | 100 × (48/155)<br>31.0 |
| (2) Asia | 44 | 28.4 |
| (3) Europe | 34 | 21.9 |
| (4) Latin America | 23 | 14.8 |
| (5) Northern America | 2 | 1.3 |
| (6) Oceania | 4 | 2.6 |
| Total | 155 | 100 |

Here '%' is the percentage of countries in a region, out of the 155 countries in the sample. his is a measure of **proportion** (that is, **relative frequency**).

Similarly, for the level of democracy, the frequency table is:

| Level of democracy | Frequency | % | Cumulative % |
|---|---|---|---|
| 0 | 35 | 22.6 | 22.6 |
| 1 | 12 | 7.7 | 30.3 |
| 2 | 4 | 2.6 | 32.9 |
| 3 | 6 | 3.9 | 36.8 |
| 4 | 5 | 3.2 | 40.0 |
| 5 | 5 | 3.2 | 43.2 |
| 6 | 12 | 7.7 | 50.9 |
| 7 | 13 | 8.4 | 59.3 |
| 8 | 16 | 10.3 | 69.6 |
| 9 | 15 | 9.7 | 79.3 |
| 10 | 32 | 20.6 | 100 |
| Total | 155 | 100 | |

'Cumulative %' for a value of the variable is the sum of the percentages for that value and all lower-numbered values.

### 1.5.1 Bar charts

A **bar chart** is the graphical equivalent of the table of frequencies. Figure 1.2 displays the region variable data as a bar chart. The relative frequencies of each region are clearly visible.

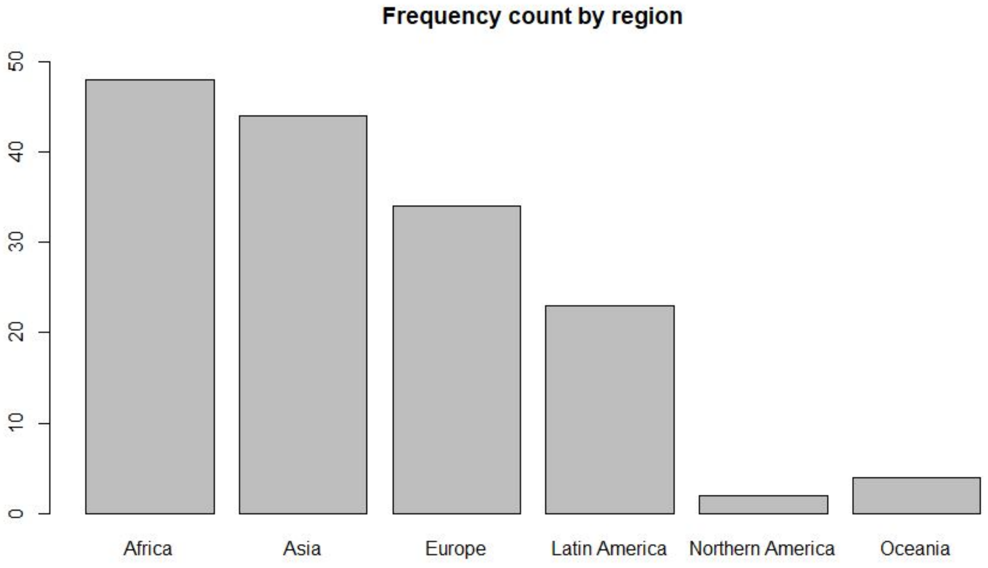


**Figure 1.2:** Example of a bar chart showing the region variable.

### 1.5.2 Sample distributions of variables with many values

If a variable has many distinct values, listing frequencies of all of them is not very practical.

A solution is to group the values into non-overlapping *intervals*, and produce a table or graph of the frequencies within the intervals. The most common graph used for this is a **histogram**.

A histogram is like a bar chart, but without gaps between bars, and often uses more bars (intervals of values) than is sensible in a table. Histograms are usually drawn using statistical software, such as R. You can let the software choose the intervals and the number of bars.

> **Example 1.4** Continuing with Example 1.1, a table of frequencies for GDP per capita where values have been grouped into non-overlapping intervals is shown below. Figure 1.3 shows a histogram of GDP per capita with a greater number of intervals to better display the sample distribution.

| GDP per capita (in $000s) | Frequency | % |
|---|---|---|
| $[0, 2)$ | 49 | 31.6 |
| $[2, 5)$ | 32 | 20.6 |
| $[5, 10)$ | 29 | 18.7 |
| $[10, 20)$ | 21 | 13.5 |
| $[20, 30)$ | 19 | 12.3 |
| $[30, 50)$ | 5 | 3.2 |
| Total | 155 | 100 |

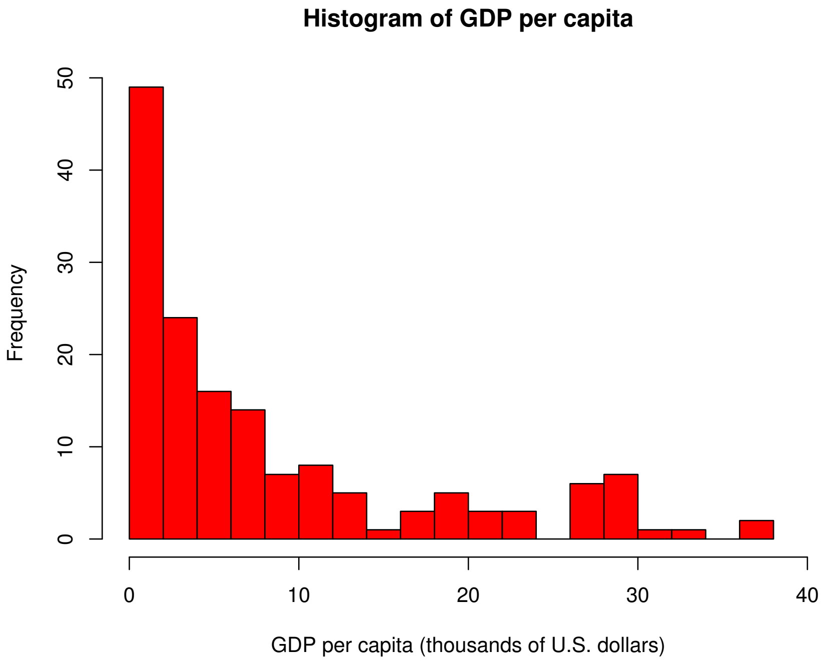


**Figure 1.3:** Histogram of GDP per capita.

### 1.5.3 Skewness of distributions

**Skewness** and **symmetry** are terms used to describe the general shape of a sample distribution.

From Figure 1.3, it is clear that a small number of countries has much larger values of GDP per capita than the majority of countries in the sample. The distribution of GDP per capita has a 'long right tail'. Such a distribution is called **positively skewed** (or skewed to the right).

A distribution with a longer left tail (i.e. toward small values) is **negatively skewed** (or skewed to the left). A distribution is **symmetric** if it is not skewed in either direction.

> **Example 1.5** Figure 1.4 shows a (more-or-less) symmetric sample distribution for diastolic blood pressure.

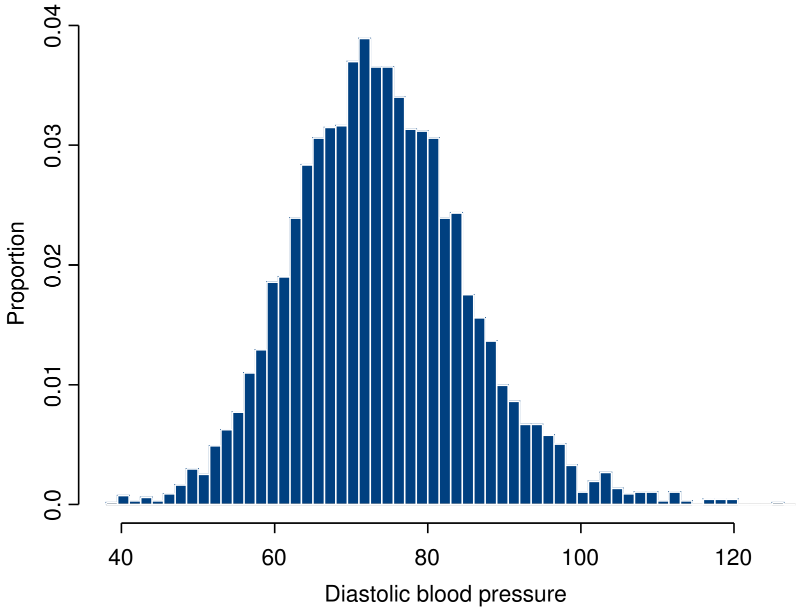


**Figure 1.4:** Diastolic blood pressures of 4,489 respondents aged 25 or over, Health Survey for England, 2002.

> **Example 1.6** Figure 1.5 shows a (slightly) negatively-skewed distribution of marks in an examination. Note the data relate to all candidates sitting the examination. Therefore, the histogram shows the *population* distribution, not a *sample* distribution.

## 1.6 Measures of central tendency

Frequency tables, bar charts and histograms aim to summarise the *whole* sample distribution of a variable. Next we consider descriptive statistics, which summarise *one* feature of the sample distribution in a single number: **summary statistics**.

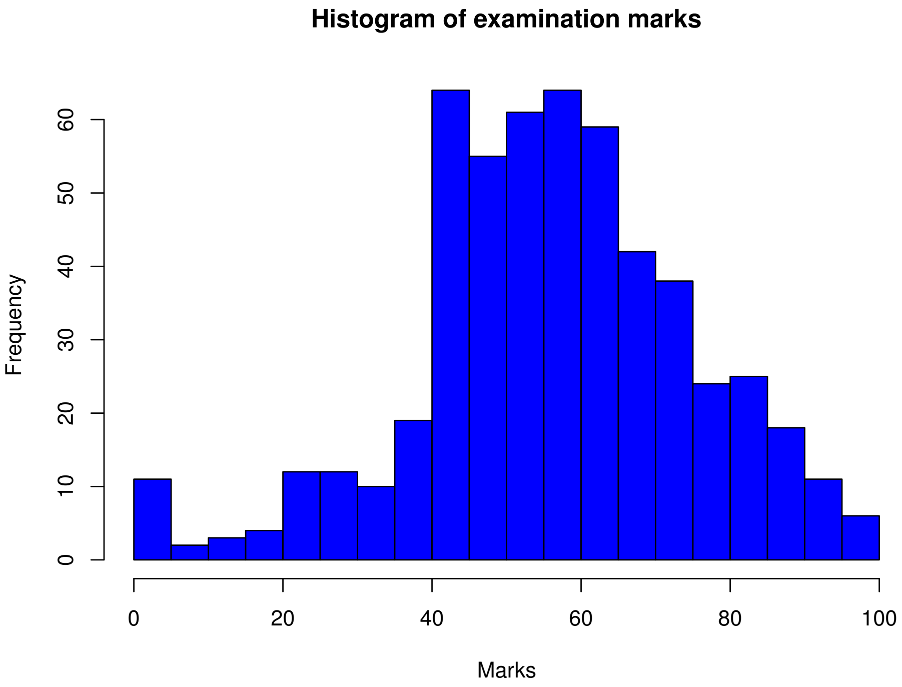


**Figure 1.5:** Final examination marks of a first-year statistics course.

We begin with **measures of central tendency**. These answer the question: where is the 'centre' or 'average' of the distribution?

We consider the following measures of central tendency:

- mean (i.e. the average, sample mean or arithmetic mean)
- median
- mode.

### 1.6.1 Notation for variables

In formulae, a generic variable is denoted by a single letter. In these course notes, usually $X$. However, any other letter ($Y$, $W$ etc.) can also be used, as long as it is used *consistently*. A letter with a subscript denotes a single observation of a variable.

> **Example 1.7** We use $X_i$ to denote the value of $X$ for unit $i$, where $i$ can take values $1, 2, \ldots, n$, and $n$ is the sample size.
>
> Therefore, the $n$ observations of $X$ in the dataset (the **sample**) are $X_1, X_2, \ldots, X_n$. These can also be written as $X_i$, for $i = 1, 2, \ldots, n$.

### 1.6.2 Summation notation

Let $X_1, X_2, \ldots, X_n$ (i.e. $X_i$, for $i = 1, 2, \ldots, n$) be a set of $n$ numbers. The sum of the numbers is written as:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n.$$

This may be written as $\sum_i X_i$, or just $\sum X_i$. Other versions of the same idea are:

- infinite sums: $\sum_{i=1}^{\infty} X_i = X_1 + X_2 + \cdots$
- sums of sets of observations other than 1 to $n$, for example:

$$\sum_{i=2}^{n/2} X_i = X_2 + X_3 + \cdots + X_{n/2}.$$

### 1.6.3 The sample mean

The **sample mean** ('arithmetic mean', 'mean' or 'average') is the most common measure of central tendency. The sample mean of a variable $X$ is denoted $\bar{X}$. It is the 'sum of the observations' divided by the 'number of observations' (sample size) expressed as:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

**Example 1.8** The mean $\bar{X} = \sum_i X_i/n$ of the numbers 1, 4 and 7 is:

$$\frac{1+4+7}{3} = \frac{12}{3} = 4.$$

**Example 1.9** For the variables in Example 1.1:

- the level of democracy has $\bar{X} = 5.3$
- GDP per capita has $\bar{X} = 8.6$ (in \$000s)
- for region the *mean* is not *mean*ingful(!), because the values of the variable do not have a *mean*ingful ordering.

The frequency table of the level of democracy is:

| Level of democracy | Frequency | | Cumulative |
|---|---|---|---|
| $X_j$ | $f_j$ | % | % |
| 0 | 35 | 22.6 | 22.6 |
| 1 | 12 | 7.7 | 30.3 |
| 2 | 4 | 2.6 | 32.9 |
| 3 | 6 | 3.9 | 36.8 |
| 4 | 5 | 3.2 | 40.0 |
| 5 | 5 | 3.2 | 43.2 |
| 6 | 12 | 7.7 | 50.9 |
| 7 | 13 | 8.4 | 59.3 |
| 8 | 16 | 10.3 | 69.6 |
| 9 | 15 | 9.7 | 79.3 |
| 10 | 32 | 20.6 | 100 |
| Total | 155 | 100 | |

If a variable has a small number of distinct values, $\bar{X}$ is easy to calculate from the frequency table. For example, the level of democracy has just 11 different values which occur in the sample $35, 12, \ldots, 32$ times each, respectively.

Suppose $X$ has $K$ different values $X_1, X_2, \ldots, X_K$, with corresponding frequencies $f_1, f_2, \ldots, f_K$. Therefore, $\sum\limits_{j=1}^{K} f_j = n$ and:

$$\bar{X} = \frac{\sum\limits_{j=1}^{K} f_j X_j}{\sum\limits_{j=1}^{K} f_j} = \frac{f_1X_1 + f_2X_2 + \cdots + f_KX_K}{f_1 + f_2 + \cdots + f_K} = \frac{f_1X_1 + f_2X_2 + \cdots + f_KX_K}{n}.$$

In our example, the mean of the level of democracy (where $K = 11$) is:

$$\bar{X} = \frac{35 \times 0 + 12 \times 1 + \cdots + 32 \times 10}{35 + 12 + \cdots + 32} = \frac{0 + 12 + \cdots + 320}{155} \approx 5.3.$$

### Why is the mean a good summary of the central tendency?

Consider the following small dataset:

| | | Deviations: | | | |
|---|---|---|---|---|---|
| | | from $\bar{X}$ $(= 4)$ | | from the median $(= 3)$ | |
| $i$ | $X_i$ | $X_i - \bar{X}$ | $(X_i - \bar{X})^2$ | $X_i - 3$ | $(X_i - 3)^2$ |
| 1 | 1 | $-3$ | 9 | $-2$ | 4 |
| 2 | 2 | $-2$ | 4 | $-1$ | 1 |
| 3 | 3 | $-1$ | 1 | 0 | 0 |
| 4 | 5 | $+1$ | 1 | $+2$ | 4 |
| 5 | 9 | $+5$ | 25 | $+6$ | 36 |
| Sum | 20<br>$\bar{X} = 4$ | 0 | 40 | $+5$ | 45 |

We see that the **sum of deviations** from the mean is 0, i.e. we have:

$$\sum_{i=1}^{n} (X_i - \bar{X}) = 0.$$

The mean is 'in the middle' of the observations $X_1, X_2, \ldots, X_n$, in the sense that positive and negative values of the deviations $X_i - \bar{X}$ cancel out, when summed over all the observations.

Also, the smallest possible value of the sum of *squared* deviations $\sum\limits_{i=1}^{n} (X_i - C)^2$ for any constant $C$ is obtained when $C = \bar{X}$.

### 1.6.4 The (sample) median

Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the sample values of $X$ when *ordered* from the smallest to the largest, known as the **order statistics**, such that:

- $X_{(1)}$ is the smallest observed value (the **minimum**) of $X$
- $X_{(n)}$ is the largest observed value (the **maximum**) of $X$.

**Median**

The (sample) **median**, $q_{50}$, of a variable $X$ is the value which is 'in the middle' of the ordered sample.

If $n$ is odd, then $\boldsymbol{q_{50} = X_{((n+1)/2)}}$.

- For example, if $n = 3$, $q_{50} = X_{(2)}$: (1) **(2)** (3).

If $n$ is even, $\boldsymbol{q_{50} = (X_{(n/2)} + X_{(n/2+1)})/2}$.

- For example, if $n = 4$, $q_{50} = (X_{(2)} + X_{(3)})/2$: (1) **(2)** **(3)** (4).

**Example 1.10** Continuing with Example 1.1, $n = 155$, so $q_{50} = X_{(78)}$. For the level of democracy, the median is 6.

From a table of frequencies, the median is the value for which the cumulative percentage first reaches 50% (or, if a cumulative % is *exactly* 50%, the average of the corresponding value of $X$ and the next highest value).

The ordered values of the level of democracy are:

| | (.0) | (.1) | (.2) | (.3) | (.4) | (.5) | (.6) | (.7) | (.8) | (.9) |
|---|---|---|---|---|---|---|---|---|---|---|
| (0.) | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (1.) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (2.) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (3.) | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| (4.) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 |
| (5.) | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 4 |
| (6.) | 4 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 6 | 6 |
| (7.) | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| (8.) | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 |
| (9.) | 7 | 7 | 7 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| (10.) | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 9 |
| (11.) | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |
| (12.) | 9 | 9 | 9 | 9 | 10 | 10 | 10 | 10 | 10 | 10 |
| (13.) | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| (14.) | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| (15.) | 10 | 10 | 10 | 10 | 10 | 10 | | | | |

The median can be determined from the frequency table of the level of democracy:

| Level of democracy<br>$X_j$ | Frequency<br>$f_j$ | % | Cumulative<br>% |
|---|---|---|---|
| 0 | 35 | 22.6 | 22.6 |
| 1 | 12 | 7.7 | 30.3 |
| 2 | 4 | 2.6 | 32.9 |
| 3 | 6 | 3.9 | 36.8 |
| 4 | 5 | 3.2 | 40.0 |
| 5 | 5 | 3.2 | 43.2 |
| **6** | 12 | 7.7 | **50.9** |
| 7 | 13 | 8.4 | 59.3 |
| 8 | 16 | 10.3 | 69.6 |
| 9 | 15 | 9.7 | 79.3 |
| 10 | 32 | 20.6 | 100 |
| Total | 155 | 100 | |

## 1.6.5 Sensitivity to outliers

For the following small ordered dataset, the mean and median are both 4:

$$1, \quad 2, \quad 4, \quad 5, \quad 8.$$

Suppose we add one observation to get the ordered sample:

$$1, \quad 2, \quad 4, \quad 5, \quad 8, \quad 100.$$

The median is now 4.5, and the mean is 20. In general, the mean is affected much more than the median by **outliers**, i.e. unusually small or large observations. Therefore, you should identify outliers early on and investigate them – perhaps there has been a data entry error, which can simply be corrected. If deemed genuine outliers, a decision has to be made about whether or not to remove them.

## 1.6.6 Skewness, means and medians

Due to its sensitivity to outliers, the mean, more than the median, is pulled toward the longer tail of the sample distribution.

- For a positively-skewed distribution, the mean is larger than the median.
- For a negatively-skewed distribution, the mean is smaller than the median.
- For an exactly symmetric distribution, the mean and median are equal.

When summarising variables with skewed distributions, it is useful to report both the mean and the median.

> **Example 1.11** For the datasets considered previously:
>
> | | Mean | Median |
> |---|---|---|
> | Level of democracy | 5.3 | 6 |
> | GDP per capita | 8.6 | 4.7 |
> | Diastolic blood pressure | 74.2 | 73.5 |
> | Examination marks | 56.6 | 57.0 |

### 1.6.7 Mode

The (sample) **mode** of a variable is the value which has the highest frequency (i.e. appears most often) in the data.

> **Example 1.12** For Example 1.1, the modal region is 1 (Africa) and the mode of the level of democracy is 0.

The mode is not very useful for continuous variables which have many different values, such as GDP per capita in Example 1.1. A variable can have several modes (i.e. be multimodal). For example, GDP per capita has modes 0.8 and 1.9, both with 5 countries out of the 155.

The mode is the only measure of central tendency which can be used even when the values of a variable have no ordering, such as for the (nominal) region variable in Example 1.1.

## 1.7 Measures of dispersion

Central tendency is not the whole story. The two sample distributions in Figure 1.6 have the same mean, but they are clearly not the same. In one (red) the values have more **dispersion** (variation) than in the other.

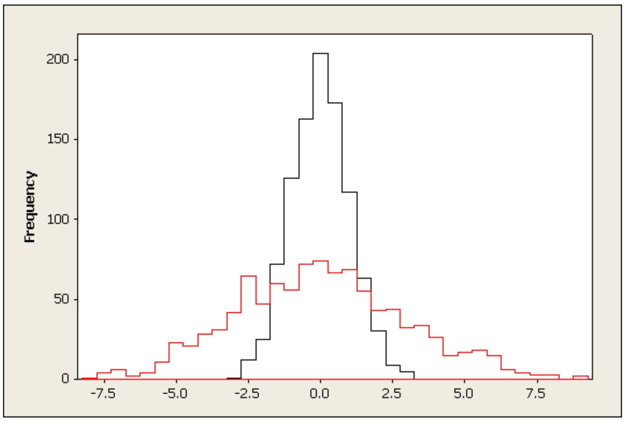


**Figure 1.6:** Two sample distributions.

**Example 1.13** A small example determining the sum of the squared deviations from the (sample) mean, used to calculate common measures of dispersion.

| | | | Deviations from $\bar{X}$ | |
|---|---|---|---|---|
| $i$ | $X_i$ | $X_i^2$ | $X_i - \bar{X}$ | $(X_i - \bar{X})^2$ |
| 1 | 1 | 1 | $-3$ | 9 |
| 2 | 2 | 4 | $-2$ | 4 |
| 3 | 3 | 9 | $-1$ | 1 |
| 4 | 5 | 25 | $+1$ | 1 |
| 5 | 9 | 81 | $+5$ | 25 |
| Sum | 20<br>$\bar{X} = 4$ | 120<br>$= \sum X_i^2$ | 0 | 40<br>$= \sum (X_i - \bar{X})^2$ |

## 1.7.1 Variance and standard deviation

The first measures of dispersion, the sample variance and its square root, the sample standard deviation, are based on $(X_i - \bar{X})^2$, i.e. the squared deviations from the mean.

**Sample variance and standard deviation**

The **sample variance** of a variable $X$, denoted $S^2$ (or $S_X^2$), is defined as:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

The **sample standard deviation** of $X$, denoted $S$ (or $S_X$), is the positive square root of the sample variance:

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}.$$

These are the most commonly-used measures of dispersion. The standard deviation is more understandable than the variance, because the standard deviation is expressed in the same units as $X$ (rather than the variance, which is expressed in squared units).

A useful rule-of-thumb for interpretation is that for many symmetric distributions, such as the 'normal' distribution:

- about 2/3 of the observations are between $\bar{X} - S$ and $\bar{X} + S$, that is, within one (sample) standard deviation about the (sample) mean
- about 95% of the observations are between $\bar{X} - 2 \times S$ and $\bar{X} + 2 \times S$, that is, within two (sample) standard deviations about the (sample) mean.

Remember that standard deviations (and variances) are *never* negative, and they are zero *only* if all the $X_i$ observations are the same (that is, there is no variation in the data).

If we are using a frequency table, we can also calculate:

$$\boldsymbol{S^2 = \frac{1}{n-1}\left(\sum_{j=1}^{K} f_j X_j^2 - n\bar{X}^2\right).}$$

**Example 1.14** Consider the following simple dataset:

| | | | Deviations from $\bar{X}$ | |
|---|---|---|---|---|
| $i$ | $X_i$ | $X_i^2$ | $X_i - \bar{X}$ | $(X_i - \bar{X})^2$ |
| 1 | 1 | 1 | $-3$ | 9 |
| 2 | 2 | 4 | $-2$ | 4 |
| 3 | 3 | 9 | $-1$ | 1 |
| 4 | 5 | 25 | $+1$ | 1 |
| 5 | 9 | 81 | $+5$ | 25 |
| Sum | 20<br>$\bar{X} = 4$ | 120<br>$= \sum X_i^2$ | 0 | 40<br>$= \sum(X_i - \bar{X})^2$ |

We have:

$$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2 = \frac{40}{4} = 10 = \frac{1}{n-1}\left(\sum X_i^2 - n\bar{X}^2\right) = \frac{120 - 5 \times 4^2}{4}$$

and $S = \sqrt{S^2} = \sqrt{10} = 3.16$.

## 1.7.2 Sample quantiles

The median, $q_{50}$, is basically the value which divides the sample into the smallest 50% of observations and the largest 50%. If we consider other percentage splits, we get other (sample) **quantiles** (percentiles), $q_c$.

**Example 1.15** Some special quantiles are given below.

- The **first quartile**, $q_{25}$ or $Q_1$, is the value which divides the sample into the smallest 25% of observations and the largest 75%.
- The **third quartile**, $q_{75}$ or $Q_3$, gives the 75%–25% split.
- The extremes in this spirit are the **minimum**, $X_{(1)}$ (the '0% quantile', so to speak), and the **maximum**, $X_{(n)}$ (the '100% quantile').

These are no longer 'in the middle' of the sample, but they are more general measures of **location** of the sample distribution.

### 1.7.3 Quantile-based measures of dispersion

**Range and interquartile range**

Two measures based on quantile-type statistics are the:

- **range**: $\boldsymbol{X_{(n)} - X_{(1)}}$ **= maximum − minimum**
- **interquartile range (IQR)**: **IQR** $\boldsymbol{= q_{75} - q_{25} = Q_3 - Q_1}$.

The range is, clearly, extremely sensitive to outliers, since it depends on nothing but the extremes of the distribution, i.e. the minimum and maximum observations. The IQR focuses on the middle 50% of the distribution, so it is completely insensitive to outliers.

### 1.7.4 Boxplots

A **boxplot** (in full, a box-and-whiskers plot) summarises some key features of a sample distribution using quantiles. The plot is comprised of the following.

- The line inside the box, which is the median.
- The box, whose edges are the first and third quartiles ($Q_1$ and $Q_3$). Hence the box captures the middle 50% of the data. Therefore, the length of the box is the interquartile range.
- The bottom whisker extends either to the minimum or up to a length of 1.5 times the interquartile range below the first quartile, whichever is closer to the first quartile.
- The top whisker extends either to the maximum or up to a length of 1.5 times the interquartile range above the third quartile, whichever is closer to the third quartile.
- Points beyond 1.5 times the interquartile range below the first quartile or above the third quartile are regarded as outliers, and plotted as individual points.

A much longer whisker (and/or outliers) in one direction relative to the other indicates a skewed distribution, as does a median line not in the middle of the box.

**Example 1.16** Figure 1.7 displays a boxplot of GDP per capita using the sample of 155 countries introduced in Example 1.1. Some summary statistics for this variable are reported below.

| | Mean | Median | Standard deviation | IQR | Range |
|---|---|---|---|---|---|
| GDP per capita | 8.6 | 4.7 | 9.5 | 9.7 | 37.3 |

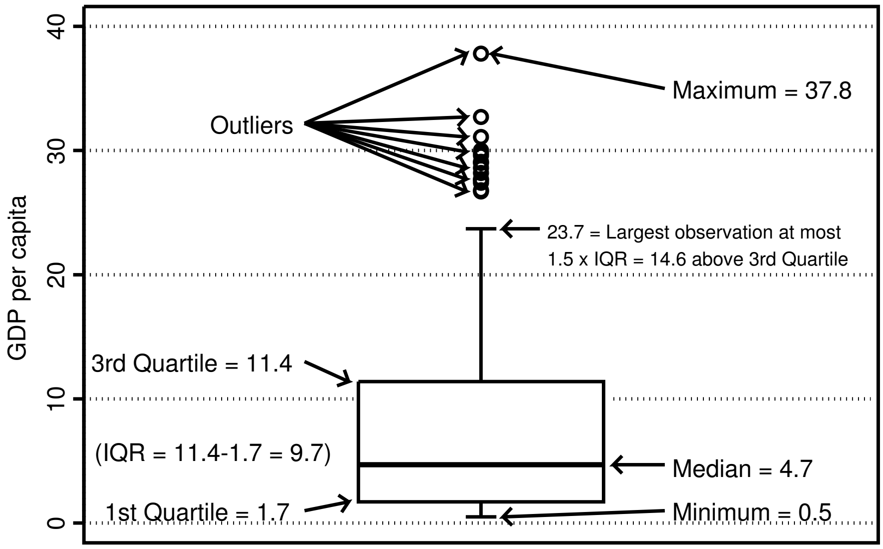


**Figure 1.7:** Boxplot of GDP per capita.

## 1.8 Associations between two variables

So far, we have tried to summarise (some aspect of) the sample distribution of *one* variable at a time.

However, we can also look at two (or more) variables together. The key question is then whether some values of one variable tend to occur frequently together with particular values of another, for example high values with high values. This would be an example of an **association** between the variables. Such associations are central to most interesting research questions, so you will hear much more about them in the future.

Some common methods of descriptive statistics for two-variable associations are introduced here, but only very briefly now and mainly through examples.

The best way to summarise two variables together depends on whether the variables have 'few' or 'many' possible values. We illustrate one method for each combination, as listed below.

- 'Many' versus 'many': scatterplots (including line plots).
- 'Few' versus 'many': side-by-side boxplots.
- 'Few' versus 'few': two-way contingency tables (cross-tabulations).

### 1.8.1 Scatterplots

A **scatterplot** shows the values of two *continuous* variables against each other, plotted as points in a two-dimensional coordinate system.

**Example 1.17** A plot of data for 164 countries is shown in Figure 1.8 which plots the following variables.

- On the horizontal axis (the $x$-axis): a World Bank measure of 'control of corruption', where *high* values indicate *low* levels of corruption.
- On the vertical axis (the $y$-axis): GDP per capita in \$.

Interpretation: it appears that virtually all countries with high levels of corruption have relatively low GDP per capita. At lower levels of corruption there is a positive association, where countries with very low levels of corruption also tend to have high GDP per capita.

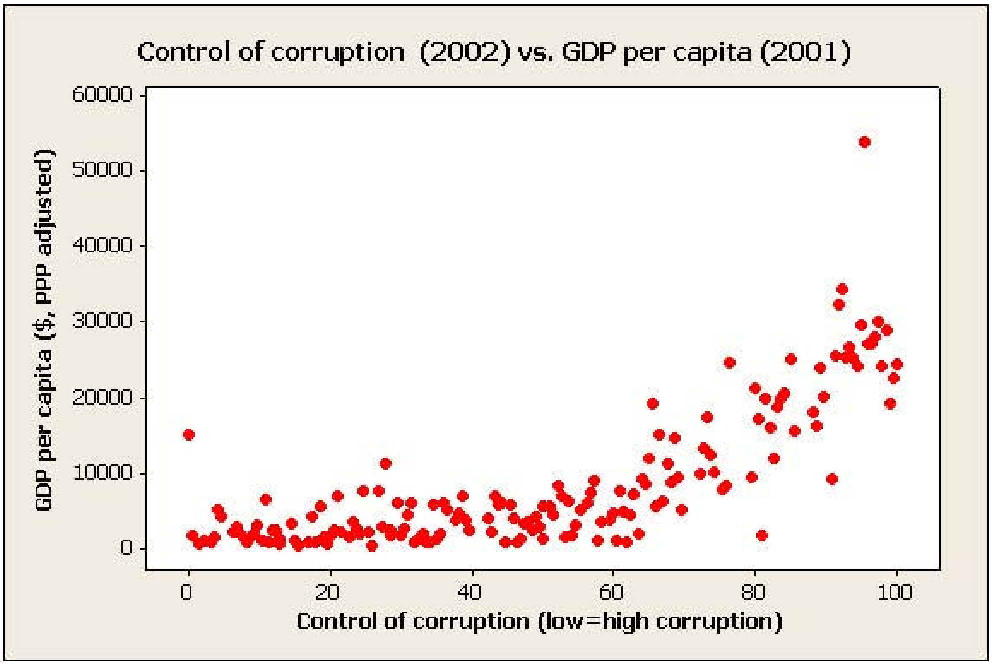


**Figure 1.8:** GDP per capita plotted against control of corruption.

### 1.8.2 Line plots (time series plots)

A common special case of a scatterplot is a **line plot** (time series plot), where the variable on the $x$-axis is time. The points are connected in time order by lines, to show how the variable on the $y$-axis changes over time.

**Example 1.18** Figure 1.9 is a time series of an index of prices of consumer goods and services in the UK for the period 1800–2009 (Office for National Statistics; scaled so that the price level in 1974 = 100). This shows the price inflation over this period.

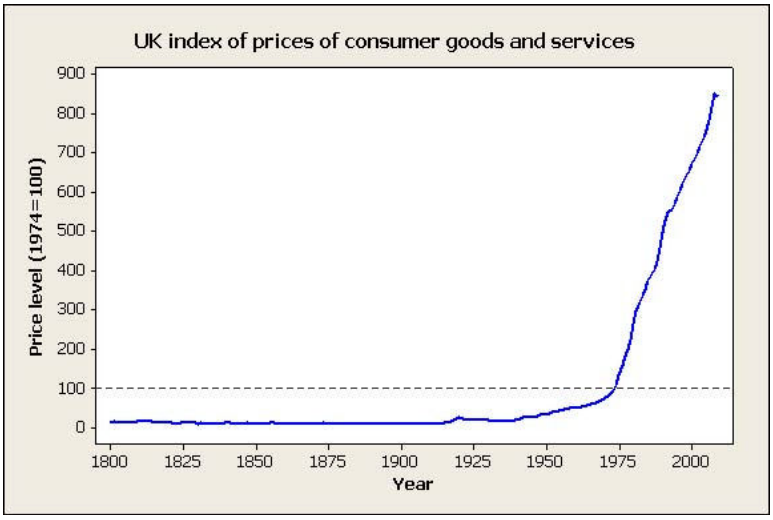


**Figure 1.9:** UK index of prices of consumer goods and services.

### 1.8.3 Side-by-side boxplots for comparisons

Boxplots are useful for *comparisons* of how the distribution of a continuous variable varies across different groups, i.e. across different levels of a discrete variable.

> **Example 1.19** Figure 1.10 shows side-by-side boxplots of GDP per capita for the different regions in Example 1.1.
>
> - GDP per capita in African countries tends to be very low. There is a handful of countries with somewhat higher GDPs per capita (shown as outliers in the plot).
> - The median for Asia is not much higher than for Africa. However, the distribution in Asia is very much skewed to the right, with a tail of countries with very high GDPs per capita.
> - The median in Europe is high, and the distribution is fairly symmetric.
> - The boxplots for Northern America and Oceania are not very useful, because they are based on very few countries (two and three countries, respectively).

### 1.8.4 Two-way contingency tables

A (two-way) **contingency table** (or **cross-tabulation**) shows the frequencies in the sample of each possible *combination* of the values of two discrete variables. Such tables often show the percentages within each *row* or *column* of the table.

> **Example 1.20** The table below reports the results from a survey of 972 private investors.[3] The variables are as follows.
>
> - Row variable: age as a discrete, grouped variable (four categories).

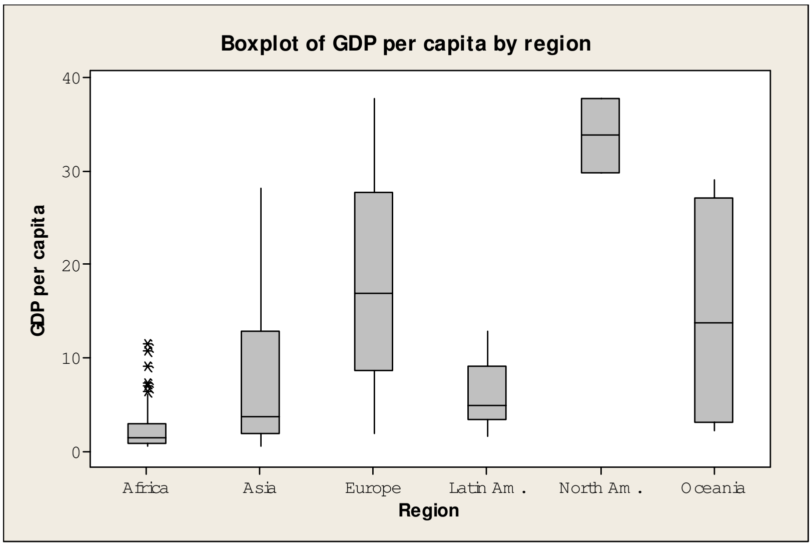


**Figure 1.10:** Side-by-side boxplots of GDP per capita by region.

- Column variable: how much importance the respondent places on short-term gains from his/her investments (four levels).

Interpretation: look at the row percentages. For example, 17.8% of those aged under 45, but only 5.2% of those aged 65 and over, think that short-term gains are 'very important'. Among the respondents, the older age groups seem to be less concerned with quick profits than the younger age groups.

| | Importance of short-term gains | | | | |
|---|---|---|---|---|---|
| | | Slightly | | Very | |
| Age group | Irrelevant | important | Important | important | Total |
| Under 45 | 37<br>(25.3) | 45<br>(30.8) | 38<br>(26.0) | 26<br>(17.8) | 146<br>(100) |
| 45–54 | 111<br>(39.4) | 77<br>(27.3) | 57<br>(20.2) | 37<br>(13.1) | 282<br>(100) |
| 55–64 | 153<br>(60.5) | 49<br>(19.4) | 31<br>(12.3) | 20<br>(7.9) | 253<br>(100) |
| 65 and over | 193<br>(66.3) | 64<br>(22.0) | 19<br>(6.5) | 15<br>(5.2) | 291<br>(100) |
| Total | 494<br>(50.8) | 235<br>(24.2) | 145<br>(14.9) | 98<br>(10.1) | 972<br>(100) |

Numbers in parentheses are percentages within the rows. For example, 25.3 = (37/146) × 100.

[3] Lewellen, W.G., R.C. Lease and G.G. Schlarbaum (1977). 'Patterns of investment strategy and behavior among individual investors'. *The Journal of Business*, 50(3), pp. 296–333.

## 1.9 Overview of chapter

This chapter has looked at different ways of presenting data visually. Which type of diagram is most appropriate will be determined by the types of data being analysed. You should be able to interpret any important features which are apparent from the diagram. This chapter has also introduced some quantitative approaches to summarising data, known as descriptive statistics. We have distinguished measures of location, dispersion and skewness. Although descriptive statistics serve as a very basic form of statistical analysis, they nevertheless are extremely useful for capturing the main characteristics of a dataset. Therefore, *any* statistical analysis of data should start with data visualisation and the calculation of descriptive statistics!

## 1.10 Key terms and concepts

- (Arithmetic) mean
- Bar chart
- Boxplot
- Continuous
- Data matrix
- Dichotomous
- Distribution
- Frequency table
- Interquartile range
- Maximum
- Measures of dispersion
- Minimum
- Nominal
- Ordinal
- Proportion
- Quartile
- Relative frequency
- Sample size
- Skewness
- Symmetry
- Variable
- Association
- Binary
- Contingency table
- Count
- Descriptive statistics
- Discrete
- Frequency
- Histogram
- Line plot
- Measures of central tendency
- Median
- Mode
- Order statistics
- Outlier
- Quantile
- Range
- Sample distribution
- Scatterplot
- Standard deviation
- Unit
- Variance

> *The average human has one breast and one testicle.*
> (Des McHale)

# Chapter 2
# Probability theory

## 2.1 Synopsis of chapter

Probability theory is very important for statistics because it provides the rules which 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 Learning outcomes

After completing this chapter, 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.

## 2.3 Introduction

Consider the following hypothetical example. A country will soon hold a referendum about whether it should leave 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 leaving 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 which 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, $\pi$, of 'Yes' voters is $(0.5083, 0.5717)$.'
- 'The null hypothesis that $\pi = 0.50$, against the alternative hypothesis that $\pi > 0.50$, 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** $\pi$.
- 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** $\pi$ 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 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 which will be observed before we actually collect the sample.

**Probability theory** is the branch of mathematics which 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 outcomes, here $\{1, 2, 3, 4, 5, 6\}$.
- **Event**: any *subset* $A$ of the sample space, for example $A = \{4, 5, 6\}$.[1]

**Probability** of an event $A$, $P(A)$, 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.

## 2.4 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.
>
> - $A = \{\text{Amy}, \text{Bob}, \text{Sam}\}$.
> - $B = \{1, 2, 3, 4, 5\}$.
> - $C = \{x \,|\, x \text{ is a prime number}\} = \{2, 3, 5, 7, 11, \ldots\}$.
> - $D = \{x \,|\, x \geq 0\}$ (that is, the set of all non-negative real numbers).

**Membership of sets and the empty set**

$\boldsymbol{x \in A}$ means that object $x$ is an element of set $A$.
$\boldsymbol{x \notin A}$ 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 $(A \cup B)^c = A^c \cap B^c$.

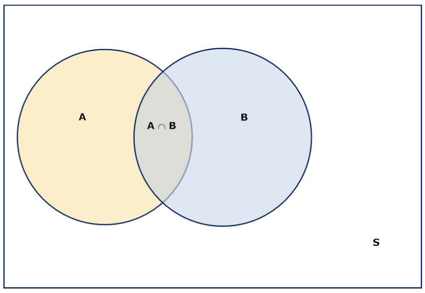


**Figure 2.1:** Venn diagram depicting $A \cup B$ (the total shaded area).

### 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.

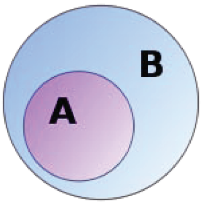


**Figure 2.2:** Venn diagram depicting a subset, where $A \subset B$.

**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.

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.

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

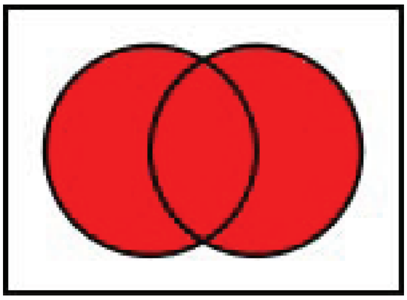


**Figure 2.3:** Venn diagram depicting the union of two sets.

**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:

$$\boldsymbol{A \cap B = \{x \mid x \in A \text{ and } x \in B\}.}$$

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

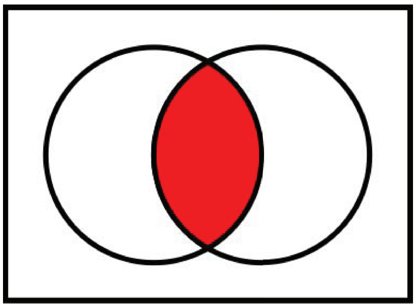


**Figure 2.4:** Venn diagram depicting the intersection of two sets.

**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:

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

and:

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

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

### 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.

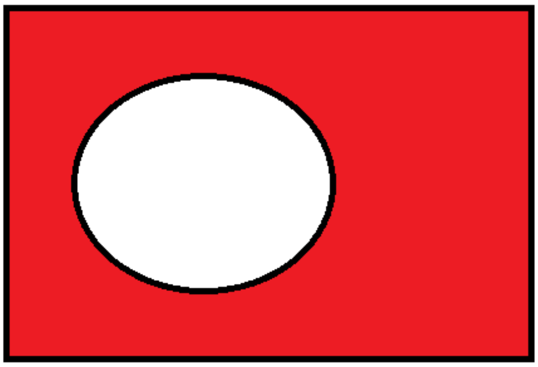


**Figure 2.5:** Venn diagram depicting the complement of a set.

We now consider some useful properties of set operators. In proofs and derivations about sets, you can use the following results without proof.

## Properties of set operators

- **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:**

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

and:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

- **De Morgan's laws:**

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

## Further properties of set operators

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

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

### Mutually exclusive events

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

$$\boldsymbol{A \cap B = \emptyset.}$$

Sets $A_1, A_2, \ldots, A_n$ are **pairwise disjoint** if all pairs of sets from them are disjoint, i.e. $\boldsymbol{A_i \cap A_j = \emptyset}$ 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 $\boldsymbol{\bigcup_{i=1}^{n} A_i = A}$, 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 $\boldsymbol{\bigcup_{i=1}^{\infty} A_i = A}$.

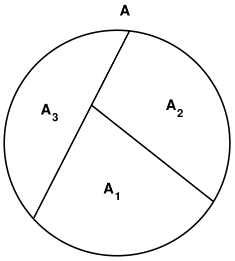
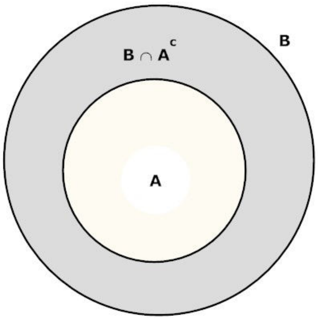


**Figure 2.6:** The partition of the set $A$ into $A_1$, $A_2$ and $A_3$.

**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 $B$.

## 2.5 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 $P(\cdot)$ 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:** $\boldsymbol{P(A) \geq 0}$ for all events $A$.

**Axiom 2:** $\boldsymbol{P(S) = 1.}$

**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:

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

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

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

- 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.

### 2.5.1 Basic properties of probability

**Probability property**

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

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

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

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

However, 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:

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

*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:

$$\boldsymbol{P(A) = P(A_1) + P(A_2) + P(A_3).}$$

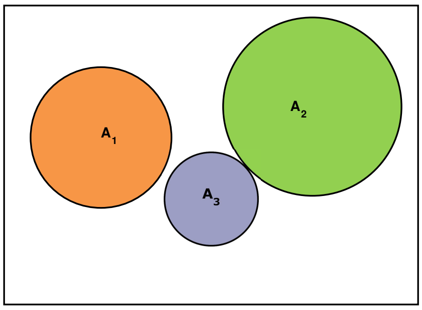


**Figure 2.7:** Venn diagram depicting three mutually exclusive sets, $A_1$, $A_2$ and $A_3$. Note although $A_2$ and $A_3$ have touching boundaries, there is no actual *intersection* and hence they are (pairwise) mutually exclusive.

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:

$$\boldsymbol{P(A^c) = 1 - P(A).}$$

*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:

$$\boldsymbol{P(A) \leq 1.}$$

*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 $\boldsymbol{A \subset B}$, then $\boldsymbol{P(A) \leq P(B)}$**.

*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 $P(B \cap A^c) \geq 0$.

■

**Probability property**

For any two events $A$ and $B$, then:

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

*Proof*: Using partitions:

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

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

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

and hence:

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

■

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. For example, we then have:

- $P(A^c) = 1 - P(A) = 1 - 0.86 = 0.14$, which is the probability that the respondent watches *less than* 1 hour of television
- $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.86 + 0.19 - 0.15 = 0.90$, which is 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 of $P(A)$ 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** Consider the following.
>
> - 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 this population.

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

## 2.6 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 likely*.

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

- 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. Therefore:

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

That is, the probability of $A$ is the *proportion* of outcomes which 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 which 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?

- The sample space is the 36 ordered pairs:

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

- The event of interest is $A = \{(1,4),(2,3),(3,2),(4,1)\}$.
- The probability is $P(A) = 4/36 = 1/9$.

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:

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

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 $P(A \cup B) = P(A) + P(B) - P(A \cap B) = (6 + 3 - 1)/36 = 8/36 = 2/9$.

### How to count the outcomes

In general, it is useful to know about three ways of counting.

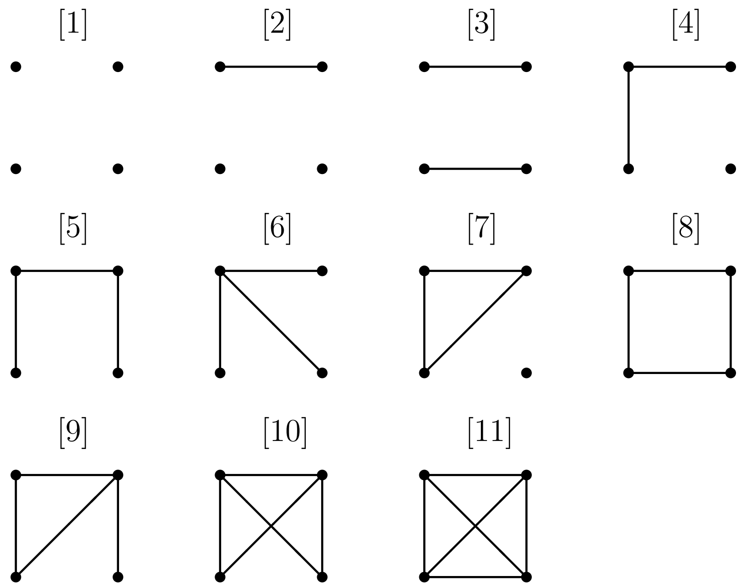


**Figure 2.8:** Friendship patterns in a four-person network.

- Listing and counting all outcomes.
- Combinatorial methods: choosing $k$ objects out of $n$ objects.
- Combining different methods: rules of sum and product.

### 2.6.1 Brute force: listing and counting

In small problems, just listing all possibilities is often quickest.

> **Example 2.15** Consider a group of four people, where each pair of people is either connected (= friends) or not. How many different *patterns* of connections are there (ignoring the identities of who is friends with whom)?
>
> The answer is 11. See the patterns in Figure 2.8.

### 2.6.2 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:

- 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.

Therefore:

- $n$ 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 $k$th object in the sequence.

Therefore, the number of possible ordered sequences of $k$ objects selected with replacement from $n$ objects is:

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

### 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, so that 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 $k$th object.

Therefore, the number of possible ordered sequences of $k$ objects selected without replacement from $n$ objects is:

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

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:

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

  different combinations, for each $k = 0, 1, 2, \ldots, n$.

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.

### Summary of the combinatorial counting rules

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

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

We have not discussed the unordered, with replacement case which is non-examinable. It is provided here only for completeness with an illustration given in Example 2.16.

**Example 2.16** We consider an outline of the proof, using $n = 5$ and $k = 3$ for illustration.

Half-graphically, let `x` denote selected values and `|` the 'walls' between different distinct values. For example:

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

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

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

**Example 2.17** 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 29 February does not exist, so that $n = 365$) in the following cases?

1. It makes a difference who has which birthday (*ordered*), i.e. Amy (1 January), Bob (5 May) and Sam (5 December) is different from Amy (5 May), Bob (5 December) and Sam (1 January), 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 (1 January), Bob (5 May) and Sam (5 December) is treated as the same as Amy (5 May), Bob (5 December) and Sam (1 January), and different people must have

different birthdays (*without replacement*). The number of different sets of birthdays is:

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

**Example 2.18** 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 |

### 2.6.3 Combining counts: rules of sum and product

Even more complex cases can be handled by combining counts.

**Rule of sum**

If an element can be selected in $m_1$ ways from set 1, *or* $m_2$ ways from set 2, … *or* $m_K$ ways from set $K$, the total number of possible selections is:

$$\boldsymbol{m_1 + m_2 + \cdots + m_K.}$$

**Rule of product**

If, in an ordered sequence of $K$ elements, element 1 can be selected in $m_1$ ways, *and* then element 2 in $m_2$ ways, … *and* then element $K$ in $m_K$ ways, the total number of possible sequences is:

$$\boldsymbol{m_1 \times m_2 \times \cdots \times m_K.}$$

**Example 2.19** (The ST102 Moodle site contains a separate note which explains playing cards and hands, and shows how to calculate the probabilities of all the hands. This is for reference only – you do not need to memorise the different types of hands!)

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

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

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

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

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

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

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

Therefore, the probability we want is:

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

How did we get the counts above?

- For full houses, shown next.
- For the others, see the ST102 Moodle site.

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

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

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

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

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

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

The following is a summary of the numbers of all types of 5-card hands, and their probabilities:

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

## 2.7 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 event has happened.

**Independence**

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

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

Independence is sometimes denoted $\boldsymbol{A \perp\!\!\!\perp B}$. 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 2.20** 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'.
>
> Therefore:
>
> - $P(A) = 30/36 = 5/6$
> - $P(B) = 30/36 = 5/6$
> - $P(A \cap B) = 25/36 = 5/6 \times 5/6 = P(A)\,P(B)$, so $A$ and $B$ are independent.

### 2.7.1 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:

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

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

**Example 2.21** 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 only has a hat, Teacher C only has a scarf and Teacher D only has gloves. One teacher out of the four is selected at random. It is shown 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. However, 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.

## 2.7.2 Independent versus mutually exclusive events

The idea of independent events is quite different from that of *mutually exclusive* (disjoint) events, as shown in Figure 2.9.

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 when $P(A) = 0$ 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.

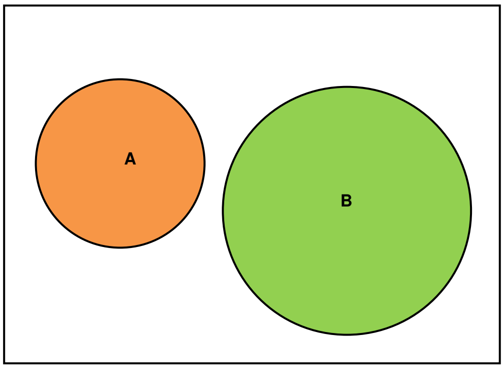


**Figure 2.9:** Venn diagram depicting mutually exclusive events.

### Conditional probability

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

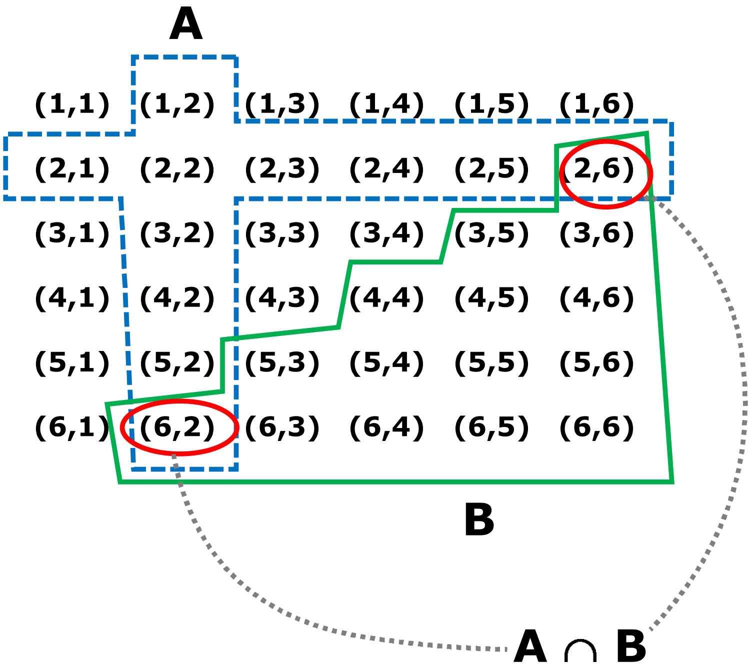
The answer is given by the conditional probability of $A$ given that $B$ has occurred, or the **conditional probability of *A* given *B*** for short, defined as:

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

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

> **Example 2.22** 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'.
>
> These are shown in Figure 2.10. Now $P(A) = 11/36 \approx 0.31$, $P(B) = 15/36$ and $P(A \cap B) = 2/36$. Therefore, the conditional probability of $A$ given $B$ is:
>
> $$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{2/36}{15/36} = \frac{2}{15} \approx 0.13.$$
>
> Learning that $B$ has occurred causes us to *revise* (update) the probability of $A$ downward, from 0.31 to 0.13.

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



| | | | | | |
|---|---|---|---|---|---|
| (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) |

**Figure 2.10:** Events $A$, $B$ and $A \cap B$ for Example 2.22.

> **Example 2.23** In Example 2.22, when we are told that the conditioning event $B$ has occurred, we know we are within the green line in Figure 2.10. 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:
>
> $$P(A \mid B) = \frac{2}{15} = \frac{\text{number of cases of } A \text{ within } B}{\text{number of cases of } B}.$$

### 2.7.3 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:

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

and:

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

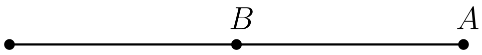
In other words, if $A$ and $B$ are independent, learning that $B$ has occurred does not change the probability of $A$, and learning that $A$ has occurred does not change the probability of $B$. This is exactly what we would expect under independence.

### 2.7.4 Chain rule of conditional probabilities

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

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

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:

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

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:

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

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 2.24.

> **Example 2.24** 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 2.25** 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, $P(A) = 1/270{,}725$.
>
> Let us try with conditional probabilities. Define $A_i$ as 'the $i$th 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
> - $P(A_3 \mid A_1, A_2) = 2/50$

> ■ $P(A_4 \,|\, A_1, A_2, A_3) = 1/49$.
>
> Putting these together with the chain rule gives:
>
> $$\begin{aligned} P(A) &= P(A_1)\, P(A_2 \,|\, A_1)\, P(A_3 \,|\, A_1, A_2)\, P(A_4 \,|\, A_1, A_2, A_3) \\ &= \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49} = \frac{24}{6{,}497{,}400} = \frac{1}{270{,}725}. \end{aligned}$$
>
> 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 2.11.

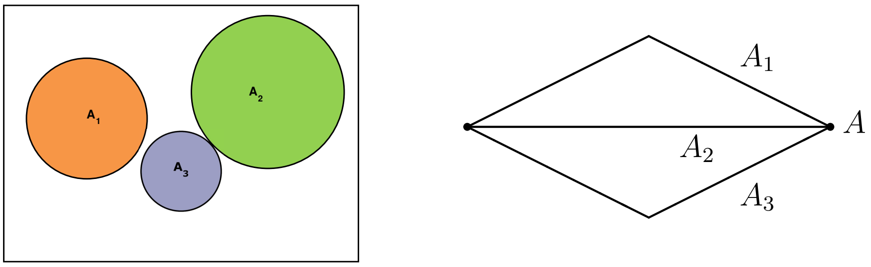


**Figure 2.11:** On the left, a Venn diagram depicting $A = A_1 \cup A_2 \cup A_3$, and on the right the 'paths' to $A$.

Both diagrams in Figure 2.11 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 2.11, where $A_1$, $A_2$ and $A_3$ are symbolised as different 'paths' to $A$.

We now develop powerful methods of calculating sums like:

$$\boldsymbol{P(A) = P(A_1) + P(A_2) + P(A_3).}$$

## 2.7.5 Total probability formula

Suppose $B_1, B_2, \ldots, B_K$ form a partition of the sample space. Therefore, $A \cap B_1$, $A \cap B_2$, ..., $A \cap B_K$ form a partition of $A$, as shown in Figure 2.12.

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:

$$P(A \cap B_i) = P(A \,|\, B_i)\, P(B_i)$$

2. add up the probabilities of the paths:

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

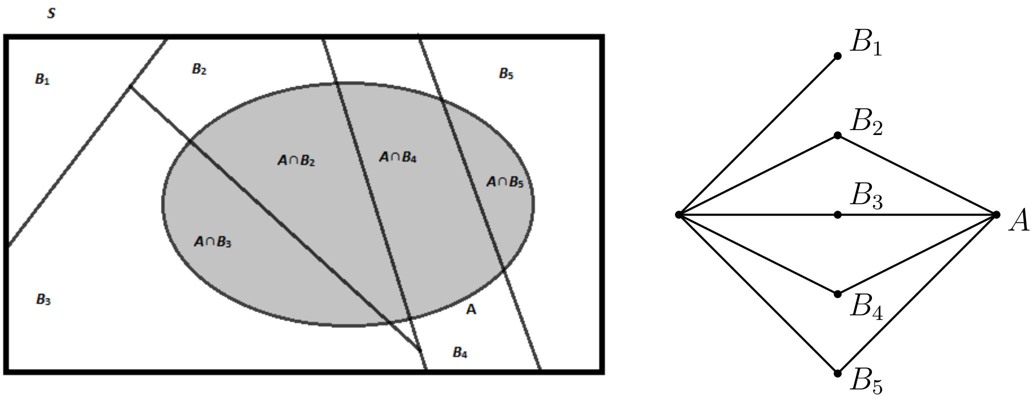
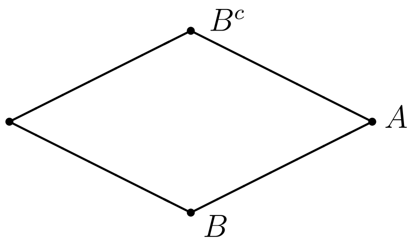


**Figure 2.12:** On the left, a Venn diagram depicting the set $A$ and the partition of $S$, and on the right the 'paths' to $A$.

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 2.26** 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 2.27** 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

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

### 2.7.6 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 occurred, as shown in Figure 2.13.

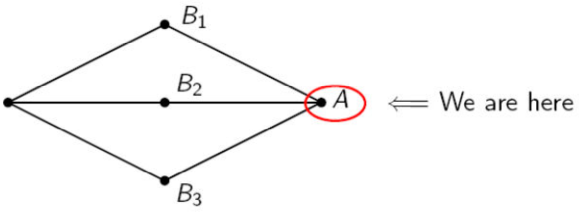


**Figure 2.13:** Paths to $A$ indicating that $A$ has occurred.

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 2.14.

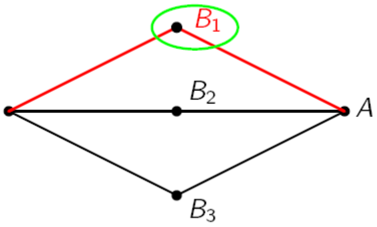


**Figure 2.14:** $A$ being achieved via $B_1$.

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:

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

which holds for each $B_j$, $j = 1, 2, \ldots, K$. This is known as **Bayes' theorem**.

**Example 2.28** Continuing with Example 2.27, 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:

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

Therefore:

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

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 2.29** *You are taking part in a gameshow. The host of the show, who is known as Monty, shows you three outwardly identical boxes. In one of them is a prize, and the other two are empty.*

*You are asked to select, but not open, one of the boxes. After you have done so, Monty, who knows where the prize is, opens one of the two remaining boxes.*

*He always opens a box he knows to be empty, and randomly chooses which box to open when he has more than one option (which happens when your initial choice contains the prize).*

*After opening the empty box, Monty gives you the choice of either switching to the other unopened box or sticking with your original choice. You then receive whatever is in the box you choose.*

*What should you do, assuming you want to win the prize?*

Suppose the three boxes are numbered 1, 2 and 3. Let us define the following events.

- $B_1$, $B_2$, $B_3$: the prize is in Box 1, 2 and 3, respectively.
- $M_1$, $M_2$, $M_3$: Monty opens Box 1, 2 and 3, respectively.

Suppose you choose Box 1 first, and then Monty opens Box 3 (the answer works the same way for all combinations of these). So Boxes 1 and 2 remain unopened.

What we want to know now are the conditional probabilities $P(B_1 \mid M_3)$ and $P(B_2 \mid M_3)$.

You should switch boxes if $P(B_2 \mid M_3) > P(B_1 \mid M_3)$, and stick with your original choice otherwise. (You would be indifferent about switching if it was the case that $P(B_2 \mid M_3) = P(B_1 \mid M_3)$.)

Suppose that you first choose Box 1, and then Monty opens Box 3. Bayes' theorem tells us that:

$$P(B_2 \mid M_3) = \frac{P(M_3 \mid B_2)\, P(B_2)}{P(M_3 \mid B_1)\, P(B_1) + P(M_3 \mid B_2)\, P(B_2) + P(M_3 \mid B_3)\, P(B_3)}.$$

We can assign values to each of these.

- The prize is initially equally likely to be in any of the boxes. Therefore, $P(B_1) = P(B_2) = P(B_3) = 1/3$.
- If the prize is in Box 1 (which you choose), Monty chooses at random between the two remaining boxes, i.e. Boxes 2 and 3. Hence $P(M_3 \mid B_1) = 1/2$.
- If the prize is in one of the two boxes you did *not* choose, Monty cannot open that box, and must open the other one. Hence $P(M_3 \mid B_2) = 1$ and so $P(M_3 \mid B_3) = 0$.

Putting these probabilities into the formula gives:

$$P(B_2 \mid M_3) = \frac{1 \times 1/3}{1/2 \times 1/3 + 1 \times 1/3 + 0 \times 1/3} = \frac{2}{3}$$

and hence $P(B_1 \mid M_3) = 1 - P(B_2 \mid M_3) = 1/3$ (because also $P(M_3 \mid B_3) = 0$ and so $P(B_3 \mid M_3) = 0$).

The same calculation applies to every combination of your first choice and Monty's choice. Therefore, you will *always* double your probability of winning the prize if you switch from your original choice to the box that Monty did not open.

The *Monty Hall problem* has been called a 'cognitive illusion', because something about it seems to mislead most people's intuition. In experiments, around 85% of people tend to get the answer wrong at first.

The most common incorrect response is that the probabilities of the remaining boxes after Monty's choice are both 1/2, so that you should not (or rather need not) switch.

This is typically based on 'no new information' reasoning. Since we know in advance that Monty *will* open one empty box, the fact that he does so appears to tell us nothing new and should not cause us to favour either of the two remaining boxes – hence a probability of 1/2 for each.

It is true that Monty's choice tells you nothing new about the probability of your *original* choice, which remains at 1/3. However, it tells us a lot about the other two boxes. First, it tells us everything about the box he chose, namely that it does not contain the prize. Second, all of the probability of that box gets 'inherited' by the box neither you nor Monty chose, which now has the probability 2/3.

**Example 2.30** You are waiting for your bag at the baggage reclaim 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 which 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 as follows.

- $P(x \mid A) = 1$ for all $x$. If your 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 2.15 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!

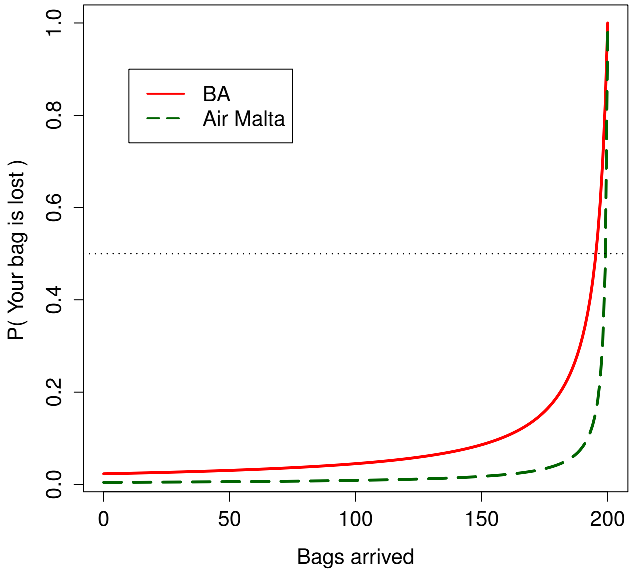


**Figure 2.15:** Plot of $P(A \mid x)$ as a function of $x$ for the two airlines in Example 2.30, Air Malta and British Airways (BA).

## 2.8 Overview of chapter

This chapter introduced some formal terminology related to probability theory. The axioms of probability were introduced, from which various other probability results were 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.

## 2.9 Key terms and concepts

- Axiom
- Binomial coefficient
- Classical probability
- Combination
- Conditional probability
- Disjoint
- Empty set
- Event
- Independence
- Mutually exclusive
- Pairwise disjoint
- Permutation
- Relative frequency
- Set
- Total probability
- Venn diagram
- Bayes' theorem
- Chain rule
- Collectively exhaustive
- Complement
- Counting
- Element
- Experiment
- Factorial
- Intersection
- Outcome
- Partition
- Probability (theory)
- Sample space
- Subset
- Union
- With(out) replacement

*There are lies, damned lies and statistics.*
(Mark Twain)

# Chapter 3
# Random variables

## 3.1 Synopsis of chapter

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 Learning outcomes

After completing this chapter, you should be able to:

- define a random variable and distinguish it from the values which 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.

## 3.3 Introduction

In Chapter 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:

$$\boldsymbol{X_1, X_2, \ldots, X_n}$$

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.

### 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 as follows.

- $\boldsymbol{P(X = x)}$ denotes the probability that (the value of) $X$ is $x$.
- $\boldsymbol{P(X > 0)}$ denotes the probability that $X$ is positive.
- $\boldsymbol{P(a < X < b)}$ 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 Chapter 1.

[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.

| Random variable | Sample |
|---|---|
| Probability distribution | Sample distribution |
| Mean (expected value) | Sample mean (average) |
| Variance | Sample variance |
| Standard deviation | Sample 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.

## 3.4 Discrete random variables

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

1. *The 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.

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

### 3.4.1 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, $X$.[3]

| Number of people<br>in the 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:

$$\boldsymbol{p(x) = P(X = x).}$$

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. $\boldsymbol{p(x) \geq 0}$ for all real numbers $x$.
2. $\boldsymbol{\sum_{x_i \in S} p(x_i) = 1}$, 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 $x$ which is not one of the possible values of $X$.

[3] Source: ONS, National report for the 2001 Census, England and Wales. Table UV51.

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

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

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.

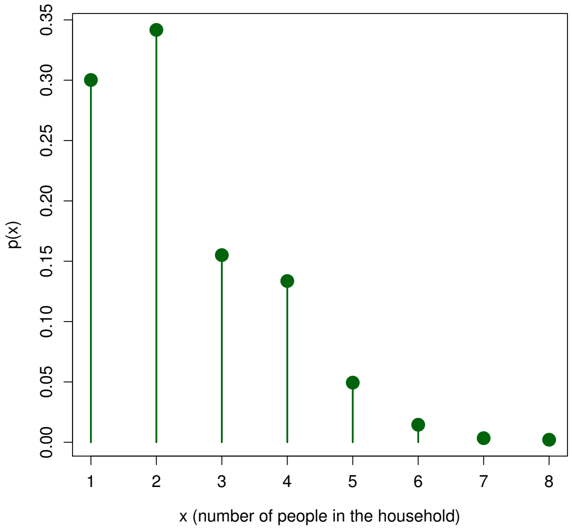


**Figure 3.1:** Probability function for Example 3.3.

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

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

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

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

**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 $\pi$ at each throw and, therefore, the probability of an unsuccessful throw is $1 - \pi$
- outcomes of different throws are independent.

Hence 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:

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

The expression of the pf involves a **parameter** $\pi$ (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 value of $\pi$. For example, Figure 3.2 shows values of $p(x)$ for two values of $\pi$ reflecting fairly good and pretty poor free-throw shooters, respectively.

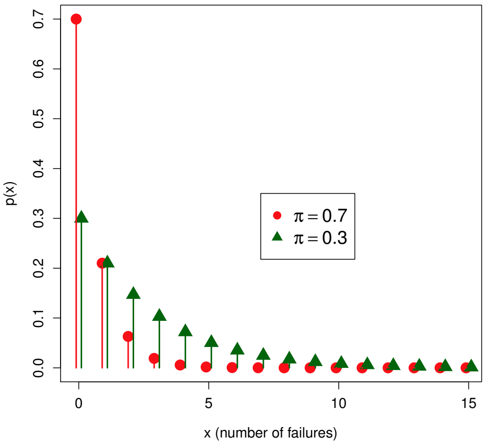


**Figure 3.2:** Probability function for Example 3.4. $\pi = 0.7$ indicates a fairly good free-throw shooter. $\pi = 0.3$ indicates a pretty poor free-throw shooter.

### 3.4.2 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:

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

For a discrete random variable it is given by:

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

i.e. the sum of the probabilities of the possible values of $X$ which 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<br>in the 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.

**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:

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

we can write:

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

The cdf is shown in graphical form in Figure 3.4.

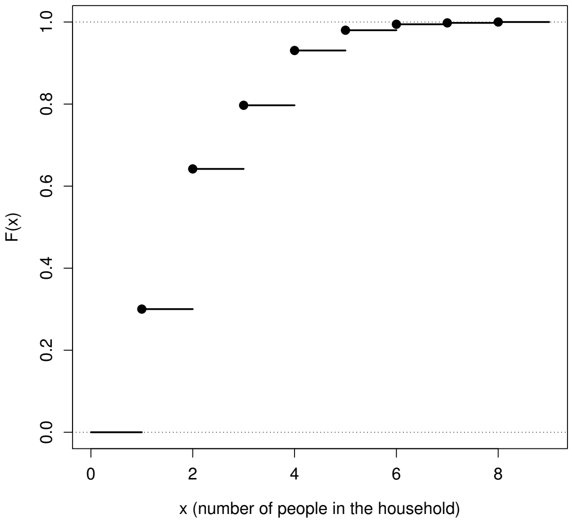


**Figure 3.3:** Cumulative distribution function for Example 3.5.

### 3.4.3 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*).

### 3.4.4 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:
>
> - $P(X = 1) = p(1) = F(1) = 0.3002$

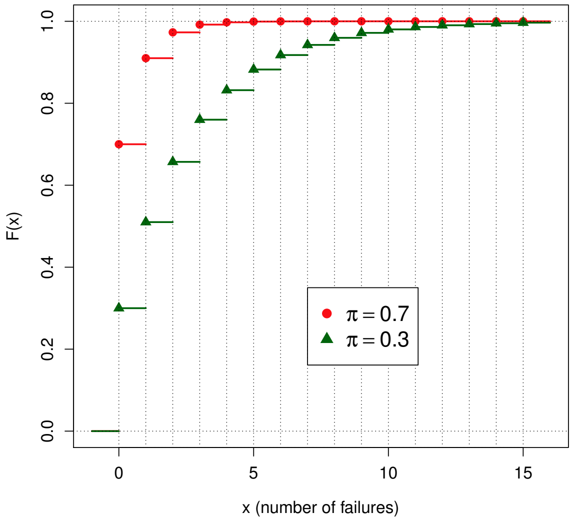


**Figure 3.4:** Cumulative distribution function for Example 3.6.

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

### 3.4.5 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:

$$\mathbf{E}(\boldsymbol{X}) = \sum_{x_i \in S} \boldsymbol{x_i p(x_i)}.$$

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 $\mu$ (the lower-case Greek letter '*mu*'), or $\mu_X$, is often used.

### 3.4.6 Expected value versus sample mean

The mean (expected value) $\text{E}(X)$ 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 $X_1, X_2, \dots, X_K$, and their frequencies in a sample are $f_1, f_2, \dots, f_K$, respectively. Therefore, the sample mean of $X$ is:

$$\bar{X} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_K x_K}{f_1 + f_2 + \cdots + f_K} = x_1 \widehat{p}(x_1) + x_2 \widehat{p}(x_2) + \cdots + x_K \widehat{p}(x_K) = \sum_{i=1}^{K} x_i \widehat{p}(x_i)$$

where:

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

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

The expected value of the random variable $X$ is:

$$\text{E}(X) = x_1 p(x_1) + x_2 p(x_2) + \cdots + x_K p(x_K) = \sum_{i=1}^{K} x_i p(x_i).$$

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

**Example 3.8** Continuing with the household size example:

| Number of people in the 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<br>$= \text{E}(X)$ |

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 $x = 0, 1, 2, \dots$, and 0 otherwise.

The expected value of $X$ is then:

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

from which we can solve:

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

Hence for example:

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

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

**Example 3.10** To illustrate the use of expected values, let us consider the game of roulette, from the point of view of the casino ('The House').

Suppose a player puts a bet of £1 on 'red'. If the ball lands on any of the 18 red numbers, the player gets that £1 back, plus another £1 from The House. If the result is one of the 18 black numbers or the green 0, the player loses the £1 to The House.

We assume that the roulette wheel is unbiased, i.e. that all 37 numbers have equal probabilities. What can we say about the probabilities and expected values of wins and losses?

Define the random variable $X$ = 'money received by The House'. Its possible values are $-1$ (the player wins) and 1 (the player loses). The probability function is:

$$p(x) = \begin{cases} 18/37 & \text{for } x = -1 \\ 19/37 & \text{for } x = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the expected value is:

$$\mathrm{E}(X) = \left(-1 \times \frac{18}{37}\right) + \left(1 \times \frac{19}{37}\right) = +0.027.$$

*On average*, The House expects to win 2.7p for every £1 which players bet on red. This expected gain is known as the *house edge*. It is positive for all possible bets in roulette.

The edge is the expected gain from a single bet. Usually, however, players bet again if they win at first – gambling can be addictive!

Consider a player who starts with £10 and bets £1 on red repeatedly until the player either has lost all of the £10 or doubled their money to £20.

It can be shown that the probability that such a player reaches £20 before they go down to £0 is about 0.368. Define $X$ = 'money received by The House', with the probability function:

$$p(x) = \begin{cases} 0.368 & \text{for } x = -10 \\ 0.632 & \text{for } x = 10 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the expected value is:

$$\mathrm{E}(X) = (-10 \times 0.368) + (10 \times 0.632) = +2.64.$$

*On average*, The House can expect to keep about 26.4% of the money which players like this bring to the table.

## 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:

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

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

**Example 3.11** The expected value of the square of $X$ is:

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

In general:

$$\mathbf{E(g(X)) \neq g(E(X))}$$

when $g(X)$ is a *non-linear* function of $X$.

> **Example 3.12** 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 which are not random variables. Therefore:

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

*Proof*: We have:

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

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:

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

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:

$$\mathbf{Var}(\boldsymbol{X}) = \mathbf{E}((\boldsymbol{X} - \mathbf{E}(\boldsymbol{X}))^2) = \sum_{\boldsymbol{x}} (\boldsymbol{x} - \mathbf{E}(\boldsymbol{X}))^2 \boldsymbol{p}(\boldsymbol{x}).$$

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

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

*Alternative notation*: the variance is often denoted $\sigma^2$ ('sigma squared') and the standard deviation by $\sigma$ ('sigma').

*An alternative formula*: the variance can also be calculated as:

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

This will be proved later.

**Example 3.13** 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<br>$= \mathrm{E}(X)$ | | 1.699<br>$= \mathrm{Var}(X)$ | | 7.259<br>$= \mathrm{E}(X^2)$ |

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

**Example 3.14** 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 $\mathrm{sd}(X) = 0.78$ for $\pi = 0.7$
- $\mathrm{Var}(X) = 0.7/(0.3)^2 = 7.78$ and $\mathrm{sd}(X) = 2.79$ for $\pi = 0.3$.

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

**Variances of linear transformations**

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

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

*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, $\mathrm{sd}(aX+b) = |a|\,\mathrm{sd}(X)$.

■

If $a = 0$, this gives:

$$\mathbf{Var}(\boldsymbol{b}) = \mathbf{0}.$$

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.

### Summary of properties of E($X$) and Var($X$)

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

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

$$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X) \quad \text{and} \quad \mathrm{sd}(aX+b) = |a|\,\mathrm{sd}(X)$$

$$\mathrm{E}(b) = b \quad \text{and} \quad \mathrm{Var}(b) = \mathrm{sd}(b) = 0.$$

We define $\mathrm{Var}(X) = \mathrm{E}((X-\mathrm{E}(X))^2) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$ and $\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$. Also, $\mathrm{Var}(X) \geq 0$ and $\mathrm{sd}(X) \geq 0$ always, and $\mathrm{Var}(X) = \mathrm{sd}(X) = 0$ only if $X$ is a constant.

## 3.4.7 Moments of a random variable

We can also define, for each $k = 1, 2, \ldots$, the following:

- the $k$th moment about zero is $\mu_k = \mathrm{E}(X^k)$
- the $k$th central moment is $\mu'_k = \mathrm{E}((X-\mathrm{E}(X))^k)$.

Clearly, $\mu_1 = \mu = \mathrm{E}(X)$ and $\mu_2' = \mathrm{Var}(X)$.

These will be mentioned again in Chapter 7.

**Example 3.15** 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 $\pi$ is a probability parameter such that $0 \leq \pi \leq 1$.

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, $F(x)$.
- Derive the expected value, $\mathrm{E}(X)$.

Note: the examination may also contain questions like this. The difficulty of such questions depends partly on the form of $p(x)$, and what kinds of manipulations are needed to work with it. So questions of this type may be very easy, or quite hard!

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. $\sum\limits_{x=0}^{n} p(x) = 1$. 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.

For the values $x = 0, 1, 2, \ldots, n$, 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:

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

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 $0, 1, 2, \ldots, n-1$ and probability parameter $\pi$.

The variance of the distribution is $\mathrm{Var}(X) = n\pi(1-\pi)$. This is not derived here, but will be proved in a different way later.

### 3.4.8 The moment generating function

**Moment generating function**

The **moment generating function** (mgf) of a discrete random variable $X$ is defined as:

$$M_X(t) = \mathrm{E}(e^{tX}) = \sum_x e^{tx} p(x).$$

$M_X(t)$ is a function of real numbers $t$. It is not a random variable itself.

The form of the mgf is *not* interesting or informative in itself. Instead, the reason we define the mgf is that it is a convenient tool for deriving means and variances of distributions, using the following results:

$$M_X'(0) = \mathrm{E}(X) \quad \text{and} \quad M_X''(0) = \mathrm{E}(X^2)$$

which also gives:

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

This is useful if the mgf is easier to derive than $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ directly.

Other moments about zero are obtained from the mgf similarly:

$$\boldsymbol{M_X^{(k)}(0) = \mathrm{E}(X^k) \quad \text{for } k = 1, 2, \ldots.}$$

**Example 3.16** In the basketball example, we considered the distribution with $p(x) = (1-\pi)^x \pi$ for $x = 0, 1, 2, \ldots$.

The mgf for this distribution is:

$$\begin{aligned} M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) &= \sum_{x=0}^{\infty} \mathrm{e}^{tx} p(x) \\ &= \sum_{x=0}^{\infty} \mathrm{e}^{tx} (1-\pi)^x \pi \\ &= \pi \sum_{x=0}^{\infty} (\mathrm{e}^t (1-\pi))^x \\ &= \frac{\pi}{1 - \mathrm{e}^t (1-\pi)} \end{aligned}$$

using the sum to infinity of a geometric series, for $t < -\ln(1-\pi)$ to ensure convergence of the sum.

From the mgf $M_X(t) = \pi/(1 - \mathrm{e}^t(1-\pi))$ we obtain:

$$\begin{aligned} M_X'(t) &= \frac{\pi(1-\pi)\mathrm{e}^t}{(1-\mathrm{e}^t(1-\pi))^2} \\ M_X''(t) &= \frac{\pi(1-\pi)\mathrm{e}^t(1-(1-\pi)\mathrm{e}^t)(1+(1-\pi)\mathrm{e}^t)}{(1-\mathrm{e}^t(1-\pi))^4} \end{aligned}$$

and hence (since $\mathrm{e}^0 = 1$):

$$\begin{aligned} M_X'(0) &= \frac{1-\pi}{\pi} = \mathrm{E}(X) \\ M_X''(0) &= \frac{(1-\pi)(2-\pi)}{\pi^2} = \mathrm{E}(X^2) \end{aligned}$$

and:

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

**Example 3.17** Consider a discrete random variable $X$ with possible values $0, 1, 2, \ldots$, a parameter $\lambda > 0$, and the following pf:

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

The mgf for this distribution is:

$$M_X(t) = \sum_{x=0}^{\infty} \mathrm{e}^{tx} \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} = \mathrm{e}^{-\lambda} \sum_{x=0}^{\infty} \frac{(\mathrm{e}^t \lambda)^x}{x!} = \mathrm{e}^{-\lambda} \mathrm{e}^{\lambda \mathrm{e}^t} = \mathrm{e}^{\lambda(\mathrm{e}^t - 1)}.$$

> Note: this uses the series expansion of the exponential function from calculus, i.e. that for any number $a$, we have:
>
> $$\mathrm{e}^a = \sum_{x=0}^{\infty} \frac{a^x}{x!} = 1 + a + \frac{a^2}{2!} + \frac{a^3}{3!} + \cdots .$$
>
> From the mgf $M_X(t) = \mathrm{e}^{\lambda(\mathrm{e}^t - 1)}$ we obtain:
>
> $$M_X'(t) = \lambda \mathrm{e}^t \mathrm{e}^{\lambda(\mathrm{e}^t-1)}$$
>
> $$M_X''(t) = \lambda \mathrm{e}^t (1 + \lambda \mathrm{e}^t) \mathrm{e}^{\lambda(\mathrm{e}^t-1)}$$
>
> and hence:
>
> $$M_X'(0) = \lambda = \mathrm{E}(X)$$
> $$M_X''(0) = \lambda(1+\lambda) = \mathrm{E}(X^2)$$
>
> and:
>
> $$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \lambda(1+\lambda) - \lambda^2 = \lambda.$$

### Other useful properties of moment generating functions

If the mgfs mentioned in these statements exist, then the following apply.

- The mgf uniquely determines a probability distribution. In other words, if for two random variables $X$ and $Y$ we have $M_X(t) = M_Y(t)$ (for points around $t = 0$), then $X$ and $Y$ have the same distribution.
- If $Y = aX + b$ where $X$ is a random variable and $a$ and $b$ are constants, then:

  $$\boldsymbol{M_Y(t) = \mathrm{e}^{bt} M_X(at).}$$

- Suppose that the random variables $X_1, X_2, \ldots, X_n$ are independent (a concept which will be defined in Chapter 5) and if we also define $Y = X_1 + X_2 + \cdots + X_n$, then:

  $$\boldsymbol{M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t)}$$

  and, in particular, if all the $X_i$s have the same distribution (of $X$), then $M_Y(t) = M_X(t)^n$.

## 3.5 Continuous random variables

A random variable (and its probability distribution) is **continuous** if it can have an uncountably infinite number of possible values.[4]

[4] 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 this course) 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.

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

> **Example 3.18** 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.
> - Therefore, the possible values of this random variable are $\{x \mid x \geq 1\}$.

Most of the concepts introduced for discrete random variables have exact or approximate analogies for continuous random variables, and many results are the same for both types. However, 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 $\boldsymbol{f(x)}$ (or $\boldsymbol{f_X(x)}$).

> **Example 3.19** Continuing the insurance example in Example 3.18, we consider a pdf of the following form:
>
> $$f(x) = \begin{cases} \alpha k^\alpha / x^{\alpha+1} & \text{for } x \geq k \\ 0 & \text{otherwise} \end{cases}$$
>
> 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.

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 random variable:

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

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

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

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

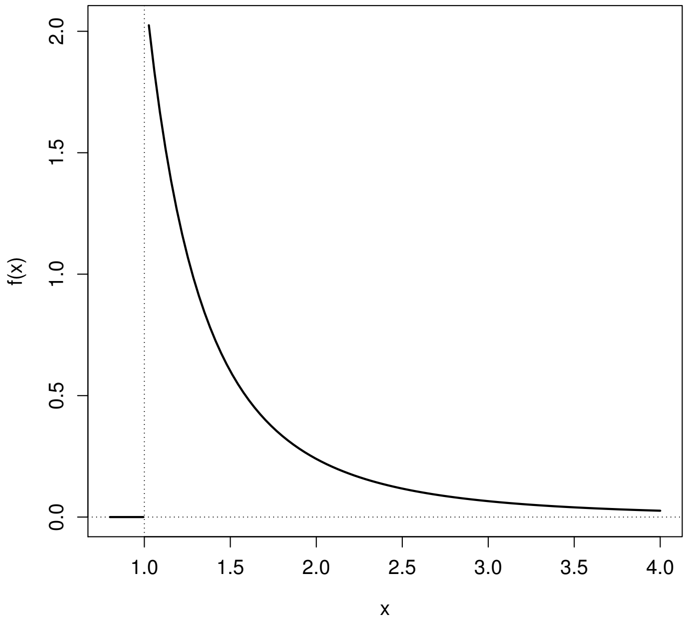


**Figure 3.5:** Probability density function for Example 3.19.

### Probabilities of intervals for continuous random variables

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

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

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

### Properties of pdfs

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

1. We require:

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

2. We require:

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

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

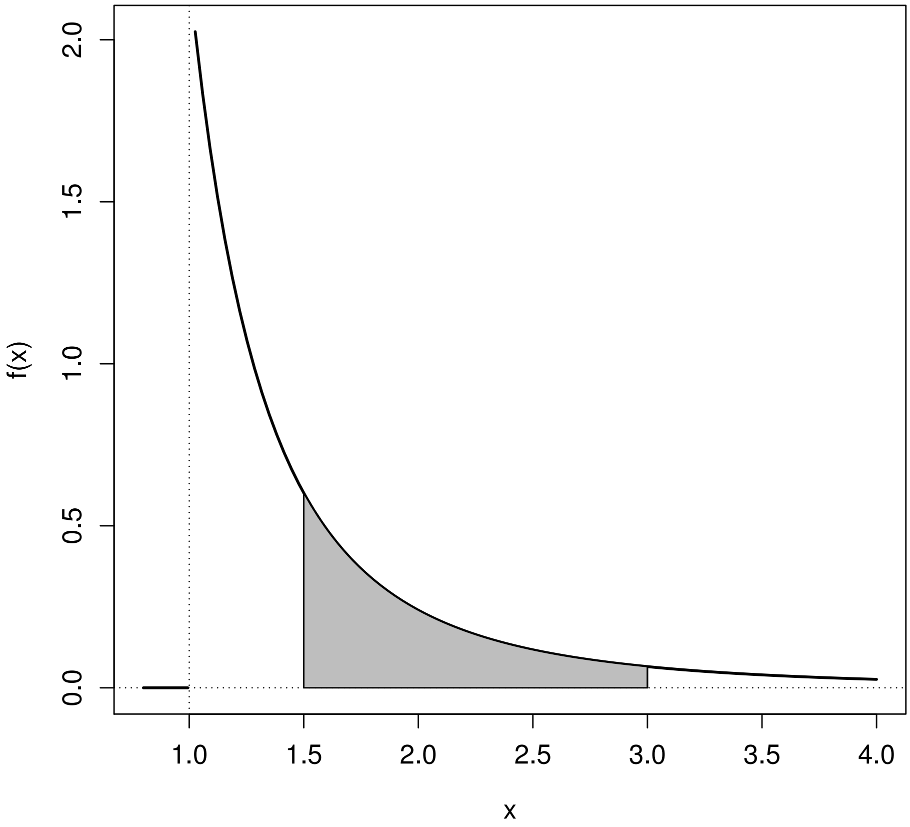


**Figure 3.6:** Probability density function showing $P(1.5 < X \leq 3)$.

**Example 3.21** Continuing with the insurance example, we check that the conditions hold for the pdf:

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

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. We have:

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

### 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:

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

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:

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

The pdf is obtained from the cdf through differentiation:

$$\boldsymbol{f(x) = F'(x).}$$

**Example 3.22** Continuing the insurance example:

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

Therefore:

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

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

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

A plot of the cdf is shown in Figure 3.7.

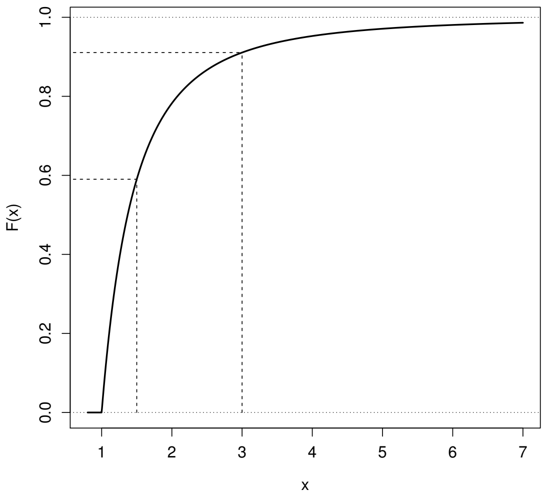


**Figure 3.7:** Cumulative distribution function for Example 3.22.

### 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:

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

**Example 3.23** 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 3.24** Consider now a continuous random variable with the following pdf:

$$f(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

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:

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

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

### Mixed distributions

A random variable can also be a *mixture* of discrete and continuous parts.

For example, consider the sizes of payments which an insurance company needs to make on *all* insurance policies of a particular type. Most policies result in no claims or claims below the deductible, so the payment for them is 0. For those policies which do result in a claim, the size of each claim is some number greater than 0.

Consider a random variable $X$ which is a mixture of two components.

- $P(X = 0) = \pi$ for some $\pi \in (0, 1)$. Here $\pi$ is the probability that a policy results in no payment.
- Among the rest, $X$ follows a continuous distribution with the probabilities distributed as $(1-\pi)f(x)$, where $f(x)$ is a continuous pdf over $x > 0$. In other words, this spreads the remaining probability $(1-\pi)$ over different non-zero values of payments. For example, we could use the Pareto distribution for this *loss distribution* $f(x)$ (or actually as a distribution of $X + k$, since the company only pays the amount above the deductible, $k$).

### 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 $g(X)$, the variance and standard deviation are the same as for discrete distributions, except that summation is replaced by integration:

$$\mathbf{E}(\boldsymbol{X}) = \int_{-\infty}^{\infty} \boldsymbol{x}\,\boldsymbol{f}(\boldsymbol{x})\,\mathbf{d}\boldsymbol{x}$$

$$\mathbf{E}(\boldsymbol{g}(\boldsymbol{X})) = \int_{-\infty}^{\infty} \boldsymbol{g}(\boldsymbol{x})\,\boldsymbol{f}(\boldsymbol{x})\,\mathbf{d}\boldsymbol{x}$$

$$\mathbf{Var}(\boldsymbol{X}) = \mathbf{E}((\boldsymbol{X}-\mathbf{E}(\boldsymbol{X}))^2) = \int_{-\infty}^{\infty} (\boldsymbol{x}-\mathbf{E}(\boldsymbol{X}))^2 \boldsymbol{f}(\boldsymbol{x})\,\mathbf{d}\boldsymbol{x} = \mathbf{E}(\boldsymbol{X}^2) - (\mathbf{E}(\boldsymbol{X}))^2$$

$$\mathbf{sd}(\boldsymbol{X}) = \sqrt{\mathbf{Var}(\boldsymbol{X})}.$$

**Example 3.25** For the Pareto distribution, introduced in Example 3.19, we have:

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

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 $\infty$ is 1. This integral converges only if $\alpha - 1 > 0$, i.e. if $\alpha > 1$.

Similarly:

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

and hence:

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

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

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

### 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 $\alpha$ 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 $\alpha$ 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.

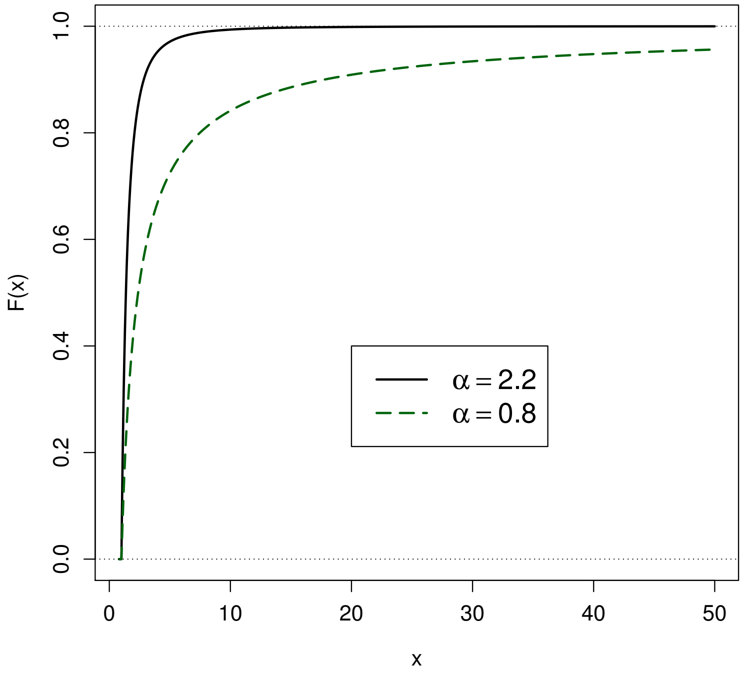


**Figure 3.8:** Pareto distribution cdfs.

**Example 3.26** Consider the exponential distribution introduced in Example 3.24. 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 $g = -\mathrm{e}^{-\lambda x}$). Therefore:

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

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

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

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

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

### 3.5.1 Moment generating functions

The moment generating function (mgf) of a continuous random variable $X$ is defined as for discrete random variables, with summation replaced by integration:

$$\boldsymbol{M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \int_{-\infty}^{\infty} \mathrm{e}^{tx} f(x)\,\mathrm{d}x.}$$

The properties of the mgf stated in Section 3.4.8 also hold for continuous distributions.

If the expected value $\mathrm{E}(\mathrm{e}^{tX})$ is infinite, the random variable $X$ does not have an mgf. For example, the Pareto distribution does not have an mgf for positive $t$.

**Example 3.27** For the exponential distribution, we have:

$$\begin{aligned}M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) &= \int_0^\infty \mathrm{e}^{tx}\lambda\mathrm{e}^{-\lambda x}\,\mathrm{d}x = \int_0^\infty \lambda\mathrm{e}^{-(\lambda-t)x}\,\mathrm{d}x \\ &= \frac{\lambda}{\lambda - t}\underbrace{\int_0^\infty (\lambda - t)\mathrm{e}^{-(\lambda-t)x}\,\mathrm{d}x}_{=1} = \frac{\lambda}{\lambda-t} \quad (\text{for } t < \lambda)\end{aligned}$$

from which we get $M_X'(t) = \lambda/(\lambda-t)^2$ and $M_X''(t) = 2\lambda/(\lambda-t)^3$, so:

$$\mathrm{E}(X) = M_X'(0) = \frac{1}{\lambda} \quad \text{and} \quad \mathrm{E}(X^2) = M_X''(0) = \frac{2}{\lambda^2}$$

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

These agree with the results derived with a bit more work in Example 3.26.

### 3.5.2 Median of a random variable

Recall 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:

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

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

A more precise general definition of the median of any probability distribution is as follows.

Let $X$ be a random variable with the cumulative distribution function $F(x)$. The median $m$ of $X$ is any number which satisfies:

$$P(X \leq m) = F(m) \geq 0.5$$

and:

$$P(X \geq m) = 1 - F(m) + P(X = m) \geq 0.5.$$

For a continuous distribution $P(X = m) = 0$ for any $m$, so this reduces to $F(m) = 0.5$.

If, for a discrete distribution, $F(x_m) = 0.5$ exactly for some value $x_m$, the median is not unique. Instead, all values from $x_m$ to the next largest observation (these included) are medians.

**Example 3.28** For the Pareto distribution we have:

$$F(x) = 1 - \left(\frac{k}{x}\right)^{\alpha} \quad \text{for } x \geq k.$$

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

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

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 3.29** For the exponential distribution we have:

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

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

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

## 3.6 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.

## 3.7 Key terms and concepts

- Binomial distribution
- Continuous
- Discrete
- Expected value
- Exponential distribution
- Median
- Outcome
- Pareto distribution
- Probability distribution
- Random variable
- Step function
- Constant
- Cumulative distribution function
- Estimators
- Experiment
- Interval
- Moment generating function
- Parameter
- Probability density function
- Probability (mass) function
- Standard deviation
- Variance

> *The death of one man is a tragedy. The death of millions is a statistic.*
> (Stalin to Churchill, Potsdam 1945)

# 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 Learning outcomes

After completing this chapter, 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.

## 4.3 Introduction

In statistical inference we will treat observations:

$$\boldsymbol{X_1, X_2, \ldots, X_n}$$

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

How to choose the probability distribution?

- Usually we do not try to invent new 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: a finite versus an infinite number of possible values
- among continuous: different sets of possible values (for example, all real numbers $x$, $x \geq 0$, 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 leaving 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 $\pi$ (the probability that $X_i = 1$) is appropriate.
> - Within the family of Bernoulli distributions, we use the one where the value of $\pi$ is our best estimate based on the observed data. This is $\widehat{\pi} = 513/950 = 0.54$.

### Distributions in the examination

For the discrete uniform, Bernoulli, binomial, Poisson, continuous uniform, exponential and normal distributions:

- you should memorise their pf/pdf, cdf (if given), mean, variance and median (if given)
- you can use these in any examination question without proof, *unless* the question directly asks you to derive them again.

For any other distributions:

- you do not need to memorise their pf/pdf or cdf; if needed for a question, these will be provided
- if a question involves means, variances or other properties of these distributions, these will either be provided, or the question will ask you to derive them.

## 4.4 Common discrete distributions

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

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

### 4.4.1 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:

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

> **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:

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

and:

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

Therefore:

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

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

### 4.4.2 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:

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

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 $\pi$. This is often written as:

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

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

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

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

and:

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

The moment generating function is:

$$\boldsymbol{M_X(t) = \sum_{x=0}^{1} \text{e}^{tx} p(x) = \text{e}^0 (1-\pi) + \text{e}^t \pi = (1-\pi) + \pi \text{e}^t.}$$

### 4.4.3 Binomial distribution

Suppose we carry out $n$ Bernoulli trials such that:

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

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

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

The binomial distribution was first encountered in Example 3.15.

**Example 4.4** A multiple choice test has 4 questions, each with 4 possible answers. James 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 James' test. $X$ follows the binomial distribution with $n = 4$ and $\pi = 0.25$, i.e. we have:

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

For example, what is the probability that James 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 1s and 0s, 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 which 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 $X \sim \mathrm{Bin}(n, \pi)$ is:

$$\boldsymbol{p(x) = \begin{cases} \binom{n}{x}\pi^x(1-\pi)^{n-x} & \text{for } x = 0, 1, 2, \ldots, n \\ 0 & \text{otherwise.} \end{cases}} \tag{4.5}$$

We have already shown that (4.5) satisfies the conditions for being a probability function in the previous chapter (see Example 3.15).

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

$$p(0) = \binom{4}{0} \times (0.25)^0 \times (0.75)^4 = 0.3164, \qquad p(1) = \binom{4}{1} \times (0.25)^1 \times (0.75)^3 = 0.4219,$$

$$p(2) = \binom{4}{2} \times (0.25)^2 \times (0.75)^2 = 0.2109, \qquad p(3) = \binom{4}{3} \times (0.25)^3 \times (0.75)^1 = 0.0469,$$

$$p(4) = \binom{4}{4} \times (0.25)^4 \times (0.75)^0 = 0.0039.$$

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

$$\mathbf{E(X) = n\pi}$$

and:

$$\mathbf{Var(X) = n\pi(1-\pi).}$$

The expected value $\mathrm{E}(X)$ was derived in the previous chapter (see Example 3.15). The variance will be derived later.

These can also be obtained from the moment generating function:

$$\boldsymbol{M_X(t) = ((1-\pi) + \pi e^t)^n.}$$

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

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 $x = 0, 1, 2, \ldots, 20$ 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 $\pi$ of the correct answer for every question, so that $X \sim \mathrm{Bin}(20, \pi)$. Figure 4.1 shows plots of the probabilities for $\pi = 0.25$, 0.5, 0.7 and 0.9.

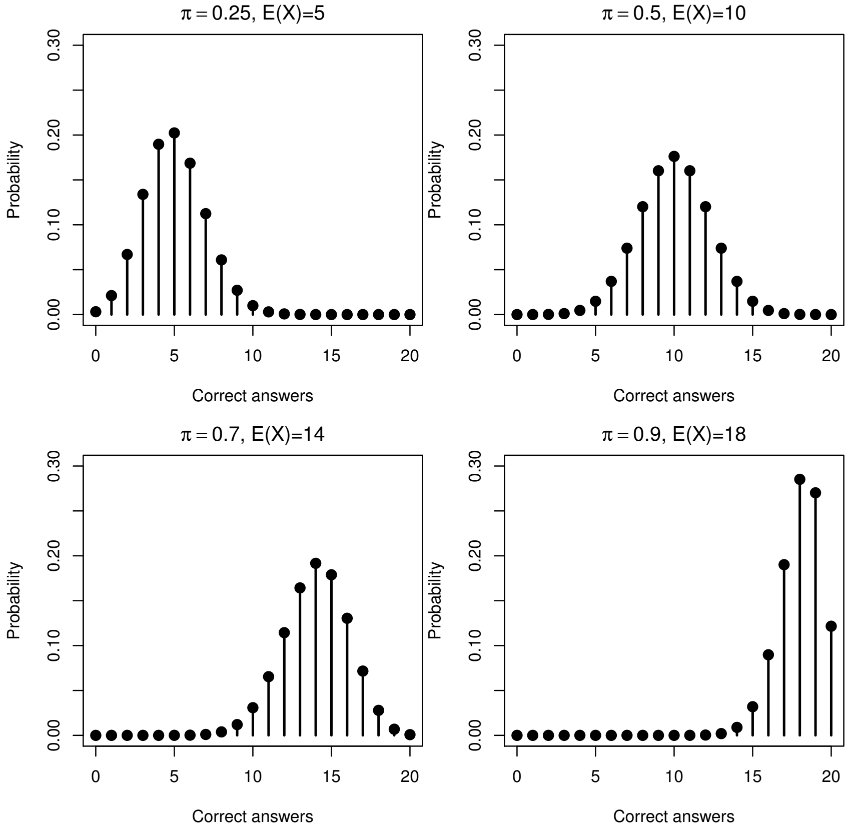


**Figure 4.1:** Probability plots for Example 4.6.

### 4.4.4 Poisson distribution

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

**Poisson distribution probability function**

The probability function of the Poisson distribution is:

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

where $\lambda > 0$ is a parameter.

If a random variable $X$ has a Poisson distribution with parameter $\lambda$, 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:

$$\text{E}(X) = \lambda$$

and:

$$\text{Var}(X) = \lambda.$$

These can also be obtained from the moment generating function (see Example 3.17):

$$M_X(t) = e^{\lambda(e^t - 1)}.$$

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 which 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 $\lambda t$ for some constant $\lambda > 0$.

In essence, these state that individual occurrences should be independent, sufficiently rare, and happen at a constant rate $\lambda$ 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 $\lambda$, i.e. $X \sim \text{Poisson}(\lambda)$.

The single parameter $\lambda$ 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:
>
> - The number of telephone calls received at a call centre *per* minute.
> - The number of accidents on a stretch of motorway *per* week.
> - The number of customers arriving at a checkout *per* minute.
> - The number of misprints *per* page of newsprint.

Because $\lambda$ 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, $Y \sim \text{Poisson}(1.5 \times 2) = \text{Poisson}(3)$.

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

Both motivations suggest that distributions with higher values of $\lambda$ 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 $X \sim \text{Poisson}(4)$.

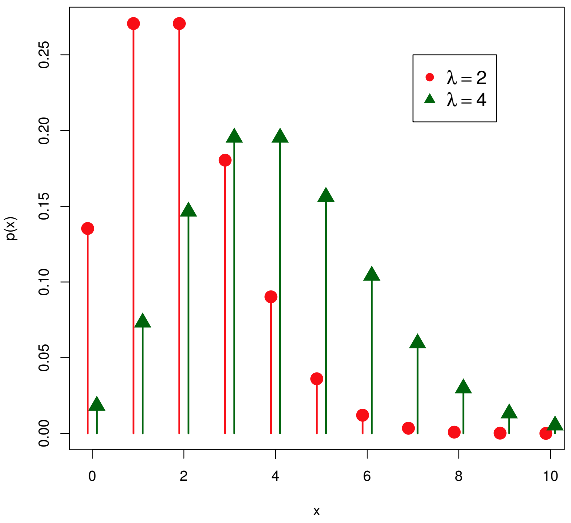


**Figure 4.2:** Probability plots for Example 4.9.

**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:

$$X \sim \text{Poisson}(1.6)$$

and:

$$Y \sim \text{Poisson}(1.6 \times 5) = \text{Poisson}(8).$$

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:

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

### 4.4.5 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($\pi$) is the same distribution as Bin($1, \pi$).

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.

### 4.4.6 Poisson approximation of the binomial distribution

Suppose that:

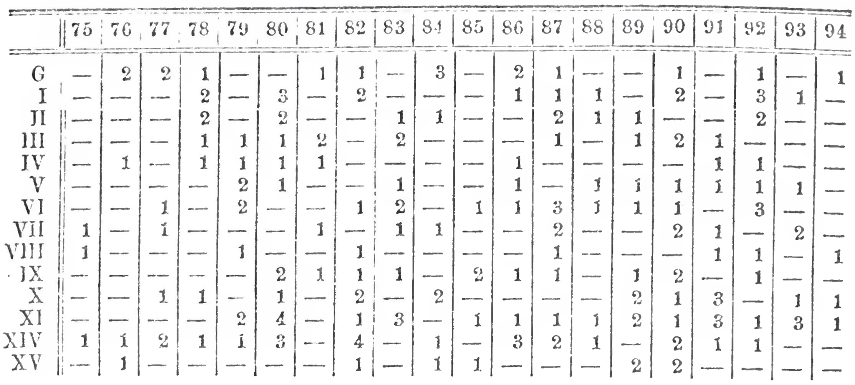
- $\boldsymbol{X \sim \mathrm{Bin}(n, \pi)}$
- **$\boldsymbol{n}$ is large and $\boldsymbol{\pi}$ is small.**

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

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

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 $X \sim \mathrm{Bin}(n, \pi)$, where:
>
> - $n$ is large – the number of men in a corps (perhaps 50,000)
> - $\pi$ is small – the probability that a man is killed by a horsekick.
>
> $X$ should be well-approximated by a Poisson distribution with some mean $\lambda$. 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 $\lambda$ for the Poisson distribution. $X \sim \mathrm{Poisson}(0.7)$ 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 spanning 1875–94. Source: Bortkiewicz (1898) *Das Gesetz der kleinen Zahlen*, Leipzig: Teubner.

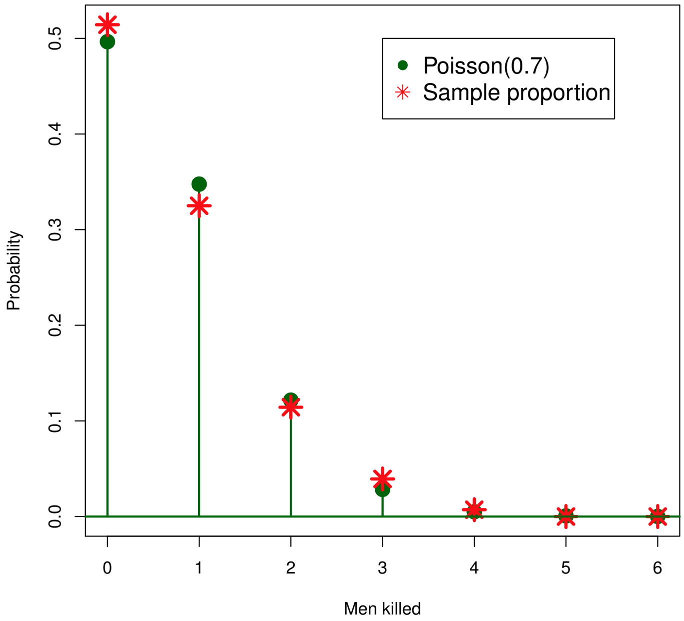


**Figure 4.4:** Fit of Poisson distribution to the data in Example 4.11.

**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:

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

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

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

### 4.4.7 Some other discrete distributions

Just their names and short comments are given here, so that you have an idea of what else there is. You may meet some of these in future courses.

- **Geometric($\pi$) distribution.**
  - Distribution of the number of failures in Bernoulli trials before the first success.
  - $\pi$ is the probability of success at each trial.
  - The sample space is $0, 1, 2, \ldots$.
  - See the basketball example in Chapter 3.
- **Negative binomial($r, \pi$) distribution.**
  - Distribution of the number of failures in Bernoulli trials before $r$ successes occur.
  - $\pi$ is the probability of success at each trial.
  - The sample space is $0, 1, 2, \ldots$.
  - Negative binomial$(1, \pi)$ is the same as Geometric$(\pi)$.
- **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.
  - The 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 $\pi_i$s 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 $\pi_i$.
  - 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 distribution is the distribution of a *multivariate* random variable, as discussed later in the course.

## 4.5 Common continuous distributions

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

- **Uniform distribution.**
- **Exponential distribution.**
- **Normal distribution.**

### 4.5.1 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 the 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:

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

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, $f(x) \ge 0$ for all $x$, and:

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

The cdf is:

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

Therefore, the probability of an interval $[x_1, x_2]$, where $a \le x_1 < x_2 \le b$, is:

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

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

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

$$\mathrm{E}(X) = \frac{a+b}{2} = \text{median of } X$$

and:

$$\text{Var}(X) = \frac{(b-a)^2}{12}.$$

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

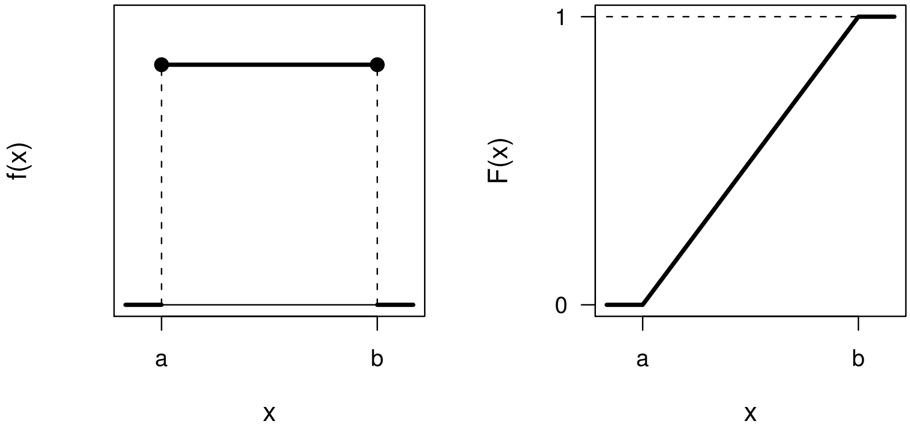


**Figure 4.5:** Continuous uniform distribution pdf (left) and cdf (right).

### 4.5.2 Exponential distribution

**Exponential distribution pdf**

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

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

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 3.24). The general shape of the pdf is that of 'exponential decay', as shown in Figure 4.6 (hence the name).

The cdf of the Exp($\lambda$) distribution is:

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

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

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

$$E(X) = \frac{1}{\lambda}$$

and:

$$\text{Var}(X) = \frac{1}{\lambda^2}.$$

These have been derived in the previous chapter (see Example 3.26). The median of the distribution, also previously derived (see Example 3.29), is:

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

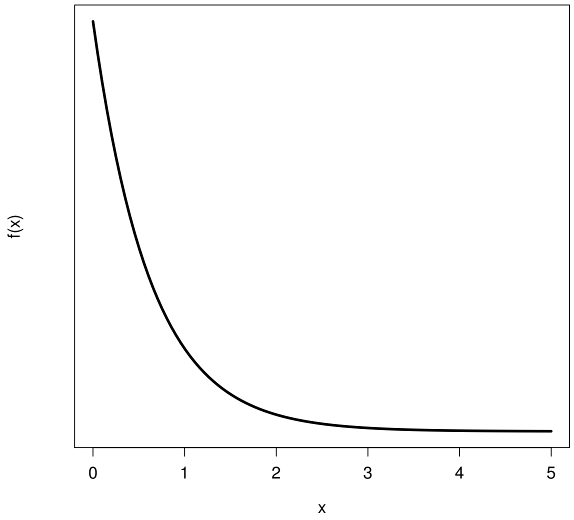


**Figure 4.6:** Exponential distribution pdf.

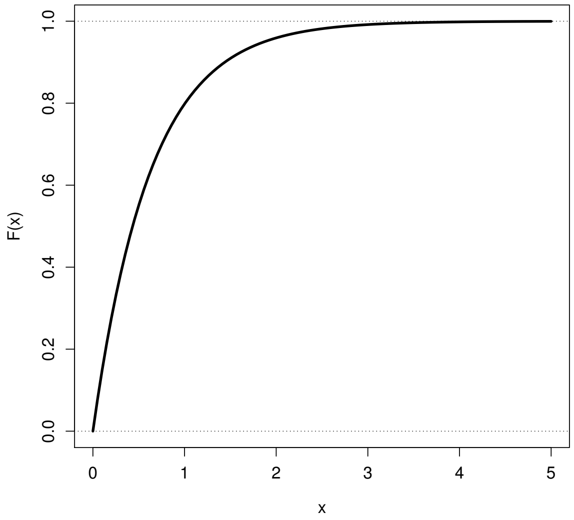


**Figure 4.7:** Exponential distribution cdf for $\lambda = 1.6$.

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

The moment generating function of the exponential distribution (derived in Example 3.27) is:

$$\boldsymbol{M_X(t) = \frac{\lambda}{\lambda - t} \quad \text{for } t < \lambda.}$$

### 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 $\lambda$, the time interval (measured in the same units of time) between two successive events has an exponential distribution with the same parameter $\lambda$.

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

- $\mathrm{E}(X) = \lambda$ for $\mathrm{Pois}(\lambda)$, i.e. a large $\lambda$ means many events per unit of time, on average.
- $\mathrm{E}(X) = 1/\lambda$ for $\mathrm{Exp}(\lambda)$, i.e. a large $\lambda$ 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 $\lambda = 1.6$.
- Therefore, 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 $(\ln 2) \times 0.625 = 0.433$.

We can also calculate probabilities of waiting times between arrivals, using the cumulative distribution function:

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - \mathrm{e}^{-1.6x} & \text{for } x \geq 0. \end{cases}$$

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.
- $P(X > 3) = 1 - F(3) = \mathrm{e}^{-1.6\times 3} = \mathrm{e}^{-4.8} = 0.0082.$

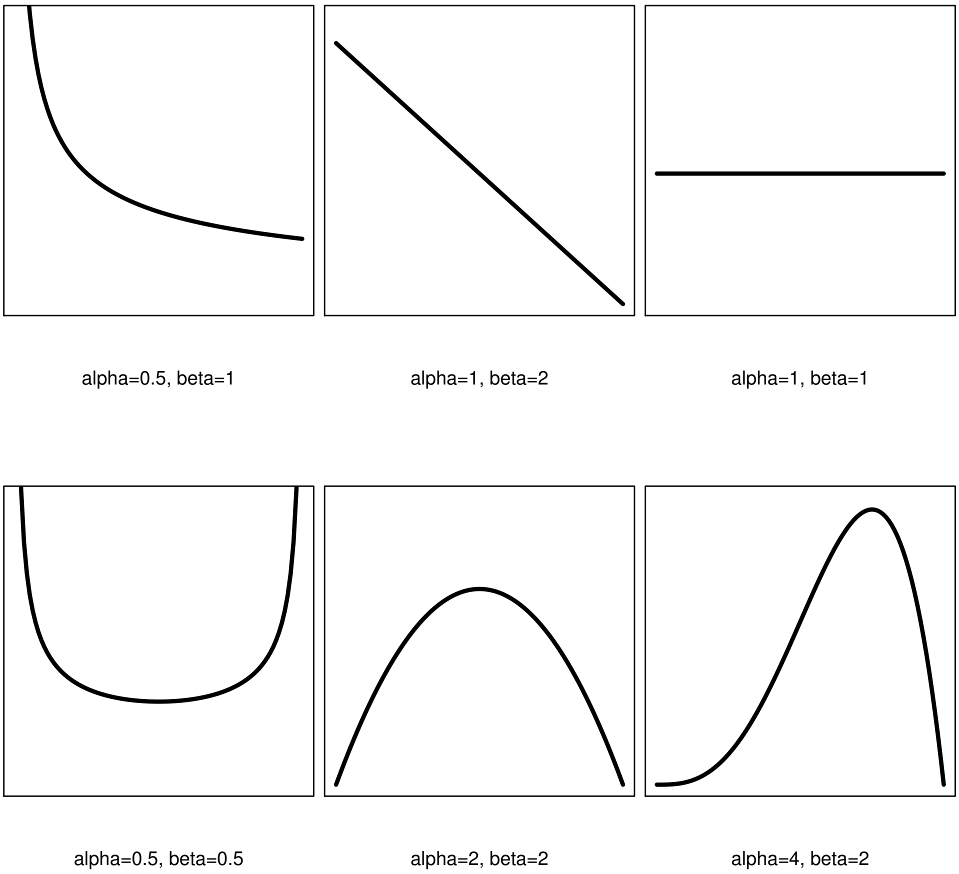
  The probability of a gap of 3 minutes or more between arrivals is very small.

### 4.5.3 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. You may meet these again in future courses.

- **Beta($\alpha, \beta$) distribution**, shown in Figure 4.8.
  - Generalising the uniform distribution, these are distributions for a closed interval, which is taken to be $[0, 1]$.
  - Therefore, the sample space is $\{x \mid 0 \leq x \leq 1\}$.
  - Unlike for the uniform distribution, the pdf is generally not flat.
  - $\mathrm{Beta}(1, 1)$ is the same as $\mathrm{Uniform}[0, 1]$.

- **Gamma($\alpha, \beta$) distribution**, shown in Figure 4.9.
  - Generalising the exponential distribution, this is a two-parameter family of skewed distributions for positive values.
  - The sample space is $\{x \,|\, x > 0\}$.
  - Gamma$(1, \beta)$ is the same as Exp$(\beta)$.



alpha=0.5, beta=1

alpha=1, beta=2

alpha=1, beta=1

alpha=0.5, beta=0.5

alpha=2, beta=2

alpha=4, beta=2

**Figure 4.8:** Beta distribution density functions.

### 4.5.4 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 which are *approximately* normal, for example heights of humans or 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

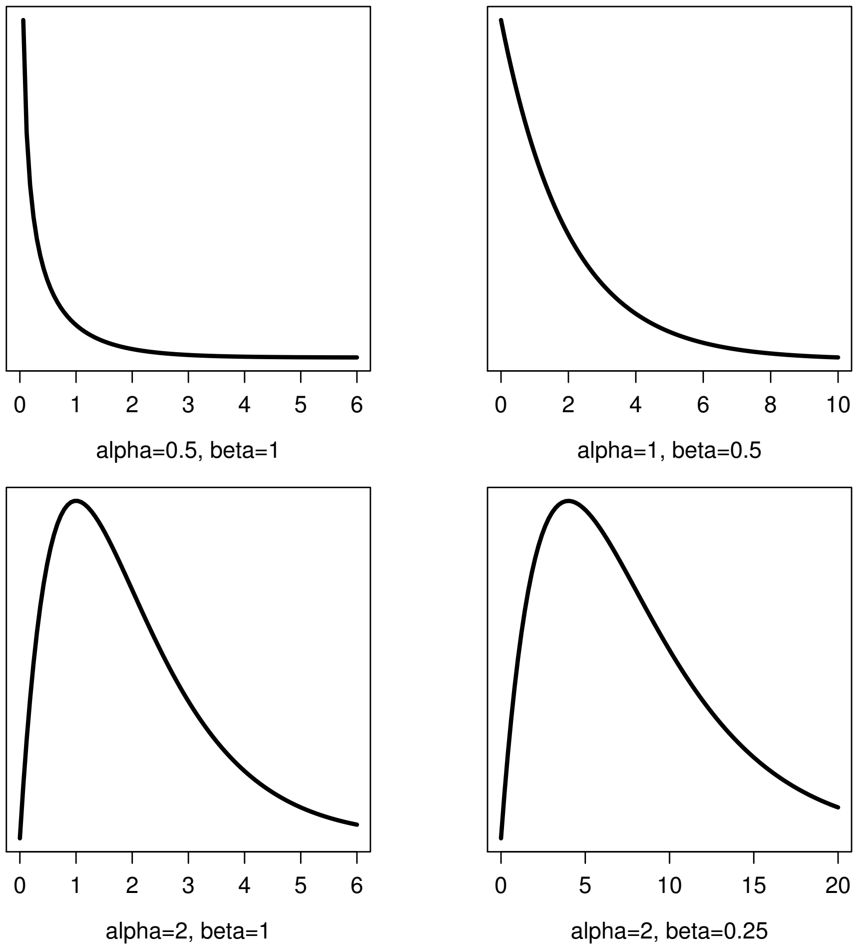


**Figure 4.9:** Gamma distribution density functions.

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:

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

where $\pi$ is the mathematical constant (i.e. $\pi = 3.14159\ldots$), and $\mu$ and $\sigma^2$ 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 $\mu$ and variance $\sigma^2$, denoted $\boldsymbol{X \sim N(\mu, \sigma^2)}$.

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.

The proof of the second point, which is somewhat elaborate, is shown in a separate note on the ST102 Moodle site. This note is not examinable!

If $X \sim N(\mu, \sigma^2)$, then:

$$\mathbf{E}(\boldsymbol{X}) = \boldsymbol{\mu}$$

and:

$$\mathbf{Var}(\boldsymbol{X}) = \boldsymbol{\sigma}^2$$

and, therefore, the standard deviation is $\text{sd}(X) = \sigma$.

A (non-examinable) proof of this is given in a separate note on the ST102 Moodle site. It uses the moment generating function of the normal distribution, which is shown to be:

$$M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right) \quad \text{for} \quad -\infty < t < \infty.$$

The mean can also be inferred from the observation that the normal pdf is **symmetric** about $\mu$. This also implies that the median of the normal distribution is $\mu$.

The normal density is the so-called 'bell curve'. The two parameters affect it as follows.

- The mean $\mu$ determines the location of the curve.
- The variance $\sigma^2$ 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.

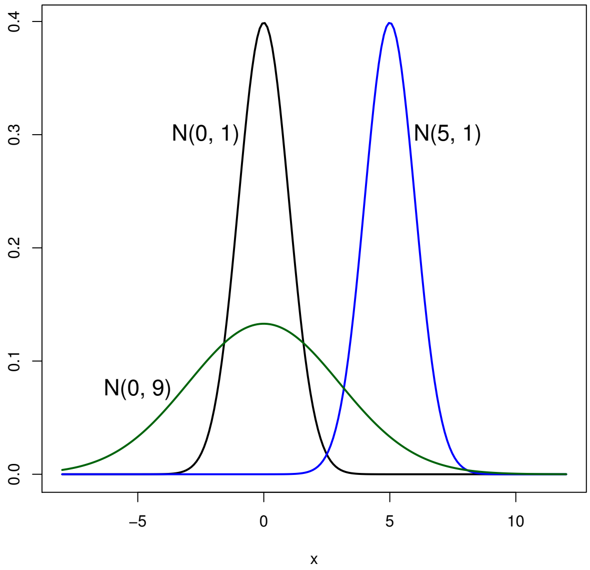


**Figure 4.10:** Various normal distributions.

## 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:

$$\boldsymbol{Y = aX + b \sim N(a\mu + b, a^2\sigma^2).} \tag{4.7}$$

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 $X$.

Let us apply (4.7) with $a = 1/\sigma$ and $b = -\mu/\sigma$, to get:

$$\boldsymbol{Z = \frac{1}{\sigma}X - \frac{\mu}{\sigma} = \frac{X - \mu}{\sigma} \sim N\left(\frac{1}{\sigma}\mu - \frac{\mu}{\sigma}, \left(\frac{1}{\sigma}\right)^2 \sigma^2\right) = N(0, 1).}$$

The transformed variable $\boldsymbol{Z = (X - \mu)/\sigma}$ 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:

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

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 $\boldsymbol{\Phi(x)}$.

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 $\Phi(z)$, the cdf of $Z \sim N(0, 1)$ (Table 3 in Murdoch and Barnes' *Statistical Tables*). This is also on the ST102 Moodle site for use in the exercises.

Since Table 3 uses the notation $\Phi(z)$ (for $z$-score), we will do so too below. $\Phi(x)$ and $\Phi(z)$ mean the same thing, of course.

Table 3 shows values of $1 - \Phi(z) = P(Z > z)$ for $z \geq 0$. 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 $\mu$ and $\sigma^2$.
2. Even for $N(0,1)$, it only shows probabilities for $z \geq 0$.

We next show how these are not really limitations, starting with '2.'.

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 also $-Z \sim N(0,1)$.

Suppose that $z \geq 0$, so that $-z \leq 0$. Table 3 shows:

$$P(Z > z) = 1 - \Phi(z) = P_z$$

which is called $P_z$ for short. From it, we also get the following probabilities.

- $P(Z \leq z) = \Phi(z) = 1 - P(Z > z) = 1 - P_z$.
- $P(Z \leq -z) = \Phi(-z) = P(-Z \geq z) = P(Z \geq z) = P(Z > z) = P_z$.
- $P(Z > -z) = 1 - \Phi(-z) = P(-Z < z) = P(Z < z) = 1 - P_z$.

In each of these, $\leq$ can be replaced by $<$, and $\geq$ by $>$ (see Section 3.5). Figure 4.11 shows tail probabilities for the standard normal distribution.

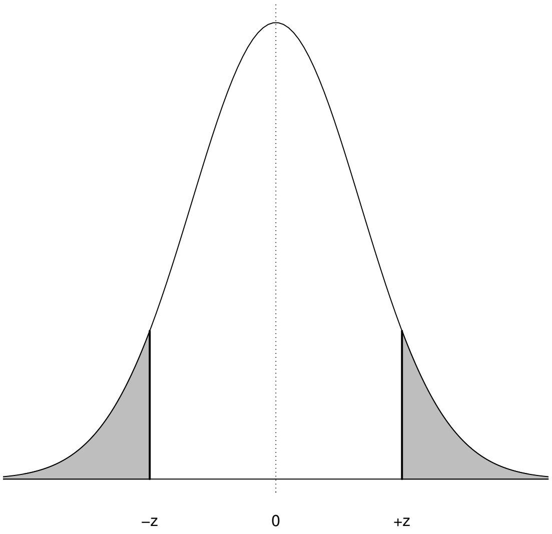


**Figure 4.11:** Tail probabilities for the standard normal distribution.

If $Z \sim N(0,1)$, for any two numbers $z_1 < z_2$, then:

$$P(z_1 < Z \leq z_2) = \Phi(z_2) - \Phi(z_1)$$

where $\Phi(z_2)$ and $\Phi(z_1)$ are obtained using the rules above.

*Reality check*: remember that:

$$\Phi(0) = P(Z \leq 0) = 0.5 = P(Z > 0) = 1 - \Phi(0).$$

So if you ever end up with results like $P(Z \leq -1) = 0.7$ or $P(Z \leq 1) = 0.2$ or $P(Z > 2) = 0.95$, these must be wrong! (See property 3 of cdfs in Section 3.4.4.)

**Example 4.15** Consider the 0.2005 value in the '0.8' row and '0.04' column of Table 3 of Murdoch and Barnes' *Statistical Tables*, which shows that:

$$1 - \Phi(0.84) = P(Z > 0.84) = 0.2005.$$

Using the results above, we then also have:

- $P(Z \leq 0.84) = \Phi(0.84) = 1 - 0.2005 = 0.7995$
- $P(Z \leq -0.84) = P(Z \geq 0.84) = 0.2005$
- $P(Z \geq -0.84) = P(Z \leq 0.84) = 0.7995$
- $P(-0.84 \leq Z \leq 0.84) = P(Z \leq 0.84) - P(Z \leq -0.84) = 0.5990.$

### Probabilities for any normal distribution

How about a normal distribution $X \sim N(\mu, \sigma^2)$, for any other $\mu$ and $\sigma^2$?

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 3 of Murdoch and Barnes' *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 4.16** 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 $\mu = 74.2$, $\sigma^2 = 127.87$ 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 3 of Murdoch and Barnes' *Statistical Tables*.

$$\begin{aligned}P(X > 90) &= P\left(\frac{X - 74.2}{11.31} > \frac{90 - 74.2}{11.31}\right)\\ &= P(Z > 1.40)\\ &= 1 - \Phi(1.40)\\ &= 1 - 0.9192\\ &= 0.0808\end{aligned}$$

and:

$$\begin{aligned}P(X < 60) &= P\left(\frac{X - 74.2}{11.31} < \frac{60 - 74.2}{11.31}\right)\\ &= P(Z < -1.26)\\ &= P(Z > 1.26)\\ &= 1 - \Phi(1.26)\\ &= 1 - 0.8962\\ &= 0.1038.\end{aligned}$$

Finally:

$$P(60 \leq X \leq 90) = P(X \leq 90) - P(X < 60) = 0.8152.$$

These probabilities are shown in Figure 4.12.

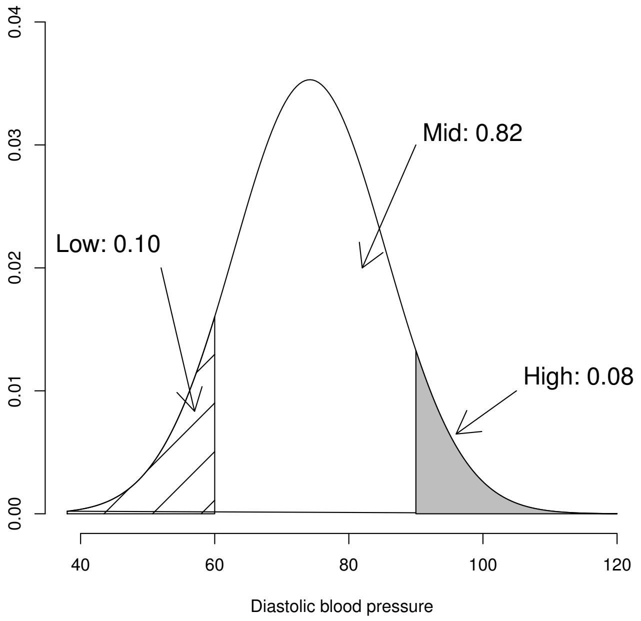


**Figure 4.12:** Distribution of blood pressure for Example 4.16.

### 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.
- $P(\mu - 1.96 \times \sigma < X < \mu + 1.96 \times \sigma) = 0.950$.
- $P(\mu - 2 \times \sigma < X < \mu + 2 \times \sigma) = 0.954$.
- $P(\mu - 2.58 \times \sigma < X < \mu + 2.58 \times \sigma) = 0.99$.
- $P(\mu - 3 \times \sigma < X < \mu + 3 \times \sigma) = 0.997$.

The first two of these are illustrated graphically in Figure 4.13.

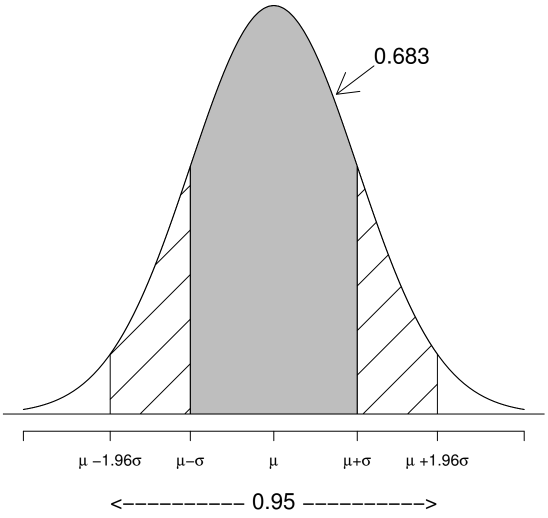


**Figure 4.13:** Some probabilities around the mean for the normal distribution.

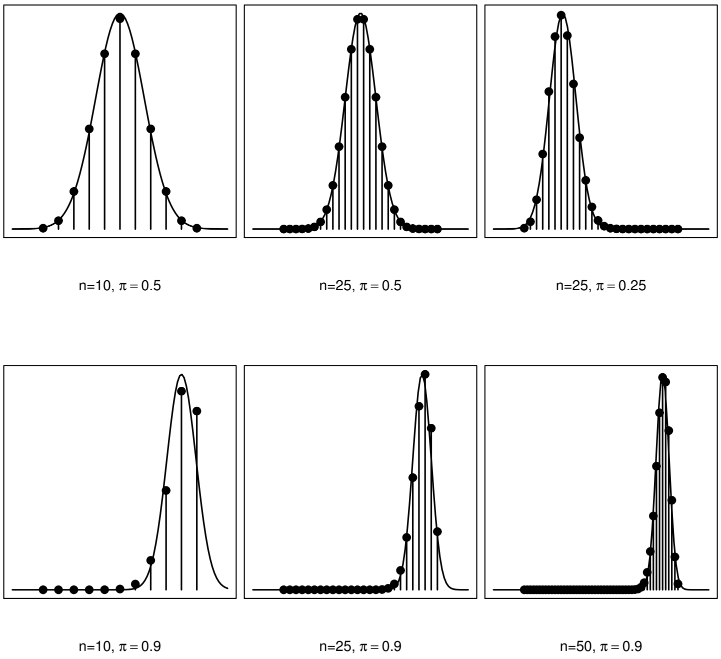
## 4.5.5 Normal approximation of the binomial distribution

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 distribution is well-approximated by the normal distribution when the number of trials $n$ is reasonably large.

For a given $n$, the approximation is best when $\pi$ 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 4.14 for different values of $n$ and $\pi$. 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 3 of Murdoch and Barnes' *Statistical Tables*.



n=10, $\pi = 0.5$ n=25, $\pi = 0.5$ n=25, $\pi = 0.25$

n=10, $\pi = 0.9$ n=25, $\pi = 0.9$ n=50, $\pi = 0.9$

**Figure 4.14:** 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 $x = 0, 1, 2, \ldots, 40$, 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:

$$\begin{aligned} P(X < 4) = P(X \leq 3) \quad &\Rightarrow \quad P(Y < 3.5) \qquad \text{(since 4 is excluded)} \\ P(X \leq 4) = P(X < 5) \quad &\Rightarrow \quad P(Y < 4.5) \qquad \text{(since 4 is included)} \\ P(1 \leq X < 6) = P(1 \leq X \leq 5) \quad &\Rightarrow \quad P(0.5 < Y < 5.5) \qquad \text{(since 1 to 5 are included).} \end{aligned}$$

**Example 4.17** 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 $X \sim \text{Bin}(1{,}000, 0.361)$.

1. What is the probability that $X \geq 400$?

   Using the normal approximation, noting $n = 1{,}000$ and $\pi = 0.361$, with $Y \sim N(1{,}000 \times 0.361, 1{,}000 \times 0.361 \times 0.639) = N(361, 230.68)$, 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 3 of Murdoch and Barnes' *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.

## 4.6 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 covered.

## 4.7 Key terms and concepts

- Bernoulli distribution
- Central limit theorem
- Continuous uniform distribution
- Exponential distribution
- Moment generating function
- Parameter
- Standardised variable
- $z$-score
- Binomial distribution
- Continuity correction
- Discrete uniform distribution
- Moment
- Normal distribution
- Poisson distribution
- Standard normal distribution

*There are two kinds of statistics, the kind you look up and the kind you make up.*
(Rex Stout)

# Chapter 5
# Multivariate random variables

## 5.1 Synopsis of chapter

Almost all applications of statistical methods deal with several measurements on the same, or connected, items. To think statistically about several measurements on a randomly selected item, you must understand some of the concepts for joint distributions of random variables.

## 5.2 Learning outcomes

After completing this chapter, you should be able to:

- arrange the probabilities for a discrete bivariate distribution in tabular form
- define marginal and conditional distributions, and determine them for a discrete bivariate distribution
- recall how to define and determine independence for two random variables
- define and compute expected values for functions of two random variables and demonstrate how to prove simple properties of expected values
- provide the definition of covariance and correlation for two random variables and calculate these.

## 5.3 Introduction

So far, we have considered **univariate** situations, that is one random variable at a time. Now we will consider **multivariate** situations, that is two or more random variables at once, and together.

In particular, we consider two somewhat different types of multivariate situations.

1. Several different variables – such as the height and weight of a person.
2. Several observations of the same variable, considered together – such as the heights of all $n$ people in a sample.

Suppose that $X_1, X_2, \ldots, X_n$ are random variables, then the vector:

$$\mathbf{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n})'$$

is a **multivariate random variable** (here $n$-variate), also known as a **random vector**. Its possible values are the vectors:

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)'$$

where each $x_i$ is a possible value of the random variable $X_i$, for $i = 1, 2, \ldots, n$.

The **joint probability distribution** of a multivariate random variable $\mathbf{X}$ is defined by the possible values $\mathbf{x}$, and their probabilities.

For now, we consider just the simplest multivariate case, a **bivariate** random variable where $n = 2$. This is sufficient for introducing most of the concepts of multivariate random variables.

For notational simplicity, we will use $X$ and $Y$ instead of $X_1$ and $X_2$. A bivariate random variable is then the pair $(X, Y)$.

> **Example 5.1** In this chapter, we consider the following examples.
>
> **Discrete** bivariate example – for a football match:
>
> - $X$ = the number of goals scored by the home team
> - $Y$ = the number of goals scored by the visiting (away) team.
>
> **Continuous** bivariate example – for a person:
>
> - $X$ = the person's height
> - $Y$ = the person's weight.

## 5.4 Joint probability functions

When the random variables in $(X_1, X_2, \ldots, X_n)$ are either all discrete or all continuous, we also call the multivariate random variable either discrete or continuous, respectively.

For a discrete multivariate random variable, the joint probability distribution is described by the **joint probability function**, defined as:

$$p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

for all vectors $(x_1, x_2, \ldots, x_n)$ of $n$ real numbers. The value $p(x_1, x_2, \ldots, x_n)$ of the joint probability function is itself a single number, not a vector.

In the bivariate case, this is:

$$p(x, y) = P(X = x, Y = y)$$

which we sometimes write as $p_{X,Y}(x, y)$ to make the random variables clear.

**Example 5.2** Consider a randomly selected football match in the English Premier League (EPL), and the two random variables:

- $X$ = the number of goals scored by the home team
- $Y$ = the number of goals scored by the visiting (away) team.

Suppose both variables have possible values 0, 1, 2 and 3 (to keep this example simple, we have recorded the small number of scores of 4 or greater also as 3).

Consider the joint distribution of $(X, Y)$. We use probabilities based on data from the 2009–10 EPL season.

Suppose the values of $p_{X,Y}(x, y) = p(x, y) = P(X = x, Y = y)$ are the following:

| | $Y = y$ | | | |
|---|---|---|---|---|
| $X = x$ | 0 | 1 | 2 | 3 |
| 0 | 0.100 | 0.031 | 0.039 | 0.031 |
| 1 | 0.100 | 0.146 | 0.092 | 0.015 |
| 2 | 0.085 | 0.108 | 0.092 | 0.023 |
| 3 | 0.062 | 0.031 | 0.039 | 0.006 |

and $p(x, y) = 0$ for all other $(x, y)$.

Note that this satisfies the conditions for a probability function.

1. $p(x, y) \geq 0$ for all $(x, y)$.
2. $\sum\limits_{x=0}^{3} \sum\limits_{y=0}^{3} p(x, y) = 0.100 + 0.031 + \cdots + 0.006 = 1.000.$

The joint probability function gives probabilities of values of $(X, Y)$, for example:

- A 1–1 draw, which is the most probable single result, has probability

$$P(X = 1, Y = 1) = p(1, 1) = 0.146.$$

- The match is a draw with probability:

$$P(X = Y) = p(0, 0) + p(1, 1) + p(2, 2) + p(3, 3) = 0.344.$$

- The match is won by the home team with probability:

$$P(X > Y) = p(1, 0) + p(2, 0) + p(2, 1) + p(3, 0) + p(3, 1) + p(3, 2) = 0.425.$$

- More than 4 goals are scored in the match with probability:

$$P(X + Y > 4) = p(2, 3) + p(3, 2) + p(3, 3) = 0.068.$$

## 5.5 Marginal distributions

Consider a multivariate discrete random variable $\mathbf{X} = (X_1, X_2, \ldots, X_n)$.

The **marginal distribution** of a subset of the variables in $\mathbf{X}$ is the (joint) distribution of this subset. The joint pf of these variables (the **marginal pf**) is obtained by summing the joint pf of $\mathbf{X}$ over the variables which are *not* included in the subset.

**Example 5.3** Consider $\mathbf{X} = (X_1, X_2, X_3, X_4)$, and the marginal distribution of the subset $(X_1, X_2)$. The marginal pf of $(X_1, X_2)$ is:

$$p_{1,2}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = \sum_{x_3} \sum_{x_4} p(x_1, x_2, x_3, x_4)$$

where the sum is of the values of the joint pf of $(X_1, X_2, X_3, X_4)$ over all possible values of $X_3$ and $X_4$.

The simplest marginal distributions are those of individual variables in the multivariate random variable.

The marginal pf is then obtained by summing the joint pf over all the *other* variables. The resulting marginal distribution is univariate, and its pf is a univariate pf.

**Marginal distributions for discrete bivariate distributions**

For the bivariate distribution of $(X, Y)$ the univariate marginal distributions are those of $X$ and $Y$ individually. Their marginal pfs are:

$$\boldsymbol{p_X(x) = \sum_{y} p(x,y)} \quad \textbf{and} \quad \boldsymbol{p_Y(y) = \sum_{x} p(x,y).}$$

**Example 5.4** Continuing with the football example introduced in Example 5.2, the joint and marginal probability functions are:

| | $Y = y$ | | | | |
|---|---|---|---|---|---|
| $X = x$ | 0 | 1 | 2 | 3 | $p_X(x)$ |
| 0 | 0.100 | 0.031 | 0.039 | 0.031 | 0.201 |
| 1 | 0.100 | 0.146 | 0.092 | 0.015 | 0.353 |
| 2 | 0.085 | 0.108 | 0.092 | 0.023 | 0.308 |
| 3 | 0.062 | 0.031 | 0.039 | 0.006 | 0.138 |
| $p_Y(y)$ | 0.347 | 0.316 | 0.262 | 0.075 | 1.000 |

and $p(x, y) = p_X(x) = p_Y(y) = 0$ for all other $(x, y)$.

For example:

$$\begin{aligned} p_X(0) = \sum_{y=0}^{3} p(0,y) &= p(0,0) + p(0,1) + p(0,2) + p(0,3) \\ &= 0.100 + 0.031 + 0.039 + 0.031 \\ &= 0.201. \end{aligned}$$

Even for a multivariate random variable, expected values $\mathrm{E}(X_i)$, variances $\mathrm{Var}(X_i)$ and medians of individual variables are obtained from the *univariate* (marginal) distributions of $X_i$, as defined in Chapter 3.

**Example 5.5** Consider again the football example.

- The expected number of goals scored by the home team is:

$$\mathrm{E}(X) = \sum_x x\, p_X(x) = 0 \times 0.201 + 1 \times 0.353 + 2 \times 0.308 + 3 \times 0.138 = 1.383.$$

- The expected number of goals scored by the visiting team is:

$$\mathrm{E}(Y) = \sum_y y\, p_Y(y) = 0 \times 0.347 + 1 \times 0.316 + 2 \times 0.262 + 3 \times 0.075 = 1.065.$$

## 5.6 Continuous multivariate distributions

If all the random variables in $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ are continuous, the joint distribution of $\mathbf{X}$ is specified by its **joint probability density function** $f(x_1, x_2, \ldots, x_n)$.

Marginal distributions are defined as in the discrete case, but with integration instead of summation.

There will be no questions on continuous multivariate joint probability density functions in the examination. Only discrete multivariate joint probability functions may appear in the examination. So just a brief example of the continuous case is given here, to give you an idea of such distributions.

**Example 5.6** For a randomly selected man (aged over 16) in England, let:

- $X$ = his height (in cm)
- $Y$ = his weight (in kg).

The univariate marginal distributions of $X$ and $Y$ are approximately normal, with:

$$X \sim N(174.9, (7.39)^2) \quad \text{and} \quad Y \sim N(84.2, (15.63)^2)$$

> and the bivariate joint distribution of $(X, Y)$ is a *bivariate normal distribution*.
>
> Plots of the univariate and bivariate probability density functions are shown in Figures 5.1, 5.2 and 5.3.

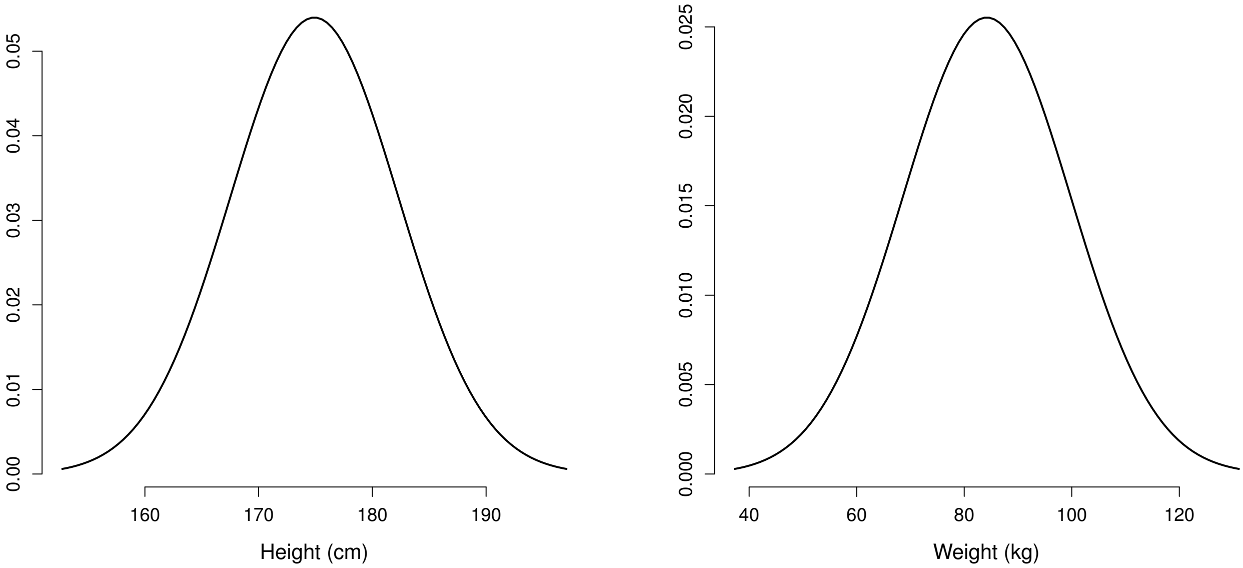


**Figure 5.1:** Univariate marginal pdfs for Example 5.6.

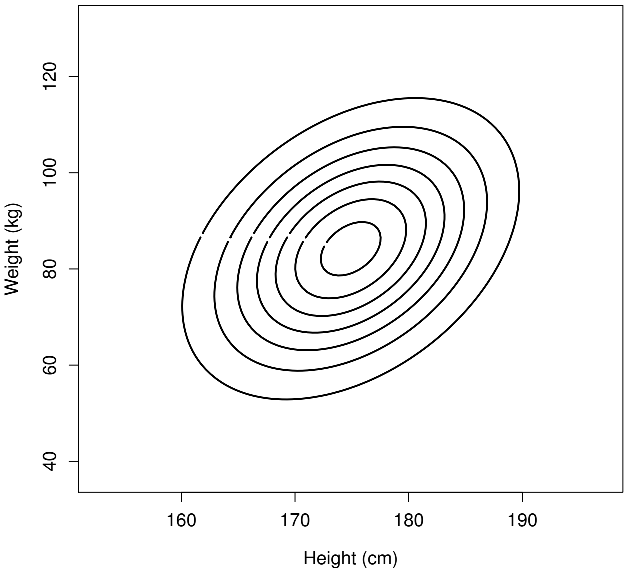


**Figure 5.2:** Bivariate joint pdf (contour plot) for Example 5.6.

## 5.7 Conditional distributions

Consider discrete variables $X$ and $Y$, with joint pf $p(x, y) = p_{X,Y}(x, y)$ and marginal pfs $p_X(x)$ and $p_Y(y)$, respectively.

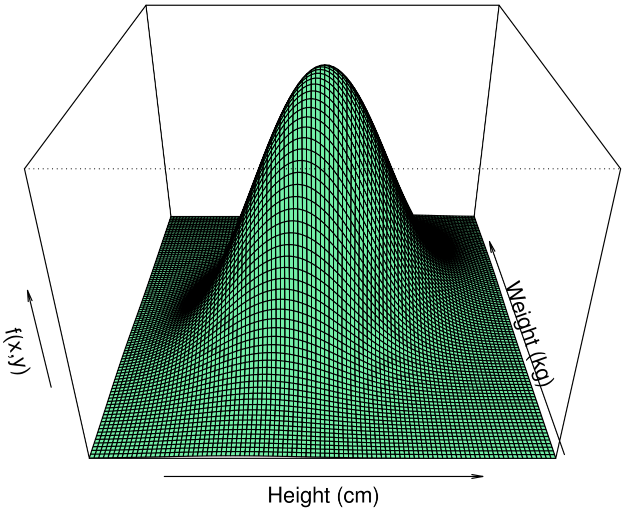


**Figure 5.3:** Bivariate joint pdf for Example 5.6.

### Conditional distributions of discrete bivariate distributions

Let $x$ be one possible value of $X$, for which $p_X(x) > 0$. The **conditional distribution** of $Y$ given that $X = x$ is the discrete probability distribution with the pf:

$$\boldsymbol{p_{Y|X}(y\,|\,x) = P(Y = y\,|\,X = x) = \frac{P(X = x \text{ and } Y = y)}{P(X = x)} = \frac{p_{X,Y}(x,y)}{p_X(x)}}$$

for any value $y$.

This is the **conditional probability function** of $Y$ given $X = x$.

**Example 5.7** Recall that in the football example the joint and marginal pfs were:

| | $Y = y$ | | | | |
|---|---|---|---|---|---|
| $X = x$ | 0 | 1 | 2 | 3 | $p_X(x)$ |
| 0 | 0.100 | 0.031 | 0.039 | 0.031 | 0.201 |
| 1 | 0.100 | 0.146 | 0.092 | 0.015 | 0.353 |
| 2 | 0.085 | 0.108 | 0.092 | 0.023 | 0.308 |
| 3 | 0.062 | 0.031 | 0.039 | 0.006 | 0.138 |
| $p_Y(y)$ | 0.347 | 0.316 | 0.262 | 0.075 | 1.000 |

We can now calculate the conditional pf of $Y$ given $X = x$ for each $x$, i.e. of away goals given home goals. For example:

$$p_{Y|X}(y\,|\,0) = p_{Y|X}(y\,|\,X = 0) = \frac{p_{X,Y}(0,y)}{p_X(0)} = \frac{p_{X,Y}(0,y)}{0.201}.$$

So, for example, $p_{Y|X}(1\,|\,0) = p_{X,Y}(0,1)/0.201 = 0.031/0.201 = 0.154$.

Calculating these for each value of $x$ gives:

|  | $p_{Y\|X}(y\,\|\,x)$ when $y$ is: |  |  |  |  |
|---|---|---|---|---|---|
| $X = x$ | 0 | 1 | 2 | 3 | Sum |
| 0 | 0.498 | 0.154 | 0.194 | 0.154 | 1.00 |
| 1 | 0.283 | 0.414 | 0.261 | 0.042 | 1.00 |
| 2 | 0.276 | 0.351 | 0.299 | 0.075 | 1.00 |
| 3 | 0.449 | 0.225 | 0.283 | 0.043 | 1.00 |

So, for example:

- if the home team scores 0 goals, the probability that the visiting team scores 1 goal is $p_{Y|X}(1\,|\,0) = 0.154$
- if the home team scores 1 goal, the probability that the visiting team wins the match is $p_{Y|X}(2\,|\,1) + p_{Y|X}(3\,|\,1) = 0.261 + 0.042 = 0.303$.

### 5.7.1 Properties of conditional distributions

Each different value of $x$ defines a different conditional distribution and conditional pf $p_{Y|X}(y\,|\,x)$. Each value of $p_{Y|X}(y\,|\,x)$ is a conditional probability of the kind previously defined. Defining events $A = \{Y = y\}$ and $B = \{X = x\}$, then:

$$\begin{aligned}P(A\,|\,B) = \frac{P(A \cap B)}{P(B)} &= \frac{P(Y = y \text{ and } X = x)}{P(X = x)}\\ &= P(Y = y\,|\,X = x)\\ &= \frac{p_{X,Y}(x,y)}{p_X(x)}\\ &= p_{Y|X}(y\,|\,x).\end{aligned}$$

A conditional distribution is itself a probability distribution, and a conditional pf is a pf. Clearly, $p_{Y|X}(y\,|\,x) \geq 0$ for all $y$, and:

$$\sum_y p_{Y|X}(y\,|\,x) = \frac{\sum_y p_{X,Y}(x,y)}{p_X(x)} = \frac{p_X(x)}{p_X(x)} = 1.$$

The conditional distribution and pf of $X$ given $Y = y$ (for any $y$ such that $p_Y(y) > 0$) is defined similarly, with the roles of $X$ and $Y$ reversed:

$$p_{X|Y}(x\,|\,y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

for any value $x$.

Conditional distributions are general and are not limited to the bivariate case. If $\mathbf{X}$ and/or $\mathbf{Y}$ are vectors of random variables, the conditional pf of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ is:

$$p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}\,|\,\mathbf{x}) = \frac{p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{p_{\mathbf{X}}(\mathbf{x})}$$

where $p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})$ is the joint pf of the random vector $(\mathbf{X},\mathbf{Y})$, and $p_{\mathbf{X}}(\mathbf{x})$ is the marginal pf of the random vector $\mathbf{X}$.

### 5.7.2 Conditional mean and variance

Since a conditional distribution is a probability distribution, it also has a mean (expected value) and variance (and median etc.).

These are known as the **conditional mean** and **conditional variance**, and are denoted, respectively, by:

$$\mathrm{E}_{Y|X}(Y\,|\,x) \quad \text{and} \quad \mathrm{Var}_{Y|X}(Y\,|\,x).$$

**Example 5.8** In the football example, we have:

$$\mathrm{E}_{Y|X}(Y\,|\,0) = \sum_y y\,p_{Y|X}(y\,|\,0) = 0\times 0.498 + 1\times 0.154 + 2\times 0.194 + 3\times 0.154 = 1.00.$$

So, if the home team scores 0 goals, the expected number of goals by the visiting team is $\mathrm{E}_{Y|X}(Y\,|\,0) = 1.00$.

$\mathrm{E}_{Y|X}(Y\,|\,x)$ for $x = 1$, 2 and 3 are obtained similarly.

Here $X$ is the number of goals by the home team, and $Y$ is the number of goals by the visiting team:

| | $p_{Y\|X}(y\,\|\,x)$ when $y$ is: | | | | |
|---|---|---|---|---|---|
| $X = x$ | 0 | 1 | 2 | 3 | $\mathrm{E}_{Y\|X}(Y\,\|\,x)$ |
| 0 | 0.498 | 0.154 | 0.194 | 0.154 | 1.00 |
| 1 | 0.283 | 0.414 | 0.261 | 0.042 | 1.06 |
| 2 | 0.276 | 0.351 | 0.299 | 0.075 | 1.17 |
| 3 | 0.449 | 0.225 | 0.283 | 0.043 | 0.92 |

Plots of the conditional means are shown in Figure 5.4.

### 5.7.3 Continuous conditional distributions

Suppose $X$ and $Y$ are continuous, with joint pdf $f_{X,Y}(x,y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$, respectively.

The conditional distribution of $Y$ given that $X = x$ is the continuous probability distribution with the pdf:

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

which is defined if $f_X(x) > 0$. For a conditional distribution of $X$ given $Y = y$, $f_{X|Y}(x\,|\,y)$ is defined similarly, with the roles of $X$ and $Y$ reversed.

Unlike in the discrete case, this is not a conditional probability. However, $f_{Y|X}(y\,|\,x)$ *is* a pdf of a continuous random variable, so the conditional distribution is itself a continuous probability distribution.

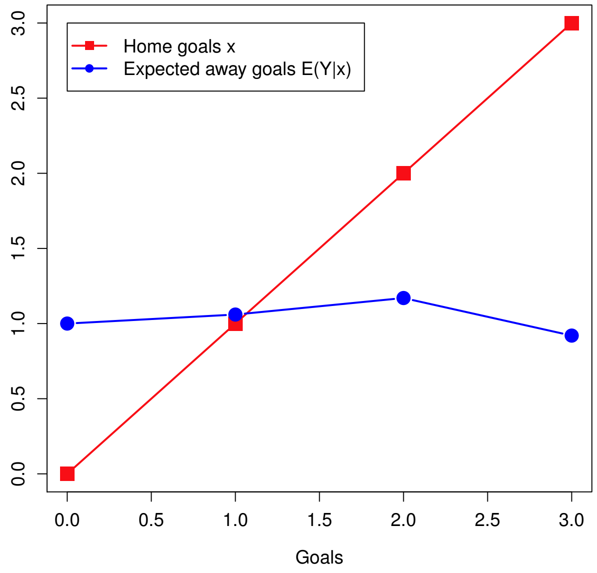


**Figure 5.4:** Conditional means for Example 5.8.

**Example 5.9** For a randomly selected man (aged over 16) in England, consider $X$ = height (in cm) and $Y$ = weight (in kg). The joint distribution of $(X, Y)$ is approximately bivariate normal (see Example 5.6).

The conditional distribution of $Y$ given $X = x$ is then a normal distribution for each $x$, with the following parameters:

$$\mathrm{E}_{Y|X}(Y \,|\, x) = -58.1 + 0.81x \quad \text{and} \quad \mathrm{Var}_{Y|X}(Y \,|\, x) = 208.$$

In other words, the conditional mean depends on $x$, but the conditional variance does not. For example:

$$\mathrm{E}_{Y|X}(Y \,|\, 160) = 71.5 \quad \text{and} \quad \mathrm{E}_{Y\,|\,X}(Y \,|\, 190) = 95.8.$$

For women, this conditional distribution is normal with the following parameters:

$$\mathrm{E}_{Y|X}(Y \,|\, x) = -23.0 + 0.58x \quad \text{and} \quad \mathrm{Var}_{Y|X}(Y \,|\, x) = 221.$$

The conditional means are shown in Figure 5.5.

## 5.8 Covariance and correlation

Suppose that the conditional distributions $p_{Y|X}(y \,|\, x)$ of a random variable $Y$ given different values $x$ of a random variable $X$ are not all the same, i.e. the conditional distribution of $Y$ 'depends on' the value of $X$.

Therefore, there is said to be an **association** (or **dependence**) between $X$ and $Y$.

If two random variables are associated (dependent), knowing the value of one (for example, $X$) will help to predict the likely value of the other (for example, $Y$).

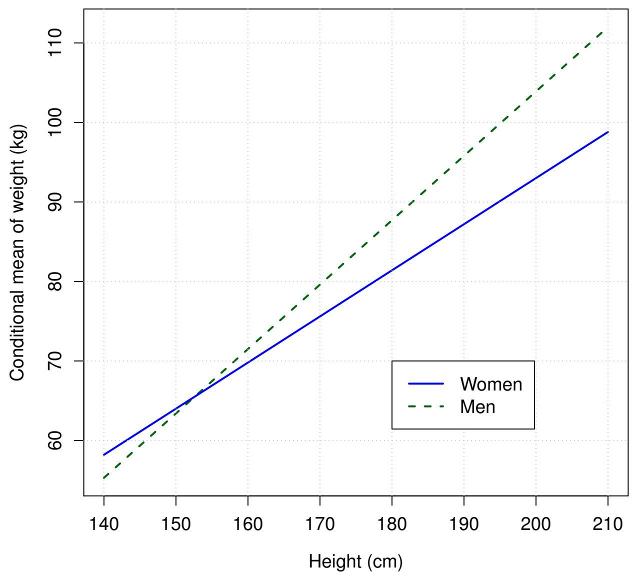


**Figure 5.5:** Conditional means for Example 5.9.

We next consider two **measures of association** which are used to summarise the *strength* of an association in a single number: covariance and correlation (scaled covariance).

## 5.8.1 Covariance

**Definition of covariance**

The **covariance** of two random variables $X$ and $Y$ is defined as:

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X) = \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))).$$

This can also be expressed as the more convenient formula:

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y).$$

This result will be proved later.

(Note that these involve expected values of products of two random variables, which have not been defined yet. We will do so later in this chapter.)

### Properties of covariance

Suppose $X$ and $Y$ are random variables, and $a$, $b$, $c$ and $d$ are constants.

- The covariance of a random variable with itself is the variance of the random variable:

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

- The covariance of a random variable and a constant is 0:

$$\mathbf{Cov}(a, X) = \mathbf{E}(aX) - \mathbf{E}(a)\,\mathbf{E}(X) = a\,\mathbf{E}(X) - a\,\mathbf{E}(X) = 0.$$

- The covariance of linear transformations of random variables is:

$$\mathbf{Cov}(aX + b, cY + d) = ac\,\mathbf{Cov}(X, Y).$$

### 5.8.2 Correlation

**Definition of correlation**

The **correlation** of two random variables $X$ and $Y$ is defined as:

$$\mathbf{Corr}(X, Y) = \mathbf{Corr}(Y, X) = \frac{\mathbf{Cov}(X, Y)}{\sqrt{\mathbf{Var}(X)\,\mathbf{Var}(Y)}} = \frac{\mathbf{Cov}(X, Y)}{\mathbf{sd}(X)\,\mathbf{sd}(Y)}.$$

When $\text{Cov}(X, Y) = 0$, then $\text{Corr}(X, Y) = 0$. When this is the case, we say that $X$ and $Y$ are **uncorrelated**.

Correlation and covariance are measures of the strength of the *linear* ('straight-line') association between $X$ and $Y$.

The further the correlation is from 0, the stronger is the linear association. The most extreme possible values of correlation are $-1$ and $+1$, which are obtained when $Y$ is an exact linear function of $X$.

- $\text{Corr}(X, Y) = +1$ when $Y = aX + b$ with $a > 0$.
- $\text{Corr}(X, Y) = -1$ when $Y = aX + b$ with $a < 0$.

If $\text{Corr}(X, Y) > 0$, we say that $X$ and $Y$ are **positively correlated**.

If $\text{Corr}(X, Y) < 0$, we say that $X$ and $Y$ are **negatively correlated**.

**Example 5.10** Recall the joint pf $p_{X,Y}(x, y)$ in the football example:

| | $Y = y$ | | | |
|---|---|---|---|---|
| $X = x$ | 0 | 1 | 2 | 3 |
| 0 | **0** | **0** | **0** | **0** |
| | 0.100 | 0.031 | 0.039 | 0.031 |
| 1 | **0** | **1** | **2** | **3** |
| | 0.100 | 0.146 | 0.092 | 0.015 |
| 2 | **0** | **2** | **4** | **6** |
| | 0.085 | 0.108 | 0.092 | 0.023 |
| 3 | **0** | **3** | **6** | **9** |
| | 0.062 | 0.031 | 0.039 | 0.006 |

Here, the numbers in bold are the values of $xy$ for each combination of $x$ and $y$. From these and their probabilities, we can derive the probability distribution of $XY$.

For example:

$$P(XY = 2) = p_{X,Y}(1,2) + p_{X,Y}(2,1) = 0.092 + 0.108 = 0.200.$$

The pf of the product $XY$ is:

| $XY = xy$ | 0 | 1 | 2 | 3 | 4 | 6 | 9 |
|---|---|---|---|---|---|---|---|
| $P(XY = xy)$ | 0.448 | 0.146 | 0.200 | 0.046 | 0.092 | 0.062 | 0.006 |

Hence:

$$\mathrm{E}(XY) = 0 \times 0.448 + 1 \times 0.146 + 2 \times 0.200 + \cdots + 9 \times 0.006 = 1.478.$$

From the marginal pfs $p_X(x)$ and $p_Y(y)$ we get:

$$\begin{aligned}
\mathrm{E}(X) &= 1.383 \\
\mathrm{E}(Y) &= 1.065 \\
\mathrm{E}(X^2) &= 2.827 \\
\mathrm{E}(Y^2) &= 2.039 \\
\mathrm{Var}(X) &= 2.827 - (1.383)^2 = 0.9143 \\
\mathrm{Var}(Y) &= 2.039 - (1.065)^2 = 0.9048.
\end{aligned}$$

Therefore, the covariance of $X$ and $Y$ is:

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = 1.478 - 1.383 \times 1.065 = 0.00511$$

and the correlation is:

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{0.00511}{\sqrt{0.9143 \times 0.9048}} = 0.00562.$$

The numbers of goals scored by the home and visiting teams are very nearly uncorrelated (i.e. not *linearly* associated).

### 5.8.3 Sample covariance and correlation

We have just introduced covariance and correlation, two new characteristics of probability distributions (*population* distributions). We now discuss their *sample* equivalents.

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a sample of $n$ pairs of observed values of two random variables $X$ and $Y$.

We can use these observations to calculate sample versions of the covariance and correlation between $X$ and $Y$. These are measures of association in the sample, i.e. descriptive statistics. They are also *estimates* of the corresponding population quantities $\mathrm{Cov}(X,Y)$ and $\mathrm{Corr}(X,Y)$. The uses of these sample measures will be discussed in more detail later in the course.

### Sample covariance

The **sample covariance** of random variables $X$ and $Y$ is calculated as:

$$\widehat{\text{Cov}(X,Y)} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of $X$ and $Y$, respectively.

### Sample correlation

The **sample correlation** of random variables $X$ and $Y$ is calculated as:

$$r = \frac{\widehat{\text{Cov}(X,Y)}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where $S_X$ and $S_Y$ are the sample standard deviations of $X$ and $Y$, respectively.

- $r$ is always between $-1$ and $+1$, and is equal to $-1$ or $+1$ only if $X$ and $Y$ are perfectly linearly related in the sample.
- $r = 0$ if $X$ and $Y$ are uncorrelated (not linearly related) in the sample.

**Example 5.11** Figure 5.6 shows different examples of scatterplots of observations of $X$ and $Y$, and different values of the sample correlation, $r$. The line shown in each plot is the best-fitting (least squares) line for the scatterplot (which will be introduced later in the course).

- In (a), $X$ and $Y$ are perfectly linearly related, and $r = 1$.
- Plots (b), (c) and (e) show relationships of different strengths.
- In (c), the variables are negatively correlated.
- In (d), there is no linear relationship, and $r = 0$.
- Plot (f) shows that $r$ can be 0 even if two variables are clearly related, if that relationship is not *linear*.

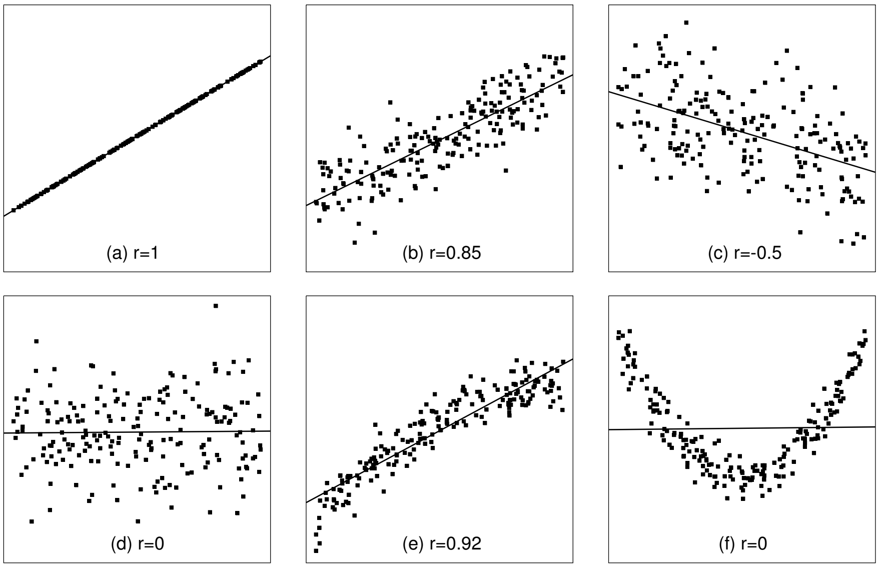


**Figure 5.6:** Scatterplots depicting various sample correlations as discussed in Example 5.11.

## 5.9 Independent random variables

Two discrete random variables $X$ and $Y$ are associated if $p_{Y|X}(y\,|\,x)$ depends on $x$. What if it does not, i.e. what if:

$$\boldsymbol{p_{Y|X}(y\,|\,x) = \frac{p_{X,Y}(x,y)}{p_X(x)} = p_Y(y) \quad \text{for all } x \text{ and } y}$$

so that knowing the value of $X$ does not help to predict $Y$?

This implies that:

$$\boldsymbol{p_{X,Y}(x,y) = p_X(x)\,p_Y(y) \quad \text{for all } x, y.} \tag{5.1}$$

$X$ and $Y$ are **independent** of each other if and only if (5.1) is true.

### Independent random variables

In general, suppose that $X_1, X_2, \ldots, X_n$ are discrete random variables. These are independent if and only if their joint pf is:

$$\boldsymbol{p(x_1, x_2, \ldots, x_n) = p_1(x_1)\,p_2(x_2)\cdots p_n(x_n)}$$

for all numbers $x_1, x_2, \ldots, x_n$, where $p_1(x_1), p_2(x_2), \ldots, p_n(x_n)$ are the univariate marginal pfs of $X_1, X_2, \ldots, X_n$, respectively.

Similarly, continuous random variables $X_1, X_2, \ldots, X_n$ are independent if and only if their joint pdf is:

$$\boldsymbol{f(x_1, x_2, \ldots, x_n) = f_1(x_1)\,f_2(x_2)\cdots f_n(x_n)}$$

for all $x_1, x_2, \ldots, x_n$, where $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$ are the univariate marginal pdfs of $X_1, X_2, \ldots, X_n$, respectively.

If two random variables are independent, they are also uncorrelated, i.e. we have:

$$\mathbf{Cov}(\boldsymbol{X},\boldsymbol{Y}) = \mathbf{0} \quad \textbf{and} \quad \mathbf{Corr}(\boldsymbol{X},\boldsymbol{Y}) = \mathbf{0}.$$

This will be proved later.

The reverse is not true, i.e. two random variables can be dependent even when their correlation is 0. This can happen when the dependence is non-linear.

> **Example 5.12** The football example is an instance of this. The conditional distributions $p_{Y|X}(y\,|\,x)$ are clearly not all the same, but the correlation is very nearly 0 (see Example 5.10).
>
> Another example is plot (f) in Figure 5.6, where the dependence is not linear, but quadratic.

### 5.9.1 Joint distribution of independent random variables

When random variables are independent, we can easily derive their joint pf or pdf as the product of their univariate marginal distributions. This is particularly simple if all the marginal distributions are the same.

> **Example 5.13** Suppose that $X_1, X_2, \ldots, X_n$ are independent, and each of them follows the Poisson distribution with the same mean $\lambda$. Therefore, the marginal pf of each $X_i$ is:
>
> $$p(x_i) = \frac{\mathrm{e}^{-\lambda}\lambda^{x_i}}{x_i!}$$
>
> and the joint pf of the random variables is:
>
> $$p(x_1, x_2, \ldots, x_n) = p(x_1)\,p(x_2)\cdots p(x_n) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} \frac{\mathrm{e}^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{\mathrm{e}^{-n\lambda}\lambda^{\sum_i x_i}}{\prod_i x_i!}.$$

> **Example 5.14** For a continuous example, suppose that $X_1, X_2, \ldots, X_n$ are independent, and each of them follows a normal distribution with the same mean $\mu$ and same variance $\sigma^2$. Therefore, the marginal pdf of each $X_i$ is:
>
> $$f(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$
>
> and the joint pdf of the variables is:
>
> $$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= f(x_1)\,f(x_2)\cdots f(x_n) = \prod_{i=1}^{n} f(x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right). \end{aligned}$$

## 5.10 Sums and products of random variables

Suppose $X_1, X_2, \ldots, X_n$ are random variables. We now go from the multivariate setting back to the univariate setting, by considering univariate *functions* of $X_1, X_2, \ldots, X_n$. In particular, we consider sums and products like:

$$\sum_{i=1}^{n} a_i X_i + b = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + b \tag{5.2}$$

and:

$$\prod_{i=1}^{n} a_i X_i = (a_1 X_1)(a_2 X_2) \cdots (a_n X_n)$$

where $a_1, a_2, \ldots, a_n$ and $b$ are constants.

Each such sum or product is itself a *univariate* random variable. The probability distribution of such a function depends on the joint distribution of $X_1, X_2, \ldots, X_n$.

**Example 5.15** In the football example, the sum $Z = X + Y$ is the total number of goals scored in a match.

Its probability function is obtained from the joint pf $p_{X,Y}(x, y)$, that is:

| $Z = z$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $p_Z(z)$ | 0.100 | 0.131 | 0.270 | 0.293 | 0.138 | 0.062 | 0.006 |

For example, $p_Z(1) = p_{X,Y}(0, 1) + p_{X,Y}(1, 0) = 0.031 + 0.100 = 0.131$. The mean of $Z$ is then $\text{E}(Z) = \sum_z z\, p_Z(z) = 2.448$.

Another example is the distribution of $XY$ (see Example 5.10).

However, what can we say about such distributions in general, in cases where we cannot derive them as easily?

### 5.10.1 Distributions of sums and products

General results for the distributions of sums and products of random variables are available as follows:

| | Sums | Products |
|---|---|---|
| Mean | Yes | Only for independent variables |
| Variance | Yes | No |
| Distributional form | Normal: Yes<br>Some other distributions: only for independent variables | No |

### 5.10.2 Expected values and variances of sums of random variables

We state, without proof, the following important result.

If $X_1, X_2, \ldots, X_n$ are random variables with means $\mathrm{E}(X_1), \mathrm{E}(X_2), \ldots, \mathrm{E}(X_n)$, respectively, and $a_1, a_2, \ldots, a_n$ and $b$ are constants, then:

$$\begin{aligned} \mathrm{E}\left(\sum_{i=1}^{n} a_i X_i + b\right) &= \mathrm{E}(a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + b) \\ &= a_1 \mathrm{E}(X_1) + a_2 \mathrm{E}(X_2) + \cdots + a_n \mathrm{E}(X_n) + b \\ &= \sum_{i=1}^{n} a_i \mathrm{E}(X_i) + b. \end{aligned} \tag{5.3}$$

Two simple special cases of this, when $n = 2$, are:

- $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$, obtained by choosing $X_1 = X$, $X_2 = Y$, $a_1 = a_2 = 1$ and $b = 0$
- $\mathrm{E}(X-Y) = \mathrm{E}(X) - \mathrm{E}(Y)$, obtained by choosing $X_1 = X$, $X_2 = Y$, $a_1 = 1$, $a_2 = -1$ and $b = 0$.

> **Example 5.16** In the football example, we have previously shown that $\mathrm{E}(X) = 1.383$, $\mathrm{E}(Y) = 1.065$ and $\mathrm{E}(X+Y) = 2.448$. So $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$, as the theorem claims.

If $X_1, X_2, \ldots, X_n$ are random variables with variances $\mathrm{Var}(X_1), \mathrm{Var}(X_2), \ldots, \mathrm{Var}(X_n)$, respectively, and covariances $\mathrm{Cov}(X_i, X_j)$ for $i \neq j$, and $a_1, a_2, \ldots, a_n$ and $b$ are constants, then:

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i + b\right) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i) + 2 \sum\sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j). \tag{5.4}$$

In particular, for $n = 2$:

- $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \times \mathrm{Cov}(X, Y)$
- $\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2 \times \mathrm{Cov}(X, Y)$.

If $X_1, X_2, \ldots, X_n$ are *independent* random variables, then $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$, and so (5.4) simplifies to:

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i). \tag{5.5}$$

In particular, for $n = 2$, when $X$ and $Y$ are independent:

- $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
- $\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

These results also hold whenever $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$, even if the random variables are not independent.

### 5.10.3 Expected values of products of independent random variables

If $X_1, X_2, \ldots, X_n$ are independent random variables and $a_1, a_2, \ldots, a_n$ are constants, then:

$$\boldsymbol{\mathrm{E}\left(\prod_{i=1}^{n} a_i X_i\right) = \mathrm{E}((a_1X_1)(a_2X_2)\cdots(a_nX_n)) = \prod_{i=1}^{n} a_i \mathrm{E}(X_i).}$$

In particular, when $X$ and $Y$ are independent:

$$\boldsymbol{\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y).}$$

There is no corresponding simple result for the means of products of *dependent* random variables. There is also no simple result for the *variances* of products of random variables, even when they are independent.

### 5.10.4 Some proofs of previous results

With these new results, we can now prove some results which were stated earlier.

Recall:

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

*Proof:*

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}((X - \mathrm{E}(X))^2) \\
&= \mathrm{E}(X^2 - 2\mathrm{E}(X)X + (\mathrm{E}(X))^2) \\
&= \mathrm{E}(X^2) - 2\,\mathrm{E}(X)\,\mathrm{E}(X) + (\mathrm{E}(X))^2 \\
&= \mathrm{E}(X^2) - 2(\mathrm{E}(X))^2 + (\mathrm{E}(X))^2 \\
&= \mathrm{E}(X^2) - (\mathrm{E}(X))^2
\end{aligned}$$

using (5.3), with $X_1 = X^2$, $X_2 = X$, $a_1 = 1$, $a_2 = -2\mathrm{E}(X)$ and $b = (\mathrm{E}(X))^2$.

■

Recall:

$$\boldsymbol{\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y).}$$

*Proof:*

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))) \\
&= \mathrm{E}(XY - \mathrm{E}(Y)X - \mathrm{E}(X)Y + \mathrm{E}(X)\,\mathrm{E}(Y)) \\
&= \mathrm{E}(XY) - \mathrm{E}(Y)\,\mathrm{E}(X) - \mathrm{E}(X)\,\mathrm{E}(Y) + \mathrm{E}(X)\,\mathrm{E}(Y) \\
&= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)
\end{aligned}$$

using (5.3), with $X_1 = XY$, $X_2 = X$, $X_3 = Y$, $a_1 = 1$, $a_2 = -\mathrm{E}(Y)$, $a_3 = -\mathrm{E}(X)$ and $b = \mathrm{E}(X)\,\mathrm{E}(Y)$.

■

Recall that if $X$ and $Y$ are independent, then:

$$\mathbf{Cov}(\boldsymbol{X},\boldsymbol{Y}) = \mathbf{Corr}(\boldsymbol{X},\boldsymbol{Y}) = \mathbf{0}.$$

*Proof:*

$$\mathbf{Cov}(\boldsymbol{X},\boldsymbol{Y}) = \mathbf{E}(\boldsymbol{XY}) - \mathbf{E}(\boldsymbol{X})\,\mathbf{E}(\boldsymbol{Y}) = \mathbf{E}(\boldsymbol{X})\,\mathbf{E}(\boldsymbol{Y}) - \mathbf{E}(\boldsymbol{X})\,\mathbf{E}(\boldsymbol{Y}) = \mathbf{0}$$

since $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$ when $X$ and $Y$ are independent.

Since $\mathrm{Corr}(X,Y) = \mathrm{Cov}(X,Y)/[\mathrm{sd}(X)\,\mathrm{sd}(Y)]$, $\mathrm{Corr}(X,Y) = 0$ whenever $\mathrm{Cov}(X,Y) = 0$.

■

### 5.10.5 Distributions of sums of random variables

We now know the *expected value* and *variance* of the sum:

$$\boldsymbol{a_1X_1 + a_2X_2 + \cdots + a_nX_n + b}$$

whatever the joint distribution of $X_1, X_2, \ldots, X_n$. This is usually *all* we can say about the distribution of this sum.

In particular, the *form* of the distribution of the sum (i.e. its pf/pdf) depends on the joint distribution of $X_1, X_2, \ldots, X_n$, and there are no simple general results about that.

For example, even if $X$ and $Y$ have distributions from the same family, the distribution of $X + Y$ is often not from that same family. However, such results are available for a few special cases.

#### Sums of independent binomial and Poisson random variables

Suppose $X_1, X_2, \ldots, X_n$ are random variables, and we consider the *unweighted* sum:

$$\boldsymbol{\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n.}$$

That is, the general sum given by (5.2), with $a_1 = a_2 = \cdots = a_n = 1$ and $b = 0$.

The following results hold when the random variables $X_1, X_2, \ldots, X_n$ are *independent*, but not otherwise.

- **If $\boldsymbol{X_i \sim \mathrm{Bin}(n_i, \pi)}$, then $\boldsymbol{\sum_i X_i \sim \mathrm{Bin}(\sum_i n_i, \pi)}$.**
- **If $\boldsymbol{X_i \sim \mathrm{Pois}(\lambda_i)}$, then $\boldsymbol{\sum_i X_i \sim \mathrm{Pois}(\sum_i \lambda_i)}$.**

### Application to the binomial distribution

An easy proof that the mean and variance of $X \sim \text{Bin}(n, \pi)$ are $\text{E}(X) = n\pi$ and $\text{Var}(X) = n\pi(1 - \pi)$ is as follows.

1. Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables, each distributed as $Z_i \sim \text{Bernoulli}(\pi) = \text{Bin}(1, \pi)$.

2. It is easy to show that $\text{E}(Z_i) = \pi$ and $\text{Var}(Z_i) = \pi(1 - \pi)$ for each $i = 1, 2, \ldots, n$ (see (4.3) and (4.4)).

3. Also, $\sum\limits_{i=1}^{n} Z_i = X \sim \text{Bin}(n, \pi)$ by the result above for sums of independent binomial random variables.

4. Therefore, using the results (5.2) and (5.5), we have:

$$\text{E}(X) = \sum_{i=1}^{n} \text{E}(Z_i) = n\pi \quad \text{and} \quad \text{Var}(X) = \sum_{i=1}^{n} \text{Var}(Z_i) = n\pi(1 - \pi).$$

### Sums of normally distributed random variables

*All* sums (linear combinations) of normally distributed random variables are also normally distributed.

Suppose $X_1, X_2, \ldots, X_n$ are normally distributed random variables, with $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots, n$, and $a_1, a_2, \ldots, a_n$ and $b$ are constants, then:

$$\boldsymbol{\sum_{i=1}^{n} a_i X_i + b \sim N(\mu, \sigma^2)}$$

where:

$$\boldsymbol{\mu = \sum_{i=1}^{n} a_i \mu_i + b \quad \text{and} \quad \sigma^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2 \sum\sum_{i<j} a_i a_j \text{Cov}(X_i, X_j).}$$

If the $X_i$s are independent (or just uncorrelated), i.e. if $\text{Cov}(X_i, X_j) = 0$ for all $i \neq j$, the variance simplifies to $\sigma^2 = \sum\limits_{i=1}^{n} a_i^2 \sigma_i^2$.

**Example 5.17** Suppose that in the population of English people aged 16 or over:

- the heights of men (in cm) follow a normal distribution with mean 174.9 and standard deviation 7.39

- the heights of women (in cm) follow a normal distribution with mean 161.3 and standard deviation 6.85.

Suppose we select one man and one woman at random and independently of each other. Denote the man's height by $X$ and the woman's height by $Y$. What is the probability that the man is at most 10 cm taller than the woman?

In other words, what is the probability that the *difference* between $X$ and $Y$ is at most 10?

Since $X$ and $Y$ are independent we have:

$$\begin{aligned} D = X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2) \\ &= N(174.9 - 161.3, (7.39)^2 + (6.85)^2) \\ &= N(13.6, (10.08)^2). \end{aligned}$$

The probability we need is:

$$\begin{aligned} P(D \leq 10) &= P\left(\frac{D - 13.6}{10.08} \leq \frac{10 - 13.6}{10.08}\right) \\ &= P(Z \leq -0.36) \\ &= P(Z \geq 0.36) \\ &= 0.3594 \end{aligned}$$

using Table 3 of Murdoch and Barnes' *Statistical Tables*.

The probability that a randomly selected man is at most 10 cm taller than a randomly selected woman is about 0.3594.

## 5.11 Overview of chapter

This chapter has introduced how to deal with more than one random variable at a time. Focusing mainly on discrete bivariate distributions, the relationships between joint, marginal and conditional distributions were explored. Sums and products of random variables concluded the chapter.

## 5.12 Key terms and concepts

- Association
- Bivariate
- Conditional distribution
- Conditional mean
- Conditional variance
- Correlation
- Covariance
- Dependence
- Independence
- Joint probability distribution
- Joint probability (density) function
- Marginal distribution
- Multivariate
- Random vector
- Uncorrelated
- Univariate

*Statistics are like bikinis. What they reveal is suggestive, but what they conceal is vital.*
(Aaron Levenstein)

# Appendix A
# Data visualisation and descriptive statistics

## A.1 (Re)vision of fundamentals

### Properties of the summation operator

Let $X_i$ and $Y_i$, for $i = 1, 2, \ldots, n$, be sets of $n$ numbers. Let $a$ denote a **constant**, i.e. a number with the same value for all $i$.

All of the following results follow simply from the properties of addition (if you are still not convinced, try them with $n = 3$).

1. $\sum\limits_{i=1}^{n} a = n \times a$.

   - *Proof:* $\sum\limits_{i=1}^{n} a = \overbrace{(a + a + \cdots + a)}^{n \text{ times}} = n \times a$.

■

2. $\sum\limits_{i=1}^{n} aX_i = a \sum\limits_{i=1}^{n} X_i$.

   - *Proof:* $\sum\limits_{i=1}^{n} aX_i = (aX_1 + aX_2 + \cdots + aX_n) = a(X_1 + X_2 + \cdots + X_n) = a \sum\limits_{i=1}^{n} X_i$.

■

3. $\sum\limits_{i=1}^{n} (X_i + Y_i) = \sum\limits_{i=1}^{n} X_i + \sum\limits_{i=1}^{n} Y_i$.

   - *Proof:* Rearranging the elements of the summation, we get:

$$\begin{aligned}\sum_{i=1}^{n} (X_i + Y_i) &= ((X_1 + Y_1) + (X_2 + Y_2) + \cdots + (X_n + Y_n)) \\ &= ((X_1 + X_2 + \cdots + X_n) + (Y_1 + Y_2 + \cdots + Y_n)) \\ &= (X_1 + X_2 + \cdots + X_n) + (Y_1 + Y_2 + \cdots + Y_n) \\ &= \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i.\end{aligned}$$

■

## Extension: double (triple etc.) summation

Sometimes sets of numbers may be indexed with two (or even more) subscripts, for example as $X_{ij}$, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$.

Summation over both indices is written as:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij} &= \sum_{i=1}^{n}(X_{i1} + X_{i2} + \cdots + X_{im}) \\ &= (X_{11} + X_{12} + \cdots + X_{1m}) + (X_{21} + X_{22} + \cdots + X_{2m}) \\ &\quad + \cdots + (X_{n1} + X_{n2} + \cdots + X_{nm}).\end{aligned}$$

The order of summation can be changed, that is:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij} = \sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}.$$

## Product notation

The analogous notation for the *product* of a set of numbers is:

$$\prod_{i=1}^{n} X_i = X_1 \times X_2 \times \cdots \times X_n.$$

The following results follow from the properties of multiplication.

1. $\prod_{i=1}^{n} aX_i = a^n \prod_{i=1}^{n} X_i$.

2. $\prod_{i=1}^{n} a = a^n$.

3. $\prod_{i=1}^{n} X_i Y_i = \left(\prod_{i=1}^{n} X_i\right)\left(\prod_{i=1}^{n} Y_i\right)$.

## The sum of deviations from the mean is 0

The mean is 'in the middle' of the observations $X_1, X_2, \ldots, X_n$, in the sense that positive and negative values of the **deviations** $X_i - \bar{X}$ cancel out, when summed over all the observations, that is:

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0.$$

*Proof*: (The proof uses the definition of $\bar{X}$ and the properties of summation introduced earlier. Note that $\bar{X}$ is a constant in the summation, because it has the same value for all $i$.)

$$\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X} = \sum_{i=1}^{n} X_i - n\bar{X} = \sum_{i=1}^{n} X_i - n\frac{\sum\limits_{i=1}^{n} X_i}{n}$$
$$= \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i = 0.$$

■

## The mean minimises the sum of squared deviations

The smallest possible value of the sum of squared deviations $\sum\limits_{i=1}^{n}(X_i - C)^2$, for any constant $C$, is obtained when $C = \bar{X}$.

*Proof*:

$$\begin{aligned}
\sum(X_i - C)^2 &= \sum(X_i \overbrace{-\bar{X} + \bar{X}}^{=0} - C)^2 \\
&= \sum((X_i - \bar{X}) + (\bar{X} - C))^2 \\
&= \sum((X_i - \bar{X})^2 + 2(X_i - \bar{X})(\bar{X} - C) + (\bar{X} - C)^2) \\
&= \sum(X_i - \bar{X})^2 + \sum 2(X_i - \bar{X})(\bar{X} - C) + \sum(\bar{X} - C)^2 \\
&= \sum(X_i - \bar{X})^2 + 2(\bar{X} - C)\overbrace{\sum(X_i - \bar{X})}^{=0} + n(\bar{X} - C)^2 \\
&= \sum(X_i - \bar{X})^2 + n(\bar{X} - C)^2 \\
&\geq \sum(X_i - \bar{X})^2
\end{aligned}$$

since $n(\bar{X} - C)^2 \geq 0$ for any choice of $C$. Equality is obtained only when $C = \bar{X}$, so that $n(\bar{X} - C)^2 = 0$.

■

## An alternative formula for the variance

The sum of squares in $S^2$ can also be expressed as:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2.$$

*Proof:* We have:

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \sum_{i=1}^{n}(X_i^2-2X_i\bar{X}+\bar{X}^2)\\ &= \sum_{i=1}^{n}X_i^2-2\bar{X}\overbrace{\sum_{i=1}^{n}X_i}^{=n\bar{X}}+\overbrace{\sum_{i=1}^{n}\bar{X}^2}^{=n\bar{X}^2}\\ &= \sum_{i=1}^{n}X_i^2-n\bar{X}^2.\end{aligned}$$

■

Therefore, the sample variance can also be calculated as:

$$S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$$

(and the standard deviation $S=\sqrt{S^2}$ again).

This formula is most convenient for calculations done by hand when summary statistics such as $\sum_i X_i$ and $\sum_i X_i^2$ are provided.

## Sample moment

**Sample moments** will be formally introduced in Chapter 7 (in Autumn term).

Let us define, for a variable $X$ and for each $k=1,2,\ldots$, the following:

- the $k$th sample moment about zero is:

$$m_k=\frac{\sum\limits_{i=1}^{n}X_i^k}{n}$$

- the $k$th central sample moment is:

$$m_k'=\frac{\sum\limits_{i=1}^{n}(X_i-\bar{X})^k}{n}.$$

In other words, these are sample averages of the powers $X_i^k$ and $(X_i-\bar{X})^k$, respectively.

Clearly:

$$\bar{X}=m_1 \quad \text{and} \quad S^2=\frac{n}{n-1}m_2'=\frac{1}{n-1}(nm_2-n(m_1)^2).$$

Moments of powers 3 and 4 are used in two more summary statistics which are described next, *for reference only*.

These are used much less often than measures of central tendency and dispersion.

### Sample skewness (non-examinable)

A measure of the **skewness** of the distibution of a variable $X$ is:

$$g_1 = \frac{m_3'}{s^3} = \frac{\sum_i (X_i - \bar{X})^3/n}{(\sum_i (X_i - \bar{X})^2/(n-1))^{3/2}}.$$

For this measure, $g_1 = 0$ for a symmetric distribution, $g_1 > 0$ for a positively-skewed distribution, and $g_1 < 0$ for a negatively-skewed distribution.

For example, $g_1 = 1.24$ for the (positively skewed) GDP per capita distribution shown in Chapter 1 of the main course notes, and $g_1 = 0.006$ for the (fairly symmetric) diastolic blood pressure distribution.

### Sample kurtosis (non-examinable)

**Kurtosis** refers to yet another characteristic of a sample distribution. This has to do with the relative sizes of the 'peak' and tails of the distribution (think about shapes of histograms).

- A distribution with high kurtosis (i.e. *leptokurtic*) has a sharp peak and a high proportion of observations in the tails far from the peak.
- A distribution with low kurtosis (i.e. *platykurtic*) is 'flat', with no pronounced peak with most of the observations spread evenly around the middle and weak tails.

A sample measure of kurtosis is:

$$g_2 = \frac{m_4'}{(m_2')^2} - 3 = \frac{\sum_i (X_i - \bar{X})^4/n}{(\sum_i (X_i - \bar{X})^2/n)^2} - 3.$$

$g_2 > 0$ for leptokurtic and $g_2 < 0$ for platykurtic distributions, and $g_2 = 0$ for the normal distribution (introduced in Chapter 4). Some software packages define a measure of kurtosis without the $-3$, i.e. 'excess kurtosis'.

### Calculation of sample quantiles (non-examinable)

This is how computer software calculates general sample quantiles (or how you can do so by hand, if you ever needed to).

Suppose we need to calculate the $c$th sample quantile, $q_c$, where $0 < c < 100$. Let $R = (n+1)c/100$, and define $r$ as the integer part of $R$ and $f = R - r$ as the fractional part (if $R$ is an integer, $r = R$ and $f = 0$). It follows that:

$$q_c = X_{(r)} + f(X_{(r+1)} - X_{(r)}) = (1-f)X_{(r)} + fX_{(r+1)}.$$

For example, if $n = 10$:

- for $q_{50}$ (the median): $R = 5.5$, $r = 5$, $f = 0.5$, and so we have:

  $$q_{50} = X_{(5)} + 0.5(X_{(6)} - X_{(5)}) = 0.5(X_{(5)} + X_{(6)})$$

  as before
- for $q_{25}$ (the first quartile): $R = 2.75$, $r = 2$, $f = 0.75$, and so:

  $$q_{25} = X_{(2)} + 0.75(X_{(3)} - X_{(2)}) = 0.25X_{(2)} + 0.75X_{(3)}.$$

## A.2 Worked example

1. Show that:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 = 2n\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right].$$

**Solution:**

Begin with the left-hand side and proceed as follows:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 = \sum_{i=1}^{n}\left[\sum_{j=1}^{n}(x_i - x_j)^2\right].$$

Now, expand the square:

$$= \sum_{i=1}^{n}\left[\sum_{j=1}^{n}(x_i^2 - 2x_ix_j + x_j^2)\right].$$

Next, sum separately inside [ ] so we have:

$$= \sum_{i=1}^{n}\left[\sum_{j=1}^{n}x_i^2 + \sum_{j=1}^{n}(-2x_ix_j) + \sum_{j=1}^{n}x_j^2\right].$$

Now, factor out $x_i$ terms inside [ ] to give:

$$= \sum_{i=1}^{n}\left[x_i^2\sum_{j=1}^{n}1 - 2x_i\sum_{j=1}^{n}x_j + \sum_{j=1}^{n}x_j^2\right].$$

Now, recall that $\bar{x} = \sum\limits_{i=1}^{n} x_i/n$, so re-write as:

$$= \sum_{i=1}^{n}\left[nx_i^2 - 2x_in\bar{x} + \sum_{j=1}^{n}x_j^2\right].$$

Next, expand the bracket:

$$= \sum_{i=1}^{n}nx_i^2 + \sum_{i=1}^{n}(-2x_in\bar{x}) + \sum_{i=1}^{n}\left(\sum_{j=1}^{n}x_j^2\right).$$

Re-arrange again:

$$= n\sum_{i=1}^{n}x_i^2 - 2n\bar{x}\sum_{i=1}^{n}x_i + \left(\sum_{j=1}^{n}x_j^2\right)\sum_{i=1}^{n}1.$$

Apply the '$\bar{x}$ trick' once more:

$$= n\sum_{i=1}^{n}x_i^2 - 2n\bar{x} \times n\bar{x} + \left(\sum_{j=1}^{n}x_j^2\right)n.$$

Factor out the common $n$ to give:

$$= n\left[\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + \sum_{j=1}^{n} x_j^2\right].$$

Without loss of generality, we can re-define the index $j$ as index $i$ so:

$$= n\left[\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + \sum_{i=1}^{n} x_i^2\right].$$

Finally, add terms, factor out $2n$, apply the '$\bar{x}$ trick' ...and you're done!

$$= 2n\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] = 2n\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right].$$

## A.3 Practice questions

Try to solve the questions before looking at the solutions – promise?! Solutions are located in Appendix F.

1. Let $Y_1$, $Y_2$ and $Y_3$ be real numbers with $\bar{Y} = (Y_1 + Y_2 + Y_3)/3$. Show that:

   (a) $\sum\limits_{j=1}^{3}(Y_j - \bar{Y}) = 0$

   (b) $\sum\limits_{j=1}^{3}\sum\limits_{k=1}^{3}(Y_j - \bar{Y})(Y_k - \bar{Y}) = 0$

   (c) $\sum\limits_{j=1}^{3} {}_{j\neq k} \sum\limits_{k=1}^{3}(Y_j - \bar{Y})(Y_k - \bar{Y}) = -\sum\limits_{j=1}^{3}(Y_j - \bar{Y})^2.$

   Hint: there are three terms in the expression of (a), nine terms in (b) and six terms in (c). Write out the terms, and try and find ways to simplify them which avoid the need for a lot of messy algebra!

2. For constants $a$ and $b$, show that:

   (a) $\bar{y} = a\bar{x} + b$, where $y_i = ax_i + b$ for $i = 1, 2, \ldots, n$

   (b) $\sum\limits_{i=1}^{n}(x_i - \bar{x})^2 = \sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2$

   (c) $\text{s.d.}_y = |a|\,\text{s.d.}_x$, where $\text{s.d.}_y$ is the standard deviation of $y$ etc.

   What are the mean and standard deviation of the set $\{x_1 + k, x_2 + k, \ldots, x_n + k\}$ where $k$ is a constant? What are the mean and standard deviation of the set $\{cx_1, cx_2, \ldots, cx_n\}$ where $c$ is a constant? Justify your answers with reference to the above results.

*The average human has one breast and one testicle.*
(Des McHale)

# Appendix B
# Probability theory

## B.1 Worked examples

1. $A$ and $B$ are independent events. Suppose that $P(A) = 2\pi$, $P(B) = \pi$ and $P(A \cup B) = 0.8$. Evaluate $\pi$.

   **Solution:**

   We have:

   $$\begin{aligned} P(A \cup B) = 0.8 &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A)\,P(B) \\ &= 2\pi + \pi - 2\pi^2. \end{aligned}$$

   Therefore:

   $$2\pi^2 - 3\pi + 0.8 = 0 \quad \Rightarrow \quad \pi = \frac{3 \pm \sqrt{9 - 6.4}}{4}.$$

   Hence $\pi = 0.346887$, since the other root is $> 1$!

2. $A$ and $B$ are events such that $P(A \mid B) > P(A)$. Prove that:

   $$P(A^c \mid B^c) > P(A^c)$$

   where $A^c$ and $B^c$ are the complements of $A$ and $B$, respectively, and $P(B^c) > 0$.

   **Solution:**

   From the definition of conditional probability:

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

   However:

   $$P(A \mid B) = \frac{P(A \cap B)}{P(B)} > P(A) \quad \text{i.e. } P(A \cap B) > P(A)\,P(B).$$

   Hence:

   $$P(A^c \mid B^c) > \frac{1 - P(A) - P(B) + P(A)\,P(B)}{1 - P(B)} = 1 - P(A) = P(A^c).$$

3. $A$, $B$ and $C$ are independent events. Prove that $A$ and $(B \cup C)$ are independent.

   **Solution:**

   Using the distributive law:

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

4. $A$ and $B$ are any two events in the sample space $S$. The binary set operator $\vee$ denotes an *exclusive union*, such that:

$$A \vee B = (A \cup B) \cap (A \cap B)^c = \{s \,|\, s \in A \text{ or } B, \text{ and } s \notin (A \cap B)\}.$$

   Show, from the axioms of probability, that:

   (a) $P(A \vee B) = P(A) + P(B) - 2 \times P(A \cap B)$

   (b) $P(A \vee B \,|\, A) = 1 - P(B \,|\, A)$.

   **Solution:**

   (a) We have:

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

   By axiom 3, noting that $(A \cap B^c)$ and $(B \cap A^c)$ are disjoint:

$$P(A \vee B) = P(A \cap B^c) + P(B \cap A^c).$$

   We can write $A = (A \cap B) \cup (A \cap B^c)$, hence (using axiom 3):

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

   Similarly, $P(B \cap A^c) = P(B) - P(A \cap B)$, hence:

$$P(A \vee B) = P(A) + P(B) - 2 \times P(A \cap B).$$

   (b) We have:

$$\begin{aligned} P(A \vee B \,|\, A) &= \frac{P((A \vee B) \cap A)}{P(A)} \\ &= \frac{P(A \cap B^c)}{P(A)} \\ &= \frac{P(A) - P(A \cap B)}{P(A)} \\ &= \frac{P(A)}{P(A)} - \frac{P(A \cap B)}{P(A)} \\ &= 1 - P(B \,|\, A). \end{aligned}$$

5. State and prove Bayes' theorem.

**Solution:**

Bayes' theorem is:

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

By definition:

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

If $\{B_i\}$, for $i = 1, 2, \ldots, K$, is a partition of the sample space $S$, then:

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

Hence the result.

6. A man has two bags. Bag A contains five keys and bag B contains seven keys. Only one of the twelve keys fits the lock which he is trying to open. The man selects a bag at random, picks out a key from the bag at random and tries that key in the lock. What is the probability that the key he has chosen fits the lock?

**Solution:**

Define a partition $\{C_i\}$, such that:

$$\begin{aligned}
C_1 &= \text{ key in bag A and bag A chosen} & \Rightarrow \quad & P(C_1) = \frac{5}{12} \times \frac{1}{2} = \frac{5}{24} \\
C_2 &= \text{ key in bag B and bag A chosen} & \Rightarrow \quad & P(C_2) = \frac{7}{12} \times \frac{1}{2} = \frac{7}{24} \\
C_3 &= \text{ key in bag A and bag B chosen} & \Rightarrow \quad & P(C_3) = \frac{5}{12} \times \frac{1}{2} = \frac{5}{24} \\
C_4 &= \text{ key in bag B and bag B chosen} & \Rightarrow \quad & P(C_4) = \frac{7}{12} \times \frac{1}{2} = \frac{7}{24}.
\end{aligned}$$

Hence we require, defining the event $F$ = 'key fits':

$$P(F) = \frac{1}{5} \times P(C_1) + \frac{1}{7} \times P(C_4) = \frac{1}{5} \times \frac{5}{24} + \frac{1}{7} \times \frac{7}{24} = \frac{1}{12}.$$

7. Continuing with Question 6, suppose the first key chosen does not fit the lock. What is the probability that the bag chosen:

(a) is bag A?

(b) contains the required key?

**Solution:**

(a) We require $P(\text{bag A} \,|\, F^c)$ which is:

$$P(\text{bag A} \,|\, F^c) = \frac{P(F^c \,|\, C_1)\, P(C_1) + P(F^c \,|\, C_2)\, P(C_2)}{\sum\limits_{i=1}^{4} P(F^c \,|\, C_i)\, P(C_i)}.$$

The conditional probabilities are:

$$P(F^c \,|\, C_1) = \frac{4}{5}, \quad P(F^c \,|\, C_2) = 1, \quad P(F^c \,|\, C_3) = 1 \quad \text{and} \quad P(F^c \,|\, C_4) = \frac{6}{7}.$$

Hence:

$$P(\text{bag A} \,|\, F^c) = \frac{4/5 \times 5/24 + 1 \times 7/24}{4/5 \times 5/24 + 1 \times 7/24 + 1 \times 5/24 + 6/7 \times 7/24} = \frac{1}{2}.$$

(b) We require $P(\text{right bag} \,|\, F^c)$ which is:

$$\begin{aligned} P(\text{right bag} \,|\, F^c) &= \frac{P(F^c \,|\, C_1)\, P(C_1) + P(F^c \,|\, C_4)\, P(C_4)}{\sum\limits_{i=1}^{4} P(F^c \,|\, C_i)\, P(C_i)} \\ &= \frac{4/5 \times 5/24 + 6/7 \times 7/24}{4/5 \times 5/24 + 1 \times 7/24 + 1 \times 5/24 + 6/7 \times 7/24} \\ &= \frac{5}{11}. \end{aligned}$$

8. Assume that a calculator has a 'random number' key and that when the key is pressed an integer between 0 and 999 inclusive is generated at random, all numbers being generated independently of one another.

(a) What is the probability that the number generated is less than 300?

(b) If two numbers are generated, what is the probability that both are less than 300?

(c) If two numbers are generated, what is the probability that the first number exceeds the second number?

(d) If two numbers are generated, what is the probability that the first number exceeds the second number, and their sum is exactly 300?

(e) If five numbers are generated, what is the probability that at least one number occurs more than once?

**Solution:**

(a) Simply $300/1{,}000 = 0.3$.

(b) Simply $0.3 \times 0.3 = 0.09$.

(c) Suppose $P(\text{first greater}) = x$, then by symmetry we have that $P(\text{second greater}) = x$. However, the probability that both are equal is (by counting):

$$\frac{\{0,0\},\{1,1\},\ldots,\{999,999\}}{1{,}000{,}000} = \frac{1{,}000}{1{,}000{,}000} = 0.001.$$

Hence $x + x + 0.001 = 1$, so $x = 0.4995$.

(d) The following cases apply $\{300,0\},\{299,1\},\ldots,\{151,149\}$, i.e. there are 150 possibilities from $(10)^6$. So the required probability is:

$$\frac{150}{1{,}000{,}000} = 0.00015.$$

(e) The probability that they are all different is:

$$\frac{999}{1{,}000} \times \frac{998}{1{,}000} \times \frac{997}{1{,}000} \times \frac{996}{1{,}000}.$$

Subtracting from 1 gives the required probability, i.e. 0.009965.

9. If $C_1, C_2, \ldots$ are events in $S$ which are pairwise mutually exclusive (i.e. $C_i \cap C_j = \emptyset$ for all $i \neq j$), then, by the axioms of probability:

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

Suppose that $A_1, A_2, \ldots$ are pairwise mutually exclusive events in $S$. Prove that a property like (B.1) also holds for conditional probabilities given some event $B$, i.e. prove that:

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

You can assume that all unions and intersections of $A_i$ and $B$ are also events in $S$.

**Solution:**

We have:

$$\begin{aligned} P\left(\left(\bigcup_{i=1}^{\infty} A_i\right) \mid B\right) &= \frac{P\left(\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B\right)}{P(B)} \\ &= \frac{P\left(\bigcup_{i=1}^{\infty} (A_i \cap B)\right)}{P(B)} \\ &= \sum_{i=1}^{\infty} \frac{P(A_i \cap B)}{P(B)} \\ &= \sum_{i=1}^{\infty} P(A_i \mid B) \end{aligned}$$

where the equation on the second line follows from (B.1) in the question, since $A_i \cap B$ are also events in $S$, and they are pairwise mutually exclusive (i.e. $(A_i \cap B \cap (A_j \cap B) = \emptyset$ for all $i \neq j$).

10. Suppose that three components numbered 1, 2 and 3 have probabilities of failure $\pi_1$, $\pi_2$ and $\pi_3$, respectively. Determine the probability of a system failure in each of the following cases where component failures are assumed to be independent.

    (a) Parallel system – the system fails if all components fail.

    (b) Series system – the system fails unless all components do *not* fail.

    (c) Mixed system – the system fails if component 1 fails or if both component 2 and component 3 fail.

    **Solution:**

    (a) Since the component failures are independent, the probability of system failure is $\pi_1\pi_2\pi_3$.

    (b) The probability that component $i$ does *not* fail is $1 - \pi_i$, hence the probability that the system does *not* fail is $(1 - \pi_1)(1 - \pi_2)(1 - \pi_3)$, and so the probability that the system fails is:

    $$1 - (1 - \pi_1)(1 - \pi_2)(1 - \pi_3).$$

    (c) Components 2 and 3 may be combined to form a notional component 4 with failure probability $\pi_2\pi_3$. So the system is equivalent to a component with failure probability $\pi_1$ and another component with failure probability $\pi_2\pi_3$, these being connected in series. Therefore, the failure probability is:

    $$1 - (1 - \pi_1)(1 - \pi_2\pi_3) = \pi_1 + \pi_2\pi_3 - \pi_1\pi_2\pi_3.$$

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

    **Solution:**

    Because there is no need to list the elementary outcome '1' twice. It is much clearer to write $S = \{1, 2\}$.

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

    **Solution:**

    The possible events are $\{a\}$, $\{b\}$, $\{c\}$, $\{a, b\}$, $\{a, c\}$, $\{b, c\}$, $\{a, b, c\}$ (the sample space $S$) and $\emptyset$.

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

    **Solution:**

    We have:

    $$A \cap S = A, \quad A \cup S = S, \quad A \cap \emptyset = \emptyset \quad \text{and} \quad A \cup \emptyset = A.$$

14. 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 $P(B \mid A)$.

**Solution:**

$S$ has 4 elementary outcomes which are equally likely, so each elementary outcome has probability $1/4$.

We have:

$$P(A \,|\, B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{c\})}{P(\{c, d\})} = \frac{1/4}{1/4 + 1/4} = \frac{1}{2}$$

and:

$$P(B \,|\, A) = \frac{P(B \cap A)}{P(A)} = \frac{P(\{c\})}{P(\{a, b, c\})} = \frac{1/4}{1/4 + 1/4 + 1/4} = \frac{1}{3}.$$

15. Suppose that we toss a fair coin twice. The sample space is given by $S = \{HH, HT, TH, TT\}$, 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.

**Solution:**

Note carefully here that we have equally likely elementary outcomes (due to the coin being fair), so that each has probability $1/4$, and the independence follows.

The event 'heads on the first toss' is $A = \{HH, HT\}$ and has probability $1/2$, because it is specified by two elementary outcomes. The event 'heads on the second toss' is $B = \{HH, TH\}$ and has probability $1/2$. The event 'heads on the first toss and the second toss' is $A \cap B = \{HH\}$ and has probability $1/4$. So the multiplication property $P(A \cap B) = 1/4 = 1/2 \times 1/2 = P(A)\,P(B)$ is satisfied, and the two events are independent.

16. Show that if $A$ and $B$ are disjoint events, and are also independent, then $P(A) = 0$ or $P(B) = 0$.[1]

**Solution:**

It is important to get the logical flow in the right direction here. We are told that $A$ and $B$ are disjoint events, that is:

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

So:

$$P(A \cap B) = 0.$$

We are also told that $A$ and $B$ are independent, that is:

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

It follows that:

$$0 = P(A)\,P(B)$$

and so either $P(A) = 0$ or $P(B) = 0$.

[1] Note that independence and disjointness are not similar ideas.

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

**Solution:**

Applying the product rule, we must have:

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

Therefore, since all subsets of two events from $A$, $B$ and $C$ must be independent, we must also have:

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

$$P(A \cap C) = P(A)\,P(C)$$

and:

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

One must check that all *four* conditions hold to verify independence of $A$, $B$ and $C$.

18. Prove the simplest version of Bayes' theorem from first principles.

**Solution:**

Applying the definition of conditional probability, we have:

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

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

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

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

**Solution:**

Here the random experiment is to choose a student at random, and to record whether the student passes ($P$) or fails ($F$), and whether the student has done their homework consistently ($C$) or has not ($N$).[2] The sample space is $S = \{PC, PN, FC, FN\}$. We use the events Pass $= \{PC, PN\}$, and Fail $= \{FC, FN\}$. We consider the sample space partitioned by Homework $= \{PC, FC\}$, and No Homework $= \{PN, FN\}$.

[2] Notice that $F = P^c$ and $N = C^c$.

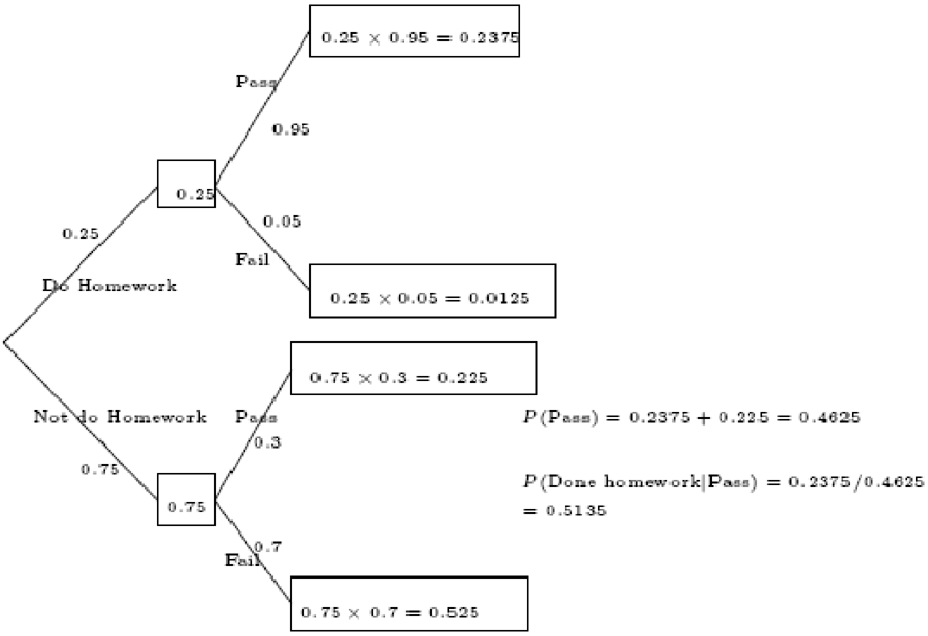
(a) The first part of the example asks for the denominator of Bayes' theorem:

$$\begin{aligned} P(\text{Pass}) &= P(\text{Pass} \,|\, \text{Homework})\, P(\text{Homework}) \\ &\quad + P(\text{Pass} \,|\, \text{No Homework})\, P(\text{No Homework}) \\ &= 0.95 \times 0.25 + 0.30 \times (1 - 0.25) \\ &= 0.2375 + 0.225 \\ &= 0.4625. \end{aligned}$$

(b) Now applying Bayes' theorem:

$$\begin{aligned} P(\text{Homework} \,|\, \text{Pass}) &= \frac{P(\text{Homework} \cap \text{Pass})}{P(\text{Pass})} \\ &= \frac{P(\text{Pass} \,|\, \text{Homework})\, P(\text{Homework})}{P(\text{Pass})} \\ &= \frac{0.95 \times 0.25}{0.4625} \\ &= 0.5135. \end{aligned}$$

Alternatively, we could arrange the calculations in a tree diagram as shown below.



20. Plagiarism is a serious problem for assessors of coursework. One check on plagiarism is to compare the coursework with a standard text. If the coursework has plagiarised the text, then there will be a 95% chance of finding exactly two phrases which are the same in both coursework and text, and a 5% chance of finding three or more phrases. 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. What is the probability that it has been plagiarised if it has exactly two phrases in the text?[3]

What if there are three or more phrases? Did you manage to get a roughly correct guess of these results before calculating?

**Solution:**

Suppose that two phrases are the same. We use Bayes' theorem:

$$P(\text{plagiarised} \,|\, \text{two the same}) = \frac{0.95 \times 0.05}{0.95 \times 0.05 + 0.5 \times 0.95} = 0.0909.$$

Finding two phrases has increased the chance the work is plagiarised from 5% to 9.1%. Did you get anywhere near 9% when guessing? Now suppose that we find three or more phrases:

$$P(\text{plagiarised} \,|\, \text{three or more the same}) = \frac{0.05 \times 0.05}{0.05 \times 0.05 + 0.5 \times 0.95} = 0.0052.$$

It seems that no plagiariser is silly enough to keep three or more phrases the same, so if we find three or more, the chance of the work being plagiarised falls from 5% to 0.5%! How close did you get by guessing?

21. $A$, $B$ and $C$ throw a die in that order until a six appears. The person who throws the first six wins. What are their respective chances of winning?

**Solution:**

We must assume that the game finishes with probability one (it would be proved in a more advanced subject). If $A$, $B$ and $C$ all throw and fail to get a six, then their respective chances of winning are as at the start of the game. We can call each completed set of three throws a round. Let us denote the probabilities of winning by $P(A)$, $P(B)$ and $P(C)$ for $A$, $B$ and $C$, respectively. Therefore:

$$\begin{aligned}
P(A) &= P(A \text{ wins on the 1st throw}) \\
&\quad + P(A \text{ wins in some round after the 1st round}) \\
&= \frac{1}{6} + P(A,\ B \text{ and } C \text{ fail on the 1st throw and } A \text{ wins after the 1st round}) \\
&= \frac{1}{6} + P(A,\ B \text{ and } C \text{ fail in the 1st round}) \\
&\quad \times P(A \text{ wins after the 1st round} \,|\, A,\ B \text{ and } C \text{ fail in the 1st round}) \\
&= \frac{1}{6} + P(\text{No six in first 3 throws})\, P(A) \\
&= \frac{1}{6} + \left(\frac{5}{6}\right)^3 P(A) \\
&= \frac{1}{6} + \left(\frac{125}{216}\right) P(A).
\end{aligned}$$

[3] Try making a guess before doing the calculation!

So $(1 - 125/216)P(A) = 1/6$, and $P(A) = 216/(91 \times 6) = 36/91$.

Similarly:

$$\begin{aligned}
P(B) &= P(B \text{ wins in the 1st round}) \\
&\qquad + P(B \text{ wins after the 1st round}) \\
&= P(A \text{ fails with the 1st throw and } B \text{ throws a six on the 1st throw}) \\
&\qquad + P(\text{All fail in the 1st round and } B \text{ wins after the 1st round}) \\
&= P(A \text{ fails with the 1st throw})\, P(B \text{ throws a six with the 1st throw}) \\
&\qquad + P(\text{All fail in the 1st round})\, P(B \text{ wins after the 1st} \,|\, \text{All fail in the 1st}) \\
&= \left(\frac{5}{6}\right)\left(\frac{1}{6}\right) + \left(\frac{5}{6}\right)^3 P(B).
\end{aligned}$$

So, $(1 - 125/216)P(B) = 5/36$, and $P(B) = 5(216)/(91 \times 36) = 30/91$.

In the same way, $P(C) = (5/6)(5/6)(1/6)(216/91) = 25/91$.

Notice that $P(A) + P(B) + P(C) = 1$. You may, on reflection, think that this rather long solution could be shortened, by considering the relative winning chances of $A$, $B$ and $C$.

22. 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.

**Solution:**

Suppose that the two players are $A$ and $B$. We calculate the probability that $A$ wins a three-, four- or five-set match, and then, since the players are evenly matched, double these probabilities for the final answer.

$$P(\text{'}A \text{ wins in 3 sets'}) = P(\text{'}A \text{ wins 1st set'} \cap \text{'}A \text{ wins 2nd set'} \cap \text{'}A \text{ wins 3rd set'}).$$

Since the sets are independent, we have:

$$\begin{aligned}
P(\text{'}A \text{ wins in 3 sets'}) &= P(\text{'}A \text{ wins 1st set'})\, P(\text{'}A \text{ wins 2nd set'})\, P(\text{'}A \text{ wins 3rd set'}) \\
&= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}.
\end{aligned}$$

Therefore, the total probability that the game lasts three sets is:

$$2 \times \frac{1}{8} = \frac{1}{4}.$$

If $A$ wins in four sets, the possible winning patterns are:

$$BAAA, \quad ABAA \quad \text{and} \quad AABA.$$

Each of these patterns has probability $(1/2)^4$ by using the same argument as in the case of 3 sets. So the probability that $A$ wins in four sets is $3 \times (1/16) = 3/16$. Therefore, the total probability of a match lasting four sets is $2 \times (3/16) = 3/8$.

The probability of a five-set match should be $1 - 3/8 - 1/4 = 3/8$, but let us check this directly. The winning patterns for $A$ in a five-set match are:

$$BBAAA, \quad BABAA, \quad BAABA, \quad ABBAA, \quad ABABA \quad \text{and} \quad AABBA.$$

Each of these has probability $(1/2)^5$ because of the independence of the sets. So the probability that $A$ wins in five sets is $6 \times (1/32) = 3/16$. Therefore, the total probability of a five-set match is $3/8$, as before.

## B.2 Practice questions

Try to solve the questions before looking at the solutions – promise?! Solutions are located in Appendix F.

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 chance of winning the game.

*There are lies, damned lies and statistics.*
(Mark Twain)

# Appendix C
# Random variables

## C.1 Worked examples

1. Toward the end of the financial year, James is considering whether to accept an offer to buy his stock option now, rather than wait until the normal exercise time. If he sells now, his profit will be £120,000. If he waits until the exercise time, his profit will be £200,000, provided that there is no crisis in the markets before that time; if there is a crisis, the option will be worthless and he would expect a net loss of £50,000. What action should he take to maximise his expected profit if the probability of crisis is:

   (a) 0.5?

   (b) 0.1?

   For what probability of a crisis would James be indifferent between the two courses of action if he wishes to maximise his expected profit?

   **Solution:**

   Let $\pi$ = probability of crisis, then:

   $$S = \text{E(profit given James sells)} = \text{£}120{,}000$$

   and:

   $$W = \text{E(profit given James waits)} = \text{£}200{,}000(1-\pi) + (-\text{£}50{,}000)\pi.$$

   (a) If $\pi = 0.5$, then $S = \text{£}120{,}000$ and $W = \text{£}75{,}000$, so $S > W$, hence James should sell now.

   (b) If $\pi = 0.1$, then $S = \text{£}120{,}000$ and $W = \text{£}175{,}000$, so $S < W$, hence James should wait until the exercise time.

   To be indifferent, we require $S = W$, i.e. we have:

   $$\text{£}200{,}000 - \text{£}250{,}000\,\pi = \text{£}120{,}000$$

   so $\pi = 8/25 = 0.32$.

2. Suppose the random variable $X$ has a *geometric* distribution with parameter $\pi$, which has the following probability function:

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

(a) Show that its moment generating function is:

$$\frac{\pi \mathrm{e}^t}{1-\mathrm{e}^t(1-\pi)}.$$

(b) Hence show that the mean of the distribution is $1/\pi$.

**Solution:**

(a) Working from the definition:

$$\begin{aligned} M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \sum_{x\in S} \mathrm{e}^{tx}\, p(x) &= \sum_{x=1}^{\infty} \mathrm{e}^{tx}\,(1-\pi)^{x-1}\pi \\ &= \sum_{x=1}^{\infty} \pi\mathrm{e}^t(\mathrm{e}^t(1-\pi))^{x-1} \\ &= \frac{\pi\mathrm{e}^t}{1-\mathrm{e}^t(1-\pi)} \end{aligned}$$

using the sum to infinity of a geometric series.

(b) Differentiating:

$$M_X'(t) = \frac{(1-\mathrm{e}^t(1-\pi))\pi\mathrm{e}^t + \pi\mathrm{e}^t(\mathrm{e}^t(1-\pi))}{(1-\mathrm{e}^t(1-\pi))^2} = \frac{\pi\mathrm{e}^t}{(1-\mathrm{e}^t(1-\pi))^2}.$$

Therefore:

$$\mathrm{E}(X) = M_X'(0) = \frac{\pi}{(1-(1-\pi))^2} = \frac{\pi}{\pi^2} = \frac{1}{\pi}.$$

3. A continuous random variable, $X$, has a probability density function, $f(x)$, defined by:

$$f(x) = \begin{cases} ax+bx^2 & \text{for } 0\le x\le 1 \\ 0 & \text{otherwise} \end{cases}$$

and $\mathrm{E}(X)=1/2$. Determine:

(a) the constants $a$ and $b$

(b) the cumulative distribution function, $F(x)$, of $X$

(c) the variance, $\mathrm{Var}(X)$.

**Solution:**

(a) We have:

$$\int_0^1 f(x)\,\mathrm{d}x = 1 \quad\Rightarrow\quad \int_0^1 ax+bx^2\,\mathrm{d}x = \left[\frac{ax^2}{2}+\frac{bx^3}{3}\right]_0^1 = 1$$

i.e. we have $a/2+b/3=1$.

Also, we know $\mathrm{E}(X) = 1/2$, hence:

$$\int_0^1 x\,(ax + bx^2)\,\mathrm{d}x = \left[\frac{ax^3}{3} + \frac{bx^4}{4}\right]_0^1 = \frac{1}{2}$$

i.e. we have:

$$\frac{a}{3} + \frac{b}{4} = \frac{1}{2} \quad \Rightarrow \quad a = 6 \quad \text{and} \quad b = -6.$$

Hence $f(x) = 6x(1-x)$ for $0 \leq x \leq 1$, and 0 otherwise.

(b) We have:

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 3x^2 - 2x^3 & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } x > 1. \end{cases}$$

(c) Finally:

$$\mathrm{E}(X^2) = \int_0^1 x^2\,(6x(1-x))\,\mathrm{d}x = \int_0^1 6x^3 - 6x^4\,\mathrm{d}x = \left[\frac{6x^4}{4} - \frac{6x^5}{5}\right]_0^1 = 0.3.$$

and so the variance is:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = 0.3 - 0.25 = 0.05.$$

4. The waiting time, $W$, of a traveller queueing at a taxi rank is distributed according to the cumulative distribution function, $G(w)$, defined by:

$$G(w) = \begin{cases} 0 & \text{for } w < 0 \\ 1 - (2/3)\exp(-w/2) & \text{for } 0 \leq w < 2 \\ 1 & \text{for } w \geq 2. \end{cases}$$

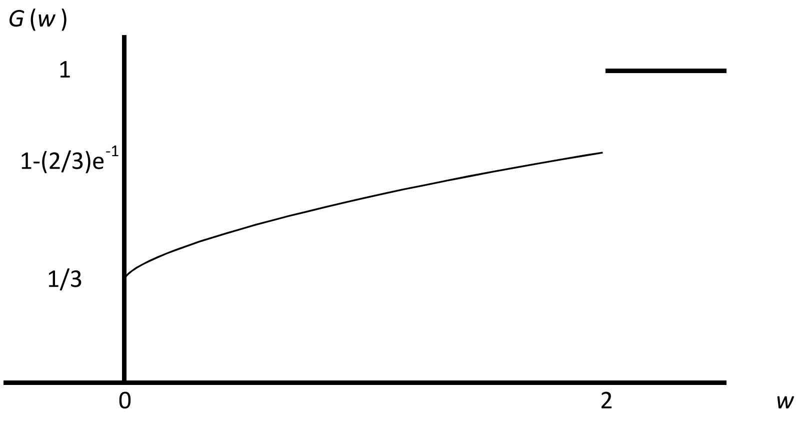
(a) Sketch the cumulative distribution function.

(b) Is the random variable $W$ discrete, continuous or mixed?

(c) Evaluate $P(W > 1)$, $P(W = 2)$, $P(W \leq 1.5 \,|\, W > 0.5)$ and $\mathrm{E}(W)$.

**Solution:**

(a) A sketch of the cumulative distribution function is:



(b) We see the distribution is mixed, with discrete 'atoms' at 0 and 2.

(c) We have:

$$P(W > 1) = 1 - G(1) = \frac{2}{3}\mathrm{e}^{-1/2}$$

$$P(W = 2) = \frac{2}{3}\mathrm{e}^{-1}$$

$$\begin{aligned}
P(W \leq 1.5 \,|\, W > 0.5) &= \frac{P(0.5 < W \leq 1.5)}{P(W > 0.5)} \\
&= \frac{G(1.5) - G(0.5)}{1 - G(0.5)} \\
&= \frac{(1 - (2/3)\mathrm{e}^{-1.5/2}) - (1 - (2/3)\mathrm{e}^{-0.5/2})}{(2/3)\mathrm{e}^{-0.5/2}} \\
&= 1 - \mathrm{e}^{-1/2}.
\end{aligned}$$

Finally, the mean is:

$$\begin{aligned}
\mathrm{E}(W) &= \frac{1}{3} \times 0 + \frac{2}{3}\mathrm{e}^{-1} \times 2 + \int_0^2 w\,\frac{1}{3}\mathrm{e}^{-w/2}\,\mathrm{d}w \\
&= \frac{4}{3}\mathrm{e}^{-1} + \left[\frac{w}{3}\frac{\mathrm{e}^{-w/2}}{-1/2}\right]_0^2 + \int_0^2 \frac{2}{3}\mathrm{e}^{-w/2}\,\mathrm{d}w \\
&= \frac{4}{3}\mathrm{e}^{-1} - \frac{4}{3}\mathrm{e}^{-1} + \left[\frac{2}{3}\frac{\mathrm{e}^{-w/2}}{-1/2}\right]_0^2 \\
&= \frac{4}{3}(1 - \mathrm{e}^{-1}).
\end{aligned}$$

5. A random variable $X$ has the following pdf:

$$f(x) = \begin{cases} 1/4 & \text{for } 0 \leq x \leq 1 \\ 3/4 & \text{for } 1 < x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Explain why $f(x)$ can serve as a pdf.

(b) Find the mean and median of the distribution.

(c) Find the variance, $\mathrm{Var}(X)$.

(d) Write down the cdf of $X$.

(e) Find $P(X = 1)$ and $P(X > 1.5 \,|\, X > 0.5)$.

(f) Derive the moment generating function of $X$.

**Solution:**

(a) Clearly, $f(x) \geq 0$ for all $x$ and $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$. This can be seen geometrically, since $f(x)$ defines two rectangles, one with base 1 and height 1/4, the other with base 1 and height 3/4, giving a total area of $1/4 + 3/4 = 1$.

(b) We have:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f(x)\,\mathrm{d}x = \int_0^1 \frac{x}{4}\,\mathrm{d}x + \int_1^2 \frac{3x}{4}\,\mathrm{d}x = \left[\frac{x^2}{8}\right]_0^1 + \left[\frac{3x^2}{8}\right]_1^2 = \frac{1}{8} + \frac{3}{2} - \frac{3}{8} = \frac{5}{4}.$$

The median is most simply found geometrically. The area to the right of the point $x = 4/3$ is 0.5, i.e. the rectangle with base $2 - 4/3 = 2/3$ and height 3/4, giving an area of $2/3 \times 3/4 = 1/2$. Hence the median is 4/3.

(c) For the variance, we proceed as follows:

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2\, f(x)\,\mathrm{d}x = \int_0^1 \frac{x^2}{4}\,\mathrm{d}x + \int_1^2 \frac{3x^2}{4}\,\mathrm{d}x = \left[\frac{x^3}{12}\right]_0^1 + \left[\frac{x^3}{4}\right]_1^2 = \frac{1}{12} + 2 - \frac{1}{4} = \frac{11}{6}.$$

Hence the variance is:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{11}{6} - \frac{25}{16} = \frac{88}{48} - \frac{75}{48} = \frac{13}{48} \approx 0.2708.$$

(d) The cdf is:

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x/4 & \text{for } 0 \leq x \leq 1 \\ 3x/4 - 1/2 & \text{for } 1 < x \leq 2 \\ 1 & \text{for } x > 2. \end{cases}$$

(e) $P(X = 1) = 0$, since the cdf is continuous, and:

$$\begin{aligned} P(X > 1.5 \,|\, X > 0.5) = \frac{P(\{X > 1.5\} \cap \{X > 0.5\})}{P(X > 0.5)} &= \frac{P(X > 1.5)}{P(X > 0.5)} \\ &= \frac{0.5 \times 0.75}{1 - 0.5 \times 0.25} \\ &= \frac{0.375}{0.875} \\ &= \frac{3}{7} \approx 0.4286. \end{aligned}$$

(f) The moment generating function is:

$$\begin{aligned} M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \int_{-\infty}^{\infty} \mathrm{e}^{tx}\, f(x)\,\mathrm{d}x &= \int_0^1 \frac{\mathrm{e}^{tx}}{4}\,\mathrm{d}x + \int_1^2 \frac{3\mathrm{e}^{tx}}{4}\,\mathrm{d}x \\ &= \left[\frac{\mathrm{e}^{tx}}{4t}\right]_0^1 + \left[\frac{3\mathrm{e}^{tx}}{4t}\right]_1^2 \\ &= \frac{1}{4t}(\mathrm{e}^t - 1) + \frac{3}{4t}(\mathrm{e}^{2t} - \mathrm{e}^t) \\ &= \frac{1}{4t}\left(3\mathrm{e}^{2t} - 2\mathrm{e}^t - 1\right). \end{aligned}$$

6. A continuous random variable $X$ has the following pdf:

$$f(x) = \begin{cases} x^3/4 & \text{for } 0 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Explain why $f(x)$ can serve as a pdf.

(b) Find the mean and mode of the distribution.

(c) Determine the cdf, $F(x)$, of $X$.

(d) Find the variance, $\mathrm{Var}(X)$.

(e) Find the skewness of $X$, given by:

$$\frac{\mathrm{E}((X - \mathrm{E}(X))^3)}{\sigma^3}.$$

(f) If a sample of five observations is drawn at random from the distribution, find the probability that all the observations exceed 1.5.

**Solution:**

(a) Clearly, $f(x) \geq 0$ for all $x$ and:

$$\int_0^2 \frac{x^3}{4}\,\mathrm{d}x = \left[\frac{x^4}{16}\right]_0^2 = 1.$$

(b) The mean is:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\,f(x)\,\mathrm{d}x = \int_0^2 \frac{x^4}{4}\,\mathrm{d}x = \left[\frac{x^5}{20}\right]_0^2 = \frac{32}{20} = 1.6$$

and the mode is 2 (where the density reaches a maximum).

(c) The cdf is:

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x^4/16 & \text{for } 0 \leq x \leq 2 \\ 1 & \text{for } x > 2. \end{cases}$$

(d) For the variance, we first find $\mathrm{E}(X^2)$, given by:

$$\mathrm{E}(X^2) = \int_0^2 x^2\,f(x)\,\mathrm{d}x = \int_0^2 \frac{x^5}{4}\,\mathrm{d}x = \left[\frac{x^6}{24}\right]_0^2 = \frac{64}{24} = \frac{8}{3}$$

$$\Rightarrow \quad \mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{8}{3} - \frac{64}{25} = \frac{8}{75} \approx 0.1067.$$

(e) The third moment about zero is:

$$\mathrm{E}(X^3) = \int_0^2 x^3\,f(x)\,\mathrm{d}x = \int_0^2 \frac{x^6}{4}\,\mathrm{d}x = \left[\frac{x^7}{28}\right]_0^2 = \frac{128}{28} \approx 4.5714.$$

Letting $\mathrm{E}(X) = \mu$, the numerator is:

$$\begin{aligned}\mathrm{E}((X - \mathrm{E}(X))^3) &= \mathrm{E}(X^3) - 3\mu\,\mathrm{E}(X^2) + 3\mu^2\mathrm{E}(X) - \mu^3 \\ &= 4.5714 - (3 \times 1.6 \times 2.6667) + (3 \times (1.6)^3) - (1.6)^3\end{aligned}$$

which is $-0.0368$, and the denominator is $(0.1067)^{3/2} = 0.0349$, hence the skewness is $-1.0544$.

(f) The probability of a single observation exceeding 1.5 is:

$$\int_{1.5}^{2} f(x)\,\mathrm{d}x = \int_{1.5}^{2} \frac{x^3}{4}\,\mathrm{d}x = \left[\frac{x^4}{16}\right]_{1.5}^{2} = 1 - 0.3164 = 0.6836.$$

So the probability of all five exceeding 1.5 is, by independence:

$$(0.6836)^5 = 0.1493.$$

7. Consider the function:

$$f(x) = \begin{cases} \lambda^2 x \mathrm{e}^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Show that this function has the characteristics of a probability density function.

(b) Evaluate $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

**Solution:**

(a) Clearly, $f(x) \geq 0$ for all $x$ since $\lambda^2 > 0$, $x \geq 0$ and $\mathrm{e}^{-\lambda x} \geq 0$.

To show, $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$, we have:

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x &= \int_{0}^{\infty} \lambda^2 x \mathrm{e}^{-\lambda x}\,\mathrm{d}x \\ &= \left[\lambda^2 x \frac{\mathrm{e}^{-\lambda x}}{-\lambda}\right]_0^{\infty} + \int_0^{\infty} \lambda^2 \frac{\mathrm{e}^{-\lambda x}}{\lambda}\,\mathrm{d}x \\ &= 0 + \int_0^{\infty} \lambda \mathrm{e}^{-\lambda x}\,\mathrm{d}x \\ &= 1 \quad (\text{provided } \lambda > 0).\end{aligned}$$

(b) For the mean:

$$\begin{aligned}\mathrm{E}(X) &= \int_0^{\infty} x\,\lambda^2 x \mathrm{e}^{-\lambda x}\,\mathrm{d}x \\ &= \Big[-x^2 \lambda \mathrm{e}^{-\lambda x}\Big]_0^{\infty} + \int_0^{\infty} 2x\lambda \mathrm{e}^{-\lambda x}\,\mathrm{d}x \\ &= 0 + \frac{2}{\lambda} \quad (\text{from the exponential distribution}).\end{aligned}$$

For the variance:

$$\mathrm{E}(X^2) = \int_0^{\infty} x^2 \lambda^2 x \mathrm{e}^{-\lambda x}\,\mathrm{d}x = \Big[-x^3 \lambda \mathrm{e}^{-\lambda x}\Big]_0^{\infty} + \int_0^{\infty} 3x^2 \lambda \mathrm{e}^{-\lambda x}\,\mathrm{d}x = \frac{6}{\lambda^2}.$$

So, $\mathrm{Var}(X) = 6/\lambda^2 - (2/\lambda)^2 = 2/\lambda^2$.

8. A random variable, $X$, has a cumulative distribution function, $F(x)$, defined by:

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

(a) Derive expressions for:

i. $P(X = 0)$

ii. $P(X = 1)$

iii. the pdf of $X$ (where it is continuous)

iv. $\mathrm{E}(X)$.

(b) Suppose that $\mathrm{E}(X) = 0.75(1 - \mathrm{e}^{-1})$. Evaluate the median of $X$ and $\mathrm{Var}(X)$.

**Solution:**

(a) We have:

i. $P(X = 0) = F(0) = 1 - a$.

ii. $P(X = 1) = \lim_{x \to 1}(F(1) - F(x)) = 1 - (1 - a\mathrm{e}^{-1}) = a\mathrm{e}^{-1}$.

iii. $f(x) = a\mathrm{e}^{-x}$, for $0 \leq x < 1$, and 0 otherwise.

iv. The mean is:

$$\begin{aligned}\mathrm{E}(X) &= 0 \times (1 - a) + 1 \times (a\mathrm{e}^{-1}) + \int_0^1 x\, a\mathrm{e}^{-x}\,\mathrm{d}x \\ &= a\mathrm{e}^{-1} + \Big[-xa\mathrm{e}^{-x}\Big]_0^1 + \int_0^1 a\mathrm{e}^{-x}\,\mathrm{d}x \\ &= a\mathrm{e}^{-1} - a\mathrm{e}^{-1} + \Big[-a\mathrm{e}^{-x}\Big]_0^1 \\ &= a(1 - \mathrm{e}^{-1}).\end{aligned}$$

(b) The median, $m$, satisfies:

$$F(m) = 0.5 = 1 - 0.75\mathrm{e}^{-m} \quad \Rightarrow \quad m = -\ln\left(\frac{2}{3}\right) = 0.4055.$$

Recall $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$, so:

$$\begin{aligned}\mathrm{E}(X^2) &= 0^2 \times (1 - a) + 1^2 \times (a\mathrm{e}^{-1}) + \int_0^1 x^2 a\mathrm{e}^{-x}\,\mathrm{d}x \\ &= a\mathrm{e}^{-1} + \Big[-x^2a\mathrm{e}^{-x}\Big]_0^1 + 2\int_0^1 xa\mathrm{e}^{-x}\,\mathrm{d}x \\ &= a\mathrm{e}^{-1} - a\mathrm{e}^{-1} + 2(a - 2a\mathrm{e}^{-1}) \\ &= 2a - 4a\mathrm{e}^{-1}.\end{aligned}$$

Hence:

$$\mathrm{Var}(X) = 2a - 4a\mathrm{e}^{-1} - a^2(1 + \mathrm{e}^{-2} - 2\mathrm{e}^{-1}) = 0.1716.$$

9. A continuous random variable, $X$, has a probability density function, $f(x)$, defined by:

$$f(x) = \begin{cases} k\sin(x) & \text{for } 0 \leq x \leq \pi \\ 0 & \text{otherwise.} \end{cases}$$

(a) Determine the constant $k$ and derive the cumulative distribution function, $F(x)$, of $X$.

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

**Solution:**

(a) We have:

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_0^{\pi} k\sin(x)\,\mathrm{d}x = 1.$$

Therefore:

$$\Big[k(-\cos(x))\Big]_0^{\pi} = 2k = 1 \quad \Rightarrow \quad k = \frac{1}{2}.$$

The cdf is hence:

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ (1-\cos(x))/2 & \text{for } 0 \leq x \leq \pi \\ 1 & \text{for } x > \pi. \end{cases}$$

(b) By symmetry, $\mathrm{E}(X) = \pi/2$. Alternatively:

$$\mathrm{E}(X) = \int_0^{\pi} \frac{1}{2}x\sin(x)\,\mathrm{d}x = \frac{1}{2}\Big[x(-\cos(x))\Big]_0^{\pi} + \int_0^{\pi}\frac{1}{2}\cos(x)\,\mathrm{d}x = \frac{\pi}{2} + \frac{1}{2}\Big[\sin(x)\Big]_0^{\pi} = \frac{\pi}{2}.$$

Next:

$$\begin{aligned} \mathrm{E}(X^2) = \int_0^{\pi} x^2\frac{1}{2}\sin(x)\,\mathrm{d}x &= \left[\frac{1}{2}x^2(-\cos(x))\right]_0^{\pi} + \int_0^{\pi} x\cos(x)\,\mathrm{d}x \\ &= \frac{\pi^2}{2} + \Big[x\sin(x)\Big]_0^{\pi} - \int_0^{\pi}\sin(x)\,\mathrm{d}x \\ &= \frac{\pi^2}{2} - \Big[-\cos(x)\Big]_0^{\pi} \\ &= \frac{\pi^2}{2} - 2. \end{aligned}$$

Therefore, the variance is:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{\pi^2}{2} - 2 - \frac{\pi^2}{4} = \frac{\pi^2}{4} - 2.$$

10. (a) Define the cumulative distribution function (cdf) of a random variable and state the principal properties of such a function.

(b) Identify which, if any, of the following functions could be a cdf under suitable choices of the constants $a$ and $b$. Explain why (or why not) each function satisfies the properties required of a cdf and the constraints which may be required in respect of the constants $a$ and $b$.

i. $F(x) = a(b-x)^2$ for $-1 \leq x \leq 1$.

ii. $F(x) = a(1-x^b)$ for $-1 \leq x \leq 1$.

iii. $F(x) = a - b\exp(-x/2)$ for $0 \leq x \leq 2$.

**Solution:**

(a) We defined the cdf to be $F(x) = P(X \leq x)$ where:

- $0 \leq F(x) \leq 1$
- $F(x)$ is non-decreasing
- $\mathrm{d}F(x)/\mathrm{d}x = f(x)$ and $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$ for continuous $X$
- $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$.

(b) i. Okay. $a = 0.25$ and $b = -1$.

ii. Not okay. At $x = 1$, $F(x) = 0$, which would mean a decreasing function.

iii. Okay. $a = b > 0$ and $b = (1 - \mathrm{e}^{-1})^{-1}$.

11. Suppose that random variable $X$ has the range $\{x_1, x_2, \ldots\}$, where $x_1 < x_2 < \cdots$. Prove the following results:

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

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

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

**Solution:**

The events $X = x_1$, $X = x_2$, $\ldots$ are disjoint, so we can write:

$$\sum_{i=1}^{\infty} p(x_i) = \sum_{i=1}^{\infty} P(X = x_i) = P(X = x_1 \cup X = x_2 \cup \cdots) = P(S) = 1.$$

In words, this result states that the sum of the probabilities of all the possible values $X$ can take is equal to 1.

For the second equation, we have:

$$F(x_k) = P(X \leq x_k) = P(X = x_k \cup X \leq x_{k-1}).$$

The two events on the right-hand side are disjoint, so:

$$F(x_k) = P(X = x_k) + P(X \leq x_{k-1}) = p(x_k) + F(x_{k-1})$$

which immediately gives the required result.

For the final result, we can write:

$$F(x_k) = P(X \leq x_k) = P(X = x_1 \cup X = x_2 \cup \cdots \cup X = x_k) = \sum_{i=1}^{k} p(x_i).$$

12. 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 for each of them to win the prize?

    **Solution:**

    Let $X$ denote the number on the winning ticket. Since all values between 1 and 100 are equally likely, $X$ has a discrete 'uniform' distribution such that:

    $$P(\text{'Carol wins'}) = P(X = 22) = p(22) = \frac{1}{100} = 0.01$$

    and:

    $$P(\text{'Janet wins'}) = P(X \leq 5) = F(5) = \frac{5}{100} = 0.05.$$

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

    **Solution:**

    We have $p(c) = 1$, so $X$ is effectively a constant, even though it is called a random variable. Its expectation is:

    $$\mathrm{E}(X) = \sum_{\forall x} x\, p(x) = cp(x) = cp(c) = c \times 1 = c. \tag{C.1}$$

    This is intuitively correct; on average, a constant must be equal to itself!

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

    **Solution:**

    We have:

    $$\mathrm{E}(X - \mathrm{E}(X)) = \mathrm{E}(X) - \mathrm{E}(\mathrm{E}(X))$$

    Since $\mathrm{E}(X)$ is just a number, as opposed to a random variable, (C.1) tells us that its expectation is equal to itself. Therefore, we can write:

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

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

**Solution:**

From the definition of variance, we have:

$$\text{Var}(X) = \text{E}((X-\mu)^2) = \sum_{\forall x} (x-\mu)^2 p(x) \geq 0$$

because the squared term $(x-\mu)^2$ is non-negative (as is $p(x)$). The only case where it is equal to 0 is when $x - \mu = 0$, that is, when $x = \mu$. Therefore, the random variable $X$ can only take the value $\mu$, and we have $p(\mu) = P(X = \mu) = 1$.

## C.2 Practice questions

Try to solve the questions before looking at the solutions – promise?! Solutions are located in Appendix F.

1. Construct suitable examples to show that for a random variable $X$:
   (a) $\text{E}(X^2) \neq (\text{E}(X))^2$ in general
   (b) $\text{E}(1/X) \neq 1/\text{E}(X)$ in general.

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

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

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

3. 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?

4. 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 which 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.

*The death of one man is a tragedy. The death of millions is a statistic.*
(Stalin to Churchill, Potsdam 1945)

# Appendix D
# Common distributions of random variables

## D.1 Worked examples

1. The random variable $X$ has a binomial distribution with parameters $n$ and $\pi$. Derive expressions for:

   (a) $\mathrm{E}(X)$

   (b) $\mathrm{E}(X(X-1))$

   (c) $\mathrm{E}(X(X-1)\cdots(X-r))$.

   **Solution:**

   (a) We have:

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

   (b) We have:

$$\begin{aligned}
\mathrm{E}(X(X-1)) &= \sum_{x=0}^{n} x(x-1)\binom{n}{x}\pi^x(1-\pi)^{n-x} \\
&= \sum_{x=2}^{n} x(x-1)\binom{n}{x}\pi^x(1-\pi)^{n-x}
\end{aligned}$$

$$\begin{aligned}
&= \sum_{x=2}^{n} \frac{n(n-1)(n-2)!}{(x-2)!\,((n-2)-(x-2))!}\pi^2\pi^{x-2}(1-\pi)^{n-x} \\
&= n(n-1)\pi^2 \sum_{x=2}^{n} \binom{n-2}{x-2}\pi^{x-2}(1-\pi)^{(n-2)-(x-2)} \\
&= n(n-1)\pi^2 \sum_{y=0}^{n-2} \binom{n-2}{y}\pi^{y}(1-\pi)^{(n-2)-y} \\
&= n(n-1)\pi^2.
\end{aligned}$$

(c) We have:

$$\begin{aligned}
&\mathrm{E}(X(X-1)\cdots(X-r)) \\
&= \sum_{x=0}^{n} x(x-1)\cdots(x-r)\binom{n}{x}\pi^x(1-\pi)^{n-x} \quad (\text{if } r<n) \\
&= \sum_{x=r+1}^{n} x(x-1)\cdots(x-r)\binom{n}{x}\pi^x(1-\pi)^{n-x} \\
&= n(n-1)\cdots(n-r)\pi^{r+1} \\
&\quad \times \sum_{x=r+1}^{n} \binom{n-(r+1)}{x-(r+1)}\pi^{x-(r+1)}(1-\pi)^{(n-(r+1))-(x-(r+1))} \\
&= n(n-1)\cdots(n-r)\pi^{r+1}.
\end{aligned}$$

2. Suppose $\{B_i\}$ is an infinite sequence of independent Bernoulli trials with:

$$P(B_i=0)=1-\pi \quad \text{and} \quad P(B_i=1)=\pi$$

for all $i$.

(a) Derive the distribution of $X_n = \sum\limits_{i=1}^{n} B_i$ and the expected value and variance of $X_n$.

(b) Let $Y = \min\{i : B_i = 1\}$. Derive the distribution of $Y$ and obtain an expression for $P(Y > y)$.

**Solution:**

(a) $X_n = \sum\limits_{i=1}^{n} B_i$ takes the values $0, 1, 2, \ldots, n$. Any sequence consisting of $x$ 1s and $n-x$ 0s has a probability $\pi^x(1-\pi)^{n-x}$ and gives a value $X_n = x$. There are $\binom{n}{x}$ such sequences, so:

$$P(X_n = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}$$

and 0 otherwise. Hence $\mathrm{E}(B_i) = \pi$ and $\mathrm{Var}(B_i) = \pi(1-\pi)$ which means $\mathrm{E}(X_n) = n\pi$ and $\mathrm{Var}(X_n) = n\pi(1-\pi)$.

(b) $Y = \min\{i : B_i = 1\}$ takes the values $1, 2, \ldots$, hence:

$$P(Y = y) = (1-\pi)^{y-1}\pi$$

and 0 otherwise. It follows that $P(Y > y) = (1-\pi)^y$.

3. A continuous random variable $X$ has the *gamma distribution*, denoted $X \sim \text{Gamma}(\alpha, \beta)$, if its probability density function (pdf) is of the form:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad \text{for } x > 0 \tag{D.1}$$

and 0 otherwise, where $\alpha > 0$ and $\beta > 0$ are parameters, and $\Gamma(\alpha)$ is the value of the *gamma function* such that:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx.$$

The gamma function has a finite value for all $\alpha > 0$. Two of its properties are that:

- $\Gamma(1) = 1$
- $\Gamma(\alpha) = (\alpha - 1)\,\Gamma(\alpha - 1)$ for all $\alpha > 1$.

(a) The function $f(x)$ defined by (1) satisfies all the conditions for being a pdf. Show that this implies the following result about an integral:

$$\int_0^\infty x^{\alpha-1} e^{-\beta x}\, dx = \frac{\Gamma(\alpha)}{\beta^\alpha} \quad \text{for any } \alpha > 0,\ \beta > 0.$$

(b) The $\text{Gamma}(1, \beta)$ distribution is the same as another distribution with a different name. What is this other distribution? Justify your answer.

(c) Show that if $X \sim \text{Gamma}(\alpha, \beta)$, the moment generating function of $X$ is:

$$M_X(t) = \left(\frac{\beta}{\beta - t}\right)^\alpha$$

which is defined when $t < \beta$.

(d) Suppose that $X \sim \text{Gamma}(\alpha, \beta)$. Derive the expected value of $X$:

i. using the pdf and the definition of the expected value

ii. using the moment generating function.

(e) If $X_1, X_2, \ldots, X_k$ are independent random variables such that $X_i \sim \text{Gamma}(\alpha_i, \beta)$ for $i = 1, 2, \ldots, k$, then:

$$\sum_{i=1}^k X_i \sim \text{Gamma}\left(\sum_{i=1}^k \alpha_i, \beta\right).$$

Using this result and the known properties of the exponential distribution, derive the expected value of $X \sim \text{Gamma}(\alpha, \beta)$ when $\alpha$ is a positive integer (i.e. $\alpha = 1, 2, \ldots$).

**Solution:**

(a) This follows immediately from the general property of pdfs that $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$, applied to the specific pdf here. We have:

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

(b) With $\alpha = 1$, the pdf becomes $f(x) = \beta\mathrm{e}^{-\beta x}$ for $x \geq 0$, and 0 otherwise. This is the pdf of the exponential distribution with parameter $\beta$, i.e. $X \sim \mathrm{Exp}(\beta)$.

(c) We have:

$$\begin{aligned} M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \int_0^\infty \mathrm{e}^{tx}\,f(x)\,\mathrm{d}x &= \int_0^\infty \mathrm{e}^{tx}\,\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-\beta x}\,\mathrm{d}x \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty \mathrm{e}^{tx}\,x^{\alpha-1}\mathrm{e}^{-\beta x}\,\mathrm{d}x \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}\mathrm{e}^{-(\beta-t)x}\,\mathrm{d}x \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \times \frac{\Gamma(\alpha)}{(\beta-t)^\alpha} \\ &= \left(\frac{\beta}{\beta-t}\right)^\alpha \end{aligned}$$

which is finite when $\beta - t > 0$, i.e. when $t < \beta$. The second-to-last step follows by substituting $\beta - t$ for $\beta$ in the result in (a).

(d) i. We have:

$$\begin{aligned} \mathrm{E}(X) = \int_{-\infty}^\infty x\,f(x)\,\mathrm{d}x &= \int_0^\infty x\,\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-\beta x}\,\mathrm{d}x \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{(\alpha+1)-1}\mathrm{e}^{-\beta x}\,\mathrm{d}x \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+1)}{\beta^{\alpha+1}} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\alpha\Gamma(\alpha)}{\beta^{\alpha+1}} \\ &= \frac{\alpha}{\beta} \end{aligned}$$

using (a) and the gamma function property stated in the question.

ii. The first derivative of $M_X(t)$ is:

$$M_X'(t) = \alpha\left(\frac{\beta}{\beta-t}\right)^{\alpha-1}\frac{\beta}{(\beta-t)^2}.$$

Therefore:

$$\mathrm{E}(X) = M_X'(0) = \frac{\alpha}{\beta}.$$

(e) When $\alpha$ is a positive integer, by the result stated in the question, we have $X = \sum_{i=1}^{\alpha} Y_i$, where $Y_1, Y_2, \ldots, Y_\alpha$ are independent random variables each distributed as Gamma$(1, \beta)$, i.e. as exponential with parameter $\beta$ as concluded in (b). The expected value of the exponential distribution can be taken as given from the lectures, so $\mathrm{E}(Y_i) = 1/\beta$ for each $i = 1, 2, \ldots, \alpha$. Therefore, using the general result on expected values of sums:

$$\mathrm{E}(X) = \mathrm{E}\left(\sum_{i=1}^{\alpha} Y_i\right) = \sum_{i=1}^{\alpha} \mathrm{E}(Y_i) = \alpha \times \frac{1}{\beta} = \frac{\alpha}{\beta}.$$

4. James enjoys playing Solitaire on his laptop. One day, he plays the game repeatedly. He has found, from experience, that the probability of success in any game is 1/3 and is independent of the outcomes of other games.

(a) What is the probability that his first success occurs in the fourth game he plays? What is the expected number of games he needs to play to achieve his first success?

(b) What is the probability of three successes in ten games? What is the expected number of successes in ten games?

(c) Use a suitable approximation to find the probability of less than 25 successes in 100 games. You should justify the use of the approximation.

(d) What is the probability that his third success occurs in the tenth game he plays?

**Solution:**

(a) $P(\text{first success in 4th game}) = (2/3)^3 \times (1/3) = 8/81 \approx 0.1$. This is a geometric distribution, for which $\mathrm{E}(X) = 1/\pi = 1/(1/3) = 3$.

(b) Use $X \sim \mathrm{Bin}(10, 1/3)$, such that $\mathrm{E}(X) = 10 \times 1/3 = 3.33$, and:

$$P(X = 3) = \binom{10}{3} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^7 \approx 0.2601.$$

(c) Approximate Bin$(100, 1/3)$ by:

$$N\left(100 \times \frac{1}{3}, 100 \times \frac{1}{3} \times \frac{2}{3}\right) = N\left(33.3, \frac{200}{9}\right).$$

The approximation seems reasonable since $n = 100$ is 'large', $\pi = 1/3$ is quite close to 0.5, $n\pi > 5$ and $n(1 - \pi) > 5$. Using a continuity correction:

$$P(X \le 24.5) = P\left(Z \le \frac{24.5 - 33.3}{\sqrt{200/9}}\right) = P(Z \le -1.87) \approx 0.0307.$$

(d) This is a negative binomial distribution (used for the trial number of the $k$th success) with a pf given by:

$$p(x) = \binom{x-1}{k-1} \pi^k (1 - \pi)^{x-k} \quad \text{for } x = k, k+1, k+2, \ldots$$

and 0 otherwise. Hence we require:

$$P(X=10)=\binom{9}{2}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^7\approx 0.0780.$$

Alternatively, you could calculate the probability of 2 successes in 9 trials, followed by a further success.

5. You may assume that 15% of individuals in a large population are left-handed.

   (a) If a random sample of 40 individuals is taken, find the probability that exactly 6 are left-handed.

   (b) If a random sample of 400 individuals is taken, find the probability that exactly 60 are left-handed by using a suitable approximation. Briefly discuss the appropriateness of the approximation.

   (c) What is the smallest possible size of a randomly chosen sample if we wish to be 99% sure of finding at least one left-handed individual in the sample?

**Solution:**

(a) Let $X \sim \text{Bin}(40, 0.15)$, hence:

$$P(X=6)=\binom{40}{6}\times(0.15)^6\times(0.85)^{34}=0.1742.$$

(b) Use a normal approximation with a continuity correction. We require $P(59.5 < X < 60.5)$, where $X \sim N(60, 51)$ since $X$ has mean $n\pi$ and variance $n\pi(1-\pi)$ with $n = 400$ and $\pi = 0.15$. Standardising, this is $2 \times P(0 < Z \leq 0.07) = 0.0558$, approximately.

Rules-of-thumb for use of the approximation are that $n$ is 'large', $\pi$ is close to 0.5, and $n\pi$ and $n(1-\pi)$ are both at least 5. The first and last of these definitely hold. There is some doubt whether a value of 0.15 can be considered close to 0.5, so use with caution!

(c) Given a sample of size $n$, $P(\text{no left-handers}) = (0.85)^n$. Therefore:

$$P(\text{at least 1 left-hander}) = 1-(0.85)^n.$$

We require $1-(0.85)^n > 0.99$, or $(0.85)^n < 0.01$. This gives:

$$100 < \left(\frac{1}{0.85}\right)^n$$

or:

$$n > \frac{\ln(100)}{\ln(1.1765)} = 28.34.$$

Rounding up, this gives a sample size of 29.

6. Show that the moment generating function (mgf) of a Poisson distribution with parameter $\lambda$ is given by:

$$M_X(t) = \exp(\lambda(\exp(t) - 1)), \quad \text{writing } \exp(\theta) \equiv \mathrm{e}^{\theta}.$$

Hence show that the mean and variance of the distribution are both $\lambda$.

**Solution:**

We have:

$$\begin{aligned} M_X(t) = \mathrm{E}(\exp(Xt)) &= \sum_{x=0}^{\infty} \exp(xt)\exp(-\lambda)\frac{\lambda^x}{x!} \\ &= \sum_{x=0}^{\infty} (\lambda\exp(t))^x \frac{\exp(-\lambda)}{x!} \\ &= \exp(-\lambda)\sum_{x=0}^{\infty} \frac{(\lambda\exp(t))^x}{x!} \\ &= \exp(-\lambda)\exp(\lambda\exp(t)) \\ &= \exp(\lambda(\exp(t) - 1)). \end{aligned}$$

We have that $M_X(0) = \exp(0) = 1$. Now, taking logs:

$$\ln M_X(t) = \lambda(\exp(t) - 1).$$

Now differentiate:

$$\frac{M_X'(t)}{M_X(t)} = \lambda\exp(t) \quad \Rightarrow \quad M_X'(t) = M_X(t)\lambda\exp(t).$$

Differentiating again, we get:

$$M_X''(t) = M_X'(t)\lambda\exp(t) + M_X(t)\lambda\exp(t).$$

We note $\mathrm{E}(X) = M_X'(0) = M_X(0)\lambda\exp(0) = \lambda$, also:

$$\mathrm{Var}(X) = M_X''(0) - (M_X'(0))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

7. In javelin throwing competitions, the throws of athlete A are normally distributed. It has been found that 15% of her throws exceed 43 metres, while 3% exceed 45 metres. What distance will be exceeded by 90% of her throws?

**Solution:**

Suppose $X \sim N(\mu, \sigma^2)$ is the random variable for throws. $P(X > 43) = 0.15$ leads to $\mu = 43 - 1.035 \times \sigma$ (using Table 3 of Murdoch and Barnes' *Statistical Tables*).

Similarly, $P(X > 45) = 0.03$ leads to $\mu = 45 - 1.88 \times \sigma$. Solving yields $\mu = 40.55$ and $\sigma = 2.367$, hence $X \sim N(40.55, (2.367)^2)$. So:

$$P(X > x) = 0.90 \quad \Rightarrow \quad \frac{x - 40.55}{2.367} = -1.28.$$

Hence $x = 37.52$ metres.

8. People entering an art gallery are counted by the attendant at the door. Assume that people arrive in accordance with a Poisson distribution, with one person arriving every 2 minutes. The attendant leaves the door unattended for 5 minutes.

   (a) Calculate the probability that:

      i. nobody will enter the gallery in this time

      ii. 3 or more people will enter the gallery in this time.

   (b) Find, to the nearest second, the length of time for which the attendant could leave the door unattended for there to be a probability of 0.90 of no arrivals in that time.

   (c) Comment briefly on the assumption of a Poisson distribution in this context.

   **Solution:**

   (a) $\lambda = 1$ for a two-minute interval, so $\lambda = 2.5$ for a five-minute interval. Therefore:

   $$P(\text{no arrivals}) = \mathrm{e}^{-2.5} = 0.0821$$

   and:

   $$P(\geq 3 \text{ arrivals}) = 1 - p_X(0) - p_X(1) - p_X(2) = 1 - \mathrm{e}^{-2.5}(1+2.5+3.125) = 0.4562.$$

   (b) For an interval of $N$ minutes, the parameter is $N/2$. We need $p(0) = 0.90$, so $\mathrm{e}^{-N/2} = 0.90$ giving $N/2 = -\ln(0.90)$ and $N = 0.21$ minutes, or 13 seconds.

   (c) The rate is unlikely to be constant: more people at lunchtimes or early evenings etc. Likely to be several arrivals in a small period – couples, groups etc. Quite unlikely the Poisson will provide a good model.

9. The random variable $Y$, representing the life-span of an electronic component, is distributed according to a probability density function $f(y)$, where $y > 0$. The *survivor function*, $\Im$, is defined as $\Im(y) = P(Y > y)$ and the *age-specific failure rate*, $\phi(y)$, is defined as $f(y)/\Im(y)$. Suppose $f(y) = \lambda \mathrm{e}^{-\lambda y}$, i.e. $Y \sim \text{Exp}(\lambda)$.

   (a) Derive expressions for $\Im(y)$ and $\phi(y)$.

   (b) Comment briefly on the implications of the *age-specific failure rate* you have derived in the context of the exponentially-distributed component life-spans.

   **Solution:**

   (a) The survivor function is:

   $$\Im(y) = P(Y > y) = \int_y^\infty \lambda \mathrm{e}^{-\lambda x}\,\mathrm{d}x = \Big[-\mathrm{e}^{-\lambda x}\Big]_y^\infty = \mathrm{e}^{-\lambda y}.$$

   The age-specific failure rate is:

   $$\phi(y) = \frac{f(y)}{\Im(y)} = \frac{\lambda \mathrm{e}^{-\lambda y}}{\mathrm{e}^{-\lambda y}} = \lambda.$$

   (b) The age-specific failure rate is constant, indicating it does not vary with age. This is unlikely to be true in practice!

10. For the binomial distribution with a probability of success of 0.25 in an individual trial, calculate the probability that, in 50 trials, there are at least 8 successes:

    (a) using the normal approximation *without* a continuity correction

    (b) using the normal approximation *with* a continuity correction.

    Compare these results with the exact probability of 0.9547 and comment.

    **Solution:**

    We seek $P(X \geq 8)$ using the normal approximation $Y \sim N(12.5, 9.375)$.

    (a) So, *without* a continuity correction:

    $$P(Y \geq 8) = P\left(Z \geq \frac{8 - 12.5}{\sqrt{9.375}}\right) = P(Z \geq -1.47) = 0.9292.$$

    The required probability could have been expressed as $P(X > 7)$, or indeed any number in $[7, 8)$, for example:

    $$P(Y > 7) = P\left(Z \geq \frac{7 - 12.5}{\sqrt{9.375}}\right) = P(Z \geq -1.80) = 0.9641.$$

    (b) *With* a continuity correction:

    $$P(Y > 7.5) = P\left(Z \geq \frac{7.5 - 12.5}{\sqrt{9.375}}\right) = P(Z \geq -1.63) = 0.9484.$$

    Compared to 0.9547, using the continuity correction yields the closer approximation.

11. 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 and which are new. 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?

    **Solution:**

    For an orange chosen at random, the event ‘mouldy’ is the union of the disjoint events ‘mouldy’ ∩ ‘new’ and ‘mouldy’ ∩ ‘old’. So:

    $$\begin{aligned} P(\text{‘mouldy’}) &= P(\text{‘mouldy’} \cap \text{‘new’}) + P(\text{‘mouldy’} \cap \text{‘old’}) \\ &= P(\text{‘mouldy’} \mid \text{‘new’})\, P(\text{‘new’}) + P(\text{‘mouldy’} \mid \text{‘old’})\, P(\text{‘old’}) \\ &= 0.1 \times 0.5 + 0.2 \times 0.5 \\ &= 0.15. \end{aligned}$$

    As the pile of oranges is very large, we can assume that the results for the five oranges will be independent, so we have 5 independent trials each with probability of ‘mouldy’ equal to 0.15. The distribution of the number of mouldy oranges will be a binomial distribution with $n = 5$ and $\pi = 0.15$.

12. Underground trains on the Northern line have a probability 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?

**Solution:**

The probability of no breakdown on one journey is $\pi = 1 - 0.05 = 0.95$, so the number of journeys without a breakdown, $X$, has a $\text{Bin}(10, 0.95)$ distribution. We want $P(X > 8)$, which is:

$$\begin{aligned}P(X > 8) &= p(9) + p(10)\\ &= \binom{10}{9} \times (0.95)^9 \times (0.05)^1 + \binom{10}{10} \times (0.95)^{10} \times (0.05)^0\\ &= 0.3151 + 0.5987\\ &= 0.9138.\end{aligned}$$

13. 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:

(a) 10 animals are injected; all 10 remain free from infection

(b) 17 animals are injected; more than 15 remain free from infection and there are 2 doubtful cases

(c) 23 animals are infected; more than 20 remain free from infection and there are three doubtful cases.

Which experiment provides the strongest evidence in favour of the serum?

**Solution:**

These experiments involve tests on different cattle, which one might expect to behave independently of one another. The probability of infection without injection with the serum might also reasonably be assumed to be the same for all cattle. So the distribution which we need here is the binomial distribution. If the serum has no effect, then the probability of infection for each of the cattle is 0.25.

One way to assess the evidence of the three experiments is to calculate the probability of the result of the experiment if the serum had no effect at all. If it has an effect, then one would expect larger numbers of cattle to remain free from infection, so the experimental results as given do provide some clue as to whether the serum has an effect, in spite of their incompleteness.

Let $X_{(n)}$ be the number of cattle infected, out of a sample of $n$. We are assuming that $X_{(n)} \sim \text{Bin}(n, 0.25)$.

(a) With 10 trials, the probability of 0 infected if the serum has no effect is:

$$P(X_{(10)} = 0) = \binom{10}{0} \times (0.75)^{10} = (0.75)^{10} = 0.0563.$$

(b) With 17 trials, the probability of more than 15 remaining uninfected if the serum has no effect is:

$$\begin{aligned} P(X_{(17)} < 2) &= P(X_{(17)} = 0) + P(X_{(17)} = 1) \\ &= \binom{17}{0} \times (0.75)^{17} + \binom{17}{1} \times (0.25)^1 \times (0.75)^{16} \\ &= (0.75)^{17} + 17 \times (0.25)^1 \times (0.75)^{16} \\ &= 0.0075 + 0.0426 \\ &= 0.0501. \end{aligned}$$

(c) With 23 trials, the probability of more than 20 remaining free from infection if the serum has no effect is:

$$\begin{aligned} P(X_{(23)} < 3) &= P(X_{(23)} = 0) + P(X_{(23)} = 1) + P(X_{(23)} = 2) \\ &= \binom{23}{0} \times (0.75)^{23} + \binom{23}{1} \times (0.25)^1 \times (0.75)^{22} \\ &\quad + \binom{23}{2} \times (0.25)^2 \times (0.75)^{21} \\ &= 0.7523 + 23 \times 0.25 \times (0.75)^{22} + \frac{23 \times 22}{2} \times (0.25)^2 \times (0.75)^{21} \\ &= 0.0013 + 0.0103 + 0.0376 \\ &= 0.0492. \end{aligned}$$

The most surprising-looking event in these three experiments is that of experiment 3, and so we can say that this experiment offered the most support for the use of the serum.

14. 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) What is the chance that there will be two or more accidents in a given week?

(c) If James goes to work there for a four-week period, what is the probability that no accidents occur while he is there?

**Solution:**

Here we have counts of random events over time, which is a typical application for the Poisson distribution. We are assuming that accidents are equally likely to occur at any time and are independent. The mean for the Poisson distribution is 0.5 per day.

Let $X$ be the number of accidents in a week. The probability of exactly two accidents in a given week is found by using the parameter $\lambda = 5 \times 0.5 = 2.5$ (5 working days a week assumed).

(a) The probability of exactly two accidents in a week is:

$$p(2) = \frac{e^{-2.5}(2.5)^2}{2!} = 0.2565.$$

(b) The probability of two or more accidents in a given week is:

$$P(X \geq 2) = 1 - p(0) - p(1) = 0.7127.$$

(c) If James goes to the industrial plant and does not change the probability of an accident simply by being there (he might bring bad luck, or be superbly safety-conscious!), then over 4 weeks there are 20 working days, and the probability of no accident comes from a Poisson random variable with mean 10. If $Y$ is the number of accidents while James is there, the probability of no accidents is:

$$p_Y(0) = \frac{e^{-10}(10)^0}{0!} = 0.0000454.$$

James is very likely to be there when there is an accident!

15. The chance that a lottery ticket has a winning number is 0.0000001.

(a) If 10,000,000 people buy tickets which are independently numbered, what is the probability there is no winner?

(b) What is the probability that there is exactly 1 winner?

(c) What is the probability that there are exactly 2 winners?

**Solution:**

The number of winning tickets, $X$, will be distributed as:

$$X \sim \text{Bin}(10{,}000{,}000, 0.0000001).$$

Since $n$ is large and $\pi$ is small, the Poisson distribution should provide a good approximation. The Poisson parameter is:

$$\lambda = n\pi = 10{,}000{,}000 \times 0.0000001 = 1$$

and so we set $X \sim \text{Pois}(1)$. We have:

$$p(0) = \frac{e^{-1}1^0}{0!} = 0.3679, \quad p(1) = \frac{e^{-1}1^1}{1!} = 0.3679 \quad \text{and} \quad p(2) = \frac{e^{-1}1^2}{2!} = 0.1839.$$

Using the exact binomial distribution of $X$, the results are:

$$p(0) = \binom{(10)^7}{0} \times ((10)^{-7})^0 \times (1 - (10)^{-7})^{(10)^7} = 0.3679$$

$$p(1) = \binom{(10)^7}{1} \times ((10)^{-7})^1 \times (1 - (10)^{-7})^{(10)^7 - 1} = 0.3679$$

$$p(2) = \binom{(10)^7}{2} \times ((10)^{-7})^2 \times (1 - (10)^{-7})^{(10)^7 - 2} = 0.1839.$$

Notice that, in this case, the Poisson approximation is correct to at least 4 decimal places.

16. Suppose that $X \sim \text{Uniform}[0,\ 1]$. Compute $P(X > 0.2)$, $P(X \geq 0.2)$ and $P(X^2 > 0.04)$.

**Solution:**

We have $a = 0$ and $b = 1$, and can use the formula for $P(c < X \leq d)$, for constants $c$ and $d$. Hence:

$$P(X > 0.2) = P(0.2 < X \leq 1) = \frac{1 - 0.2}{1 - 0} = 0.8.$$

Also:

$$P(X \geq 0.2) = P(X = 0.2) + P(X > 0.2) = 0 + P(X > 0.2) = 0.8.$$

Finally:

$$P(X^2 > 0.04) = P(X < -0.2) + P(X > 0.2) = 0 + P(X > 0.2) = 0.8.$$

17. 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?

**Solution:**

The distribution of $X$ is Exp(1/3), so the probability is:

$$P(X > 4) = 1 - F(4) = 1 - (1 - \mathrm{e}^{-(1/3)\times 4}) = 1 - 0.7364 = 0.2636.$$

18. 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.

**Solution:**

The values of interest are 169 and 190. The corresponding $z$-values are:

$$z_1 = \frac{169 - 175}{6} = -1 \quad \text{and} \quad z_2 = \frac{190 - 175}{6} = 2.5.$$

Using values from Table 3 of Murdoch and Barnes' *Statistical Tables*, we have:

$$\begin{aligned} P(X < 169) &= P(Z < -1) = \Phi(-1) \\ &= 1 - \Phi(1) = 1 - 0.8413 = 0.1587 \end{aligned}$$

$$\begin{aligned} P(X > 190) &= P(Z > 2.5) = 1 - \Phi(2.5) \\ &= 1 - 0.9938 = 0.0062 \end{aligned}$$

and:

$$\begin{aligned} P(169 < X < 190) &= P(-1 < Z < 2.5) = \Phi(2.5) - \Phi(-1) \\ &= 0.9938 - 0.1587 = 0.8351. \end{aligned}$$

19. 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 $B$?

**Solution:**

The standardised $z$-value giving 5% in the upper tail is 1.6449, and for 10% it is 1.2816. So, converting to the scale for IQ scores, the values are:

$$1.6449 \times 15 = 24.6735 \quad \text{and} \quad 1.2816 \times 15 = 19.224.$$

Write the means according to $A$ and $B$ as $\mu_A$ and $\mu_B$, respectively. Therefore:

$$\mu_A + 24.6735 = 134.6735$$

so:

$$\mu_A = 110$$

whereas:

$$\mu_B + 19.224 = 109.224$$

so $\mu_B = 90$. The difference $\mu_A - \mu_B = 110 - 90 = 20$.

## D.2 Practice questions

Try to solve the questions before looking at the solutions – promise?! Solutions are located in Appendix F.

1. 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 parameters as follows.

   - Piston diameters: mean 10.42 cm, standard deviation 0.03 cm.
   - Cylinder diameters: mean 10.52 cm, standard deviation 0.04 cm.

   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)?

   (a) 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?

   (b) Calculate both of these probabilities:

   i. exactly

   ii. using a Poisson approximation.

   Discuss the appropriateness of using this approximation.

2. If $X$ has the discrete uniform distribution such that $P(X = i) = 1/k$ for $i = 1, 2, \ldots, k$, show that its moment generating function is:

$$M_X(t) = \frac{\mathrm{e}^t(1 - \mathrm{e}^{kt})}{k(1 - \mathrm{e}^t)}.$$

(Do not attempt to find the mean and variance using the mgf.)

3. Let $f(z)$ be defined as:

$$f(z) = \frac{1}{2}\mathrm{e}^{-|z|} \quad \text{for all real values of } z.$$

(a) Sketch $f(z)$ and explain why it can serve as the pdf for a random variable $Z$.

(b) Determine the moment generating function of $Z$.

(c) Use the mgf to find $\mathrm{E}(Z)$, $\mathrm{Var}(Z)$, $\mathrm{E}(Z^3)$ and $\mathrm{E}(Z^4)$.

(You may assume that $-1 < t < 1$, for the mgf, which will ensure convergence.)

4. Show that for a binomial random variable $X \sim \mathrm{Bin}(n, \pi)$, then:

$$\mathrm{E}(X) = n\pi \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!\,(n-x)!}\pi^{x-1}(1-\pi)^{n-x}.$$

Hence find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$. (The wording of the question implies that you use the result which you have just proved. Other methods of derivation will not be accepted!)

5. 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.

6. James goes fishing every Saturday. The number of fish he catches follows a Poisson distribution. On a proportion $\pi$ of the days he goes fishing, he does not catch anything. He makes it a rule to take home the first, and then every other, fish which he catches, i.e. the first, third, fifth fish etc.

(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 $(1 - \pi^2)/2$.

*There are two kinds of statistics, the kind you look up and the kind you make up.*

(Rex Stout)

# Appendix E
# Multivariate random variables

## E.1 Worked examples

1. $X$ and $Y$ are independent random variables with distributions as follows:

| $X = x$ | 0 | 1 | 2 |
|---|---|---|---|
| $p_X(x)$ | 0.4 | 0.2 | 0.4 |

| $Y = y$ | 1 | 2 |
|---|---|---|
| $p_Y(y)$ | 0.4 | 0.6 |

The random variables $W$ and $Z$ are defined by $W = 2X$ and $Z = Y - X$, respectively.

(a) Compute the joint distribution of $W$ and $Z$.

(b) Evaluate $P(W = 2 \,|\, Z = 1)$, $\mathrm{E}(W \,|\, Z = 0)$ and $\mathrm{Cov}(W, Z)$.

**Solution:**

(a) The joint distribution (with marginal probabilities) is:

| | | $W = w$ | | | |
|---|---|---|---|---|---|
| | | 0 | 2 | 4 | $p_Z(z)$ |
| | $-1$ | 0.00 | 0.00 | 0.16 | 0.16 |
| $Z = z$ | 0 | 0.00 | 0.08 | 0.24 | 0.32 |
| | 1 | 0.16 | 0.12 | 0.00 | 0.28 |
| | 2 | 0.24 | 0.00 | 0.00 | 0.24 |
| | $p_W(w)$ | 0.40 | 0.20 | 0.40 | 1.00 |

(b) It is straightforward to see that:

$$P(W = 2 \,|\, Z = 1) = \frac{P(W = 2 \cap Z = 1)}{P(Z = 1)} = \frac{0.12}{0.28} = \frac{3}{7}.$$

For $\mathrm{E}(W \,|\, Z = 0)$, we have:

$$\mathrm{E}(W \,|\, Z = 0) = \sum_w w\, P(W = w \,|\, Z = 0) = 0 \times \frac{0}{0.32} + 2 \times \frac{0.08}{0.32} + 4 \times \frac{0.24}{0.32} = 3.5.$$

We see $\mathrm{E}(W) = 2$ (by symmetry), and:

$$\mathrm{E}(Z) = -1 \times 0.16 + 0 \times 0.32 + 1 \times 0.28 + 2 \times 0.24 = 0.6.$$

Also:

$$\mathrm{E}(WZ) = \sum_w \sum_z wz\, p(w, z) = -4 \times 0.16 + 2 \times 0.12 = -0.4$$

hence:

$$\mathrm{Cov}(W, Z) = \mathrm{E}(WZ) - \mathrm{E}(W)\,\mathrm{E}(Z) = -0.4 - 2 \times 0.6 = -1.6.$$

2. The joint probability distribution of the random variables $X$ and $Y$ is:

| | | $X = x$ | | |
|---|---|---|---|---|
| | | $-1$ | $0$ | $1$ |
| $Y = y$ | $-1$ | $0.05$ | $0.15$ | $0.10$ |
| | $0$ | $0.10$ | $0.05$ | $0.25$ |
| | $1$ | $0.10$ | $0.05$ | $0.15$ |

(a) Identify the marginal distributions of $X$ and $Y$ and the conditional distribution of $X$ given $Y = 1$.

(b) Evaluate $\mathrm{E}(X \mid Y = 1)$ and the correlation coefficient of $X$ and $Y$.

(c) Are $X$ and $Y$ independent random variables?

**Solution:**

(a) The marginal and conditional distributions are, respectively:

| $X = x$ | $-1$ | $0$ | $1$ |
|---|---|---|---|
| $p_X(x)$ | $0.25$ | $0.25$ | $0.50$ |

| $Y = y$ | $-1$ | $0$ | $1$ |
|---|---|---|---|
| $p_Y(y)$ | $0.30$ | $0.40$ | $0.30$ |

| $X = x \mid Y = 1$ | $-1$ | $0$ | $1$ |
|---|---|---|---|
| $p_{X\mid Y=1}(x \mid Y = 1)$ | $1/3$ | $1/6$ | $1/2$ |

(b) From the conditional distribution we see:

$$\mathrm{E}(X \mid Y = 1) = -1 \times \frac{1}{3} + 0 \times \frac{1}{6} + 1 \times \frac{1}{2} = \frac{1}{6}.$$

$\mathrm{E}(Y) = 0$ (by symmetry), and so $\mathrm{Var}(Y) = \mathrm{E}(Y^2) = 0.6$.

$\mathrm{E}(X) = 0.25$ and:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = 0.75 - (0.25)^2 = 0.6875.$$

(Note that $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ are not strictly necessary here!)

Next:

$$\begin{aligned}\mathrm{E}(XY) &= \sum_x \sum_y xy\, p(x, y) \\ &= (-1)(-1)(0.05) + (1)(-1)(0.1) + (-1)(1)(0.1) + (1)(1)(0.15) \\ &= 0.\end{aligned}$$

So:

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = 0 \quad \Rightarrow \quad \mathrm{Corr}(X, Y) = 0.$$

(c) $X$ and $Y$ are not independent random variables since, for example:

$$P(X = 1, Y = -1) = 0.1 \neq P(X = 1)\,P(Y = -1) = 0.5 \times 0.3 = 0.15.$$

3. $X_1, X_2, \ldots, X_n$ are independent Bernoulli random variables. The probability function of $X_i$ is given by:

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

where:

$$\pi_i = \frac{\mathrm{e}^{i\theta}}{1+\mathrm{e}^{i\theta}}$$

for $i = 1, 2, \ldots, n$. Derive the joint probability function, $p(x_1, x_2, \ldots, x_n)$.

**Solution:**

Since the $X_i$s are independent (but not identically distributed) random variables, we have:

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i).$$

So, the joint probability function is:

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \left(\frac{1}{1+\mathrm{e}^{i\theta}}\right)^{1-x_i} \left(\frac{\mathrm{e}^{i\theta}}{1+\mathrm{e}^{i\theta}}\right)^{x_i} = \prod_{i=1}^{n} \left(\frac{\mathrm{e}^{i\theta x_i}}{1+\mathrm{e}^{i\theta}}\right) = \frac{\mathrm{e}^{\theta \sum\limits_{i=1}^{n} i x_i}}{\prod\limits_{i=1}^{n}(1+\mathrm{e}^{i\theta})}.$$

4. $X_1, X_2, \ldots, X_n$ are independent random variables with the common probability density function:

$$f(x) = \begin{cases} \lambda^2 x \mathrm{e}^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Derive the joint probability density function, $f(x_1, x_2, \ldots, x_n)$.

**Solution:**

Since the $X_i$s are independent (and identically distributed) random variables, we have:

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i).$$

So, the joint probability density function is:

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \lambda^2 x_i \mathrm{e}^{-\lambda x_i} = \lambda^{2n} \prod_{i=1}^{n} x_i \mathrm{e}^{-\lambda x_1 - \lambda x_2 - \cdots - \lambda x_n} = \lambda^{2n} \left(\prod_{i=1}^{n} x_i\right) \mathrm{e}^{-\lambda \sum\limits_{i=1}^{n} x_i}.$$

5. $X_1, X_2, \ldots, X_n$ are independent random variables with the common probability function:

$$p(x) = \binom{m}{x} \frac{\theta^x}{(1+\theta)^m} \quad \text{for } x = 0, 1, 2, \ldots, m$$

and 0 otherwise. Derive the joint probability function, $p(x_1, x_2, \ldots, x_n)$.

**Solution:**

Since the $X_i$s are independent (and identically distributed) random variables, we have:

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i).$$

So, the joint probability function is:

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \binom{m}{x_i} \frac{\theta^{x_i}}{(1+\theta)^m} = \left(\prod_{i=1}^{n} \binom{m}{x_i}\right) \frac{\theta^{x_1}\theta^{x_2}\cdots\theta^{x_n}}{(1+\theta)^{nm}} = \left(\prod_{i=1}^{n} \binom{m}{x_i}\right) \frac{\theta^{\sum\limits_{i=1}^{n} x_i}}{(1+\theta)^{nm}}.$$

6. The random variables $X_1$ and $X_2$ are independent and have the common distribution given in the table below:

| $X = x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $p_X(x)$ | 0.2 | 0.4 | 0.3 | 0.1 |

The random variables $W$ and $Y$ are defined by $W = \max(X_1, X_2)$ and $Y = \min(X_1, X_2)$.

(a) Calculate the table of probabilities which defines the joint distribution of $W$ and $Y$.

(b) Find:

i. the marginal distribution of $W$

ii. the conditional distribution of $Y$ given $W = 2$

iii. $\mathrm{E}(Y \mid W = 2)$ and $\mathrm{Var}(Y \mid W = 2)$

iv. $\mathrm{Cov}(W, Y)$.

**Solution:**

(a) The joint distribution of $W$ and $Y$ is:

| | | $W = w$ | | | |
|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 |
| $Y = y$ | 0 | $(0.2)^2$ | $2(0.2)(0.4)$ | $2(0.2)(0.3)$ | $2(0.2)(0.1)$ |
| | 1 | 0 | $(0.4)(0.4)$ | $2(0.4)(0.3)$ | $2(0.4)(0.1)$ |
| | 2 | 0 | 0 | $(0.3)(0.3)$ | $2(0.3)(0.1)$ |
| | 3 | 0 | 0 | 0 | $(0.1)(0.1)$ |
| | | $(0.2)^2$ | $(0.8)(0.4)$ | $(1.5)(0.3)$ | $(1.9)(0.1)$ |

which is:

| | | $W = w$ | | | |
|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 |
| $Y = y$ | 0 | 0.04 | 0.16 | 0.12 | 0.04 |
| | 1 | 0.00 | 0.16 | 0.24 | 0.08 |
| | 2 | 0.00 | 0.00 | 0.09 | 0.06 |
| | 3 | 0.00 | 0.00 | 0.00 | 0.01 |
| | | 0.04 | 0.32 | 0.45 | 0.19 |

(b) i. Hence the marginal distribution of $W$ is:

| $W = w$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $p_W(w)$ | 0.04 | 0.32 | 0.45 | 0.19 |

ii. The conditional distribution of $Y \mid W = 2$ is:

| $Y = y \mid W = 2$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $p_{Y\mid W=2}(y \mid W = 2)$ | $4/15$<br>$= 0.2\dot{6}$ | $8/15$<br>$= 0.5\dot{3}$ | $2/10$<br>$= 0.2$ | 0<br>0 |

iii. We have:

$$\mathrm{E}(Y \mid W = 2) = 0 \times \frac{4}{15} + 1 \times \frac{8}{15} + 2 \times \frac{2}{10} + 3 \times 0 = 0.9\dot{3}$$

and:

$$\mathrm{Var}(Y \mid W = 2) = \mathrm{E}(Y^2 \mid W = 2) - (\mathrm{E}(Y \mid W = 2))^2 = 1.\dot{3} - (0.9\dot{3})^2 = 0.4622.$$

iv. $\mathrm{E}(WY) = 1.69$, $\mathrm{E}(W) = 1.79$ and $\mathrm{E}(Y) = 0.81$, therefore:

$$\mathrm{Cov}(W, Y) = \mathrm{E}(WY) - \mathrm{E}(W)\,\mathrm{E}(Y) = 1.69 - 1.79 \times 0.81 = 0.2401.$$

7. Consider two random variables $X$ and $Y$. $X$ can take the values $-1$, 0 and 1, and $Y$ can take the values 0, 1 and 2. The joint probabilities for each pair are given by the following table:

| | $X = -1$ | $X = 0$ | $X = 1$ |
|---|---|---|---|
| $Y = 0$ | 0.10 | 0.20 | 0.10 |
| $Y = 1$ | 0.10 | 0.05 | 0.10 |
| $Y = 2$ | 0.10 | 0.05 | 0.20 |

(a) Calculate the marginal distributions and expected values of $X$ and $Y$.

(b) Calculate the covariance of the random variables $U$ and $V$, where $U = X + Y$ and $V = X - Y$.

(c) Calculate $\mathrm{E}(V \mid U = 1)$.

**Solution:**

(a) The marginal distribution of $X$ is:

| $X = x$ | $-1$ | 0 | 1 |
|---|---|---|---|
| $p_X(x)$ | 0.3 | 0.3 | 0.4 |

The marginal distribution of $Y$ is:

| $Y = y$ | 0 | 1 | 2 |
|---|---|---|---|
| $p_Y(y)$ | 0.40 | 0.25 | 0.35 |

Hence:

$$\mathrm{E}(X) = -1 \times 0.3 + 0 \times 0.3 + 1 \times 0.4 = 0.1$$

and:

$$\mathrm{E}(Y) = 0 \times 0.40 + 1 \times 0.25 + 2 \times 0.35 = 0.95.$$

(b) We have:

$$\begin{aligned}\operatorname{Cov}(U,V) &= \operatorname{Cov}(X+Y,X-Y)\\ &= \mathrm{E}((X+Y)(X-Y)) - \mathrm{E}(X+Y)\,\mathrm{E}(X-Y)\\ &= \mathrm{E}(X^2-Y^2) - (\mathrm{E}(X)+\mathrm{E}(Y))(\mathrm{E}(X)-\mathrm{E}(Y))\\ \mathrm{E}(X^2) &= ((-1)^2\times 0.3) + (0^2\times 0.3) + (1^2\times 0.4) = 0.7\\ \mathrm{E}(Y^2) &= (0^2\times 0.4) + (1^2\times 0.25) + (2^2\times 0.35) = 1.65\end{aligned}$$

hence:

$$\operatorname{Cov}(U,V) = (0.7-1.65) - (0.1+0.95)(0.1-0.95) = -0.0575.$$

(c) $U=1$ is achieved for $(X,Y)$ pairs $(-1,2)$, $(0,1)$ or $(1,0)$. The corresponding values of $V$ are $-3$, $-1$ and 1. We have:

$$\begin{aligned}P(U=1) &= 0.1+0.05+0.1 = 0.25\\ P(V=-3\,|\,U=1) &= \frac{0.1}{0.25} = \frac{2}{5}\\ P(V=-1\,|\,U=1) &= \frac{0.05}{0.25} = \frac{1}{5}\\ P(V=1\,|\,U=1) &= \frac{0.1}{0.25} = \frac{2}{5}\end{aligned}$$

hence:

$$\mathrm{E}(V\,|\,U=1) = \left(-3\times\frac{2}{5}\right) + \left(-1\times\frac{1}{5}\right) + \left(1\times\frac{2}{5}\right) = -1.$$

8. Two refills for a ballpoint pen are selected at random from a box containing three blue refills, two red refills and three green refills. Define the following random variables:

$X =$ the number of blue refills selected

$Y =$ the number of red refills selected.

(a) Show that $P(X=1,Y=1)=3/14$.

(b) Form the table showing the joint probability distribution of $X$ and $Y$.

(c) Calculate $\mathrm{E}(X)$, $\mathrm{E}(Y)$ and $\mathrm{E}(X\,|\,Y=1)$.

(d) Find the covariance between $X$ and $Y$.

(e) Are $X$ and $Y$ independent random variables? Give a reason for your answer.

**Solution:**

(a) With the obvious notation $B$ = blue and $R$ = red:

$$P(X=1,Y=1)=P(BR)+P(RB)=\frac{3}{8}\times\frac{2}{7}+\frac{2}{8}\times\frac{3}{7}=\frac{3}{14}.$$

(b) We have:

| | | $X=x$ | | |
|---|---|---|---|---|
| | | 0 | 1 | 2 |
| | 0 | 3/28 | 9/28 | 3/28 |
| $Y=y$ | 1 | 3/14 | 3/14 | 0 |
| | 2 | 1/28 | 0 | 0 |

(c) The marginal distribution of $X$ is:

| $X=x$ | 0 | 1 | 2 |
|---|---|---|---|
| $p_X(x)$ | 10/28 | 15/28 | 3/28 |

Hence:

$$\mathrm{E}(X)=0\times\frac{10}{28}+1\times\frac{15}{28}+2\times\frac{3}{28}=\frac{3}{4}.$$

The marginal distribution of $Y$ is:

| $Y=y$ | 0 | 1 | 2 |
|---|---|---|---|
| $p_Y(y)$ | 15/28 | 12/28 | 1/28 |

Hence:

$$\mathrm{E}(Y)=0\times\frac{15}{28}+1\times\frac{12}{28}+2\times\frac{1}{28}=\frac{1}{2}.$$

The conditional distribution of $X$ given $Y=1$ is:

| $X=x\,|\,Y=1$ | 0 | 1 |
|---|---|---|
| $p_{X|Y=1}(x\,|\,y=1)$ | 1/2 | 1/2 |

Hence:

$$\mathrm{E}(X\,|\,Y=1)=0\times\frac{1}{2}+1\times\frac{1}{2}=\frac{1}{2}.$$

(d) The distribution of $XY$ is:

| $XY=xy$ | 0 | 1 |
|---|---|---|
| $p_{XY}(xy)$ | 22/28 | 6/28 |

Hence:

$$\mathrm{E}(XY)=0\times\frac{22}{28}+1\times\frac{6}{28}=\frac{3}{14}$$

and:

$$\mathrm{Cov}(X,Y)=\mathrm{E}(XY)-\mathrm{E}(X)\,\mathrm{E}(Y)=\frac{3}{14}-\frac{3}{4}\times\frac{1}{2}=-\frac{9}{56}.$$

(e) Since $\mathrm{Cov}(X,Y)\neq 0$, a necessary condition for independence fails to hold. The random variables are not independent.

9. Show that the marginal distributions of a bivariate distribution are not enough to define the bivariate distribution itself.

**Solution:**

Here we must show that there are two distinct bivariate distributions with the same marginal distributions. It is easiest to think of the simplest case where $X$ and $Y$ each take only two values, say 0 and 1.

Suppose the marginal distributions of $X$ and $Y$ are the same, with $p(0) = p(1) = 0.5$. One possible bivariate distribution with these marginal distributions is the one for which there is independence between $X$ and $Y$. This has $p_{X,Y}(x, y) = p_X(x)\,p_Y(y)$ for all $x, y$. Writing it in full:

$$p_{X,Y}(0,0) = p_{X,Y}(1,0) = p_{X,Y}(0,1) = p_{X,Y}(1,1) = 0.5 \times 0.5 = 0.25.$$

The table of probabilities for this choice of independence is shown in the first table below.

Trying some other value for $p_{X,Y}(0,0)$, like 0.2, gives the second table below.

| $X/Y$ | 0 | 1 |
|---|---|---|
| 0 | 0.25 | 0.25 |
| 1 | 0.25 | 0.25 |

| $X/Y$ | 0 | 1 |
|---|---|---|
| 0 | 0.2 | 0.3 |
| 1 | 0.3 | 0.2 |

The construction of these probabilities is done by making sure the row and column totals are equal to 0.5, and so we now have a second distribution with the same marginal distributions as the first.

This example is very simple, but one can almost always construct many bivariate distributions with the same marginal distributions even for continuous random variables.

10. Show that if:

$$P(X \leq x \cap Y \leq y) = (1 - \mathrm{e}^{-x})(1 - \mathrm{e}^{-2y})$$

for all $x, y \geq 0$, then $X$ and $Y$ are independent random variables, each with an exponential distribution.

**Solution:**

The right-hand side of the result given is the product of the cdf of an exponential random variable $X$ with mean 1 and the cdf of an exponential random variable $Y$ with mean 2. So the result follows from the definition of independent random variables.

11. There are different ways to write the covariance. Show that:

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

and:

$$\mathrm{Cov}(X, Y) = \mathrm{E}((X - \mathrm{E}(X))Y) = \mathrm{E}(X(Y - \mathrm{E}(Y))).$$

**Solution:**

Working directly from the definition:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}((X-\mathrm{E}(X))(Y-\mathrm{E}(Y)))\\ &= \mathrm{E}(XY - X\,\mathrm{E}(Y) - \mathrm{E}(X)Y + \mathrm{E}(X)\,\mathrm{E}(Y))\\ &= \mathrm{E}(XY) + \mathrm{E}(-X\,\mathrm{E}(Y)) + \mathrm{E}(-\mathrm{E}(X)Y) + \mathrm{E}(\mathrm{E}(X)\,\mathrm{E}(Y))\\ &= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) - \mathrm{E}(X)\,\mathrm{E}(Y) + \mathrm{E}(X)\,\mathrm{E}(Y)\\ &= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y).\end{aligned}$$

For the second part, we begin with the right-hand side:

$$\begin{aligned}\mathrm{E}((X-\mathrm{E}(X))Y) &= \mathrm{E}(XY - \mathrm{E}(X)Y)\\ &= \mathrm{E}(XY) + \mathrm{E}(-\mathrm{E}(X)Y)\\ &= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)\\ &= \mathrm{Cov}(X,Y).\end{aligned}$$

The remaining result follows by an argument symmetric with the last one.

12. Suppose that $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$, and that $X$ and $Y$ have correlation coefficient $\rho$. Show that it follows from $\mathrm{Var}(X-\rho Y) \geq 0$ that $\rho^2 \leq 1$.

**Solution:**

We have:

$$0 \leq \mathrm{Var}(X-\rho Y) = \mathrm{Var}(X) - 2\rho\,\mathrm{Cov}(X,Y) + \rho^2\mathrm{Var}(Y) = 1 - 2\rho^2 + \rho^2 = (1-\rho^2).$$

Hence $1-\rho^2 \geq 0$, and so $\rho^2 \leq 1$.

13. The distribution of a random variable $X$ is:

| $X = x$ | $-1$ | $0$ | $1$ |
|---|---|---|---|
| $P(X = x)$ | $a$ | $b$ | $a$ |

Show that $X$ and $X^2$ are uncorrelated.

**Solution:**

This is an example of two random variables $X$ and $Y = X^2$ which are uncorrelated, but obviously dependent. The bivariate distribution of $(X,Y)$ in this case is *singular* because of the complete functional dependence between them.

We have:

$$\begin{aligned}\mathrm{E}(X) &= -1\times a + 0\times b + 1\times a = 0\\ \mathrm{E}(X^2) &= +1\times a + 0\times b + 1\times a = 2a\\ \mathrm{E}(X^3) &= -1\times a + 0\times b + 1\times a = 0\end{aligned}$$

and we must show that the covariance is zero:

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = \mathrm{E}(X^3) - \mathrm{E}(X)\,\mathrm{E}(X^2) = 0 - 0 \times 2a = 0.$$

There are many possible choices for $a$ and $b$ which give a valid probability distribution, for instance $a = 0.25$ and $b = 0.5$.

14. A fair coin is thrown $n$ times, each throw being independent of the ones before. Let $R$ = 'the number of heads', and $S$ = 'the number of tails'. Find the covariance of $R$ and $S$. What is the correlation of $R$ and $S$?

**Solution:**

One can go about this in a straightforward way. If $X_i$ is the number of heads and $Y_i$ is the number of tails on the $i$th throw, then the distribution of $X_i$ and $Y_i$ is given by:

| $X/Y$ | 0 | 1 |
|---|---|---|
| 0 | 0 | 0.5 |
| 1 | 0.5 | 0 |

From this table, we compute the following:

$$\mathrm{E}(X_i) = \mathrm{E}(Y_i) = 0 \times 0.5 + 1 \times 0.5 = 0.5$$

$$\mathrm{E}(X_i^2) = \mathrm{E}(Y_i^2) = 0 \times 0.5 + 1 \times 0.5 = 0.5$$

$$\mathrm{Var}(X_i) = \mathrm{Var}(Y_i) = 0.5 - (0.5)^2 = 0.25$$

$$\mathrm{E}(X_iY_i) = 0 \times 0.5 + 0 \times 0.5 = 0$$

$$\mathrm{Cov}(X_i,Y_i) = \mathrm{E}(X_iY_i) - \mathrm{E}(X_i)\,\mathrm{E}(Y_i) = 0 - 0.25 = -0.25.$$

Now, since $R = \sum_i X_i$ and $S = \sum_i Y_i$, we can add covariances of independent $X_i$s and $Y_i$s, just like means and variances, then:

$$\mathrm{Cov}(R,S) = -0.25n.$$

Since $R + S = n$ is a fixed quantity, there is a complete linear dependence between $R$ and $S$. We have $R = n - S$, so the correlation between $R$ and $S$ should be $-1$. This can be checked directly since:

$$\mathrm{Var}(R) = \mathrm{Var}(S) = 0.25n$$

(add the variances of the $X_i$s or $Y_i$s). The correlation between $R$ and $S$ works out as $-0.25n/0.25n = -1$.

15. Suppose that $X$ and $Y$ have a bivariate distribution. Find the covariance of the new random variables $W = aX + bY$ and $V = cX + dY$ where $a$, $b$, $c$ and $d$ are constants.

**Solution:**

The covariance of $W$ and $V$ is:

$$\begin{aligned}\mathrm{E}(WV) - \mathrm{E}(W)\,\mathrm{E}(V) &= \mathrm{E}(acX^2 + bdY^2 + (ad+bc)XY) \\ &\quad - (ac\,\mathrm{E}(X)^2 + bd\,\mathrm{E}(Y)^2 + (ad+bc)\,\mathrm{E}(X)\,\mathrm{E}(Y)) \\ &= ac(\mathrm{E}(X^2) - \mathrm{E}(X)^2) + bd(\mathrm{E}(Y^2) - \mathrm{E}(Y)^2) \\ &\quad + (ad+bc)(\mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)) \\ &= ac\sigma_X^2 + bd\sigma_Y^2 + (ad+bc)\sigma_{XY}.\end{aligned}$$

16. Following on from Question 15, show that, if the variances of $X$ and $Y$ are the same, then $W = X + Y$ and $V = X - Y$ are uncorrelated.

**Solution:**

Here we have $a = b = c = 1$ and $d = -1$. Substituting into the formula found above:

$$\sigma_{WV} = \sigma_X^2 - \sigma_Y^2 = 0.$$

There is no assumption here that $X$ and $Y$ are independent. It is not true that $W$ and $V$ are independent without further restrictions on $X$ and $Y$.

## E.2 Practice questions

Try to solve the questions before looking at the solutions – promise?! Solutions are located in Appendix F.

1. (a) For random variables $X$ and $Y$, show that:

$$\mathrm{Cov}(X+Y, X-Y) = \mathrm{Var}(X) - \mathrm{Var}(Y).$$

(b) For random variables $X$ and $Y$, and constants $a$, $b$, $c$ and $d$, show that:

$$\mathrm{Cov}(a+bX, c+dY) = bd\,\mathrm{Cov}(X,Y).$$

2. Let $X_1, X_2, \ldots, X_k$ be independent random variables, and $a_1, a_2, \ldots, a_k$ be constants. Show that:

(a) $\mathrm{E}\left(\sum\limits_{i=1}^{k} a_i X_i\right) = \sum\limits_{i=1}^{k} a_i \mathrm{E}(X_i)$

(b) $\mathrm{Var}\left(\sum\limits_{i=1}^{k} a_i X_i\right) = \sum\limits_{i=1}^{k} a_i^2 \mathrm{Var}(X_i).$

3. $X$ and $Y$ are discrete random variables which can assume values 0, 1 and 2 only.

$$P(X = x, Y = y) = A(x+y) \text{ for some constant } A \text{ and } x,\, y \in \{0, 1, 2\}.$$

(a) Draw up a table to describe the joint distribution of $X$ and $Y$ and find the value of the constant $A$.

(b) Describe the marginal distributions of $X$ and $Y$.

(c) Give the conditional distribution of $X \mid Y = 1$ and find $\mathrm{E}(X \mid Y = 1)$.

(d) Are $X$ and $Y$ independent? Give a reason for your answer.

*Statistics are like bikinis. What they reveal is suggestive, but what they conceal is vital.*
(Aaron Levenstein)

# Appendix F
# Solutions to Practice questions

## F.1 Chapter 1 – Data visualisation and descriptive statistics

1. (a) We have:

$$\sum_{j=1}^{3}(Y_j - \bar{Y}) = (Y_1 - \bar{Y}) + (Y_2 - \bar{Y}) + (Y_3 - \bar{Y}).$$

However:

$$3\bar{Y} = Y_1 + Y_2 + Y_3$$

hence:

$$\sum_{j=1}^{3}(Y_j - \bar{Y}) = 3\bar{Y} - 3\bar{Y} = 0.$$

(b) We have:

$$\begin{aligned}\sum_{j=1}^{3}\sum_{k=1}^{3}(Y_j - \bar{Y})(Y_k - \bar{Y}) &= (Y_1 - \bar{Y})((Y_1 - \bar{Y}) + (Y_2 - \bar{Y}) + (Y_3 - \bar{Y}))\\ &\quad + (Y_2 - \bar{Y})((Y_1 - \bar{Y}) + (Y_2 - \bar{Y}) + (Y_3 - \bar{Y}))\\ &\quad + (Y_3 - \bar{Y})((Y_1 - \bar{Y}) + (Y_2 - \bar{Y}) + (Y_3 - \bar{Y}))\\ &= ((Y_1 - \bar{Y}) + (Y_2 - \bar{Y}) + (Y_3 - \bar{Y}))^2\end{aligned}$$

and $3\bar{Y} = Y_1 + Y_2 + Y_3$ as above, so:

$$\sum_{j=1}^{3}\sum_{k=1}^{3}(Y_j - \bar{Y})(Y_k - \bar{Y}) = 0^2 = 0.$$

(c) We have:

$$\sum_{j=1}^{3}\sum_{k=1}^{3}(Y_j - \bar{Y})(Y_k - \bar{Y}) = \sum_{j=1}^{3}\ {}_{j\neq k}\sum_{k=1}^{3}(Y_j - \bar{Y})(Y_k - \bar{Y}) + \sum_{j=1}^{3}(Y_j - \bar{Y})^2$$

We have written the nine terms in the left-hand expression as the sum of the six terms for which $j \neq k$, and the three terms for which $j = k$.

However, we showed in (b) that the left-hand expression is in fact 0, so:

$$0 = \sum_{j=1}^{3} {}_{j\neq k} \sum_{k=1}^{3} (Y_j - \bar{Y})(Y_k - \bar{Y}) + \sum_{j=1}^{3} (Y_j - \bar{Y})^2$$

from which the result follows.

2. (a) We have:

$$\bar{y} = \frac{\sum\limits_{i=1}^{n} y_i}{n} = \frac{\sum\limits_{i=1}^{n} (ax_i + b)}{n} = \frac{a\sum\limits_{i=1}^{n} x_i + nb}{n} = a\bar{x} + b.$$

(b) Multiply out the square within the summation sign and then evaluate the three expressions, remembering that $\bar{x}$ is a constant with respect to summation and can be taken outside the summation sign as a common factor, i.e. we have:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\end{aligned}$$

hence the result. Recall that $\sum\limits_{i=1}^{n} x_i = n\bar{x}$.

(c) It is probably best to work with variances to avoid the square roots. The variance of $y$ values, say $s_y^2$, is given by:

$$\begin{aligned}s_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(ax_i + b - (a\bar{x} + b))^2 \\ &= a^2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= a^2 s_x^2.\end{aligned}$$

The result follows on taking the square root, observing that the standard deviation cannot be a negative quantity.

Adding a constant $k$ to each value of a dataset adds $k$ to the mean and leaves the standard deviation unchanged. This corresponds to a transformation $y_i = ax_i + b$ with $a = 1$ and $b = k$. Apply (a) and (c) with these values.

Multiplying each value of a dataset by a constant $c$ multiplies the mean by $c$ and also the standard deviation by $|c|$. This corresponds to a transformation $y_i = cx_i$ with $a = c$ and $b = 0$. Apply (a) and (c) with these values.

## F.2 Chapter 2 – Probability theory

1. (a) We know $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

   Consider $A \cup B \cup C$ as $(A \cup B) \cup C$ (i.e. as the union of the two sets $A \cup B$ and $C$) and then apply the result above to obtain:

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

   Now $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ – a Venn diagram can be drawn to check this.

   So:

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

   using the earlier result again for $A \cap C$ and $B \cap C$.

   Now $(A \cap C) \cap (B \cap C) = A \cap B \cap C$ and if we apply the earlier result once more for $A$ and $B$, we obtain:

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

   which is the required result.

   (b) Use the result that if $X \subset Y$ then $P(X) \leq P(Y)$ for events $X$ and $Y$.

   Since $A \subset A \cup B$ and $B \subset A \cup B$, we have $P(A) \leq P(A \cup B)$ and $P(B) \leq P(A \cup B)$.

   Adding these inequalities, $P(A) + P(B) \leq 2P(A \cup B)$ so:

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

   Similarly, $A \cap B \subset A$ and $A \cap B \subset B$, so $P(A \cap B) \leq P(A)$ and $P(A \cap B) \leq P(B)$.

   Adding, $2P(A \cap B) \leq P(A) + P(B)$ so:

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

2. (a) We know that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. For independent events $A$ and $B$, $P(A \cap B) = P(A)\,P(B)$, so $P(A \cup B) = P(A) + P(B) - P(A)\,P(B)$ gives $0.75 = p + 2p - 2p^2$, or $2p^2 - 3p + 0.75 = 0$.

   Solving the quadratic equation gives:

   $$p = \frac{3 - \sqrt{3}}{4} \approx 0.317$$

   suppressing the irrelevant case for which $p > 1$.

   Since $A$ and $B$ are independent, $P(A \mid B) = P(A) = p = 0.317$.

(b) For mutually exclusive events, $P(A \cup B) = P(A) + P(B)$, so $0.75 = p + 2p$, leading to $p = 0.25$.

Here $P(A \cap B) = 0$, so $P(A \mid B) = P(A \cap B)/P(B) = 0$.

3. (a) We are given that $A$ and $B$ are independent, so $P(A \cap B) = P(A)\,P(B)$. We need to show a similar result for $A^c$ and $B^c$, namely we need to show that $P(A^c \cap B^c) = P(A^c)\,P(B^c)$.

Now $A^c \cap B^c = (A \cup B)^c$ from basic set theory (draw a Venn diagram), hence:

$$\begin{aligned}
P(A^c \cap B^c) &= P((A \cup B)^c) \\
&= 1 - P(A \cup B) \\
&= 1 - (P(A) + P(B) - P(A \cap B)) \\
&= 1 - P(A) - P(B) + P(A \cap B) \\
&= 1 - P(A) - P(B) + P(A)\,P(B) \quad \text{(independence assumption)} \\
&= (1 - P(A))(1 - P(B)) \quad \text{(factorising)} \\
&= P(A^c)\,P(B^c) \quad \text{(as required).}
\end{aligned}$$

(b) To show that $X^c$ and $Y^c$ are not necessarily mutually exclusive when $X$ and $Y$ are mutually exclusive, the best approach is to find a counterexample. Attempts to 'prove' the result directly are likely to be logically flawed.

Look for a simple example. Suppose we roll a die. Let $X = \{6\}$ be the event of obtaining a 6, and let $Y = \{5\}$ be the event of obtaining a 5. Obviously $X$ and $Y$ are mutually exclusive, but $X^c = \{1, 2, 3, 4, 5\}$ and $Y^c = \{1, 2, 3, 4, 6\}$ have $X^c \cap Y^c \neq \emptyset$, so $X^c$ and $Y^c$ are not mutually exclusive.

4. (a) $A$ will win the game without deuce if he or she wins four points, including the last point, before $B$ wins three points. This can occur in three ways.

- $A$ wins four straight points, i.e. $AAAA$ with probability $(2/3)^4 = 16/81$.
- $B$ wins just one point in the game. There are ${}^4C_1$ ways for this to happen, namely $BAAAA$, $ABAAA$, $AABAA$ and $AAABA$. Each has probability $(1/3)(2/3)^4$, so the probability of one of these outcomes is given by $4(1/3)(2/3)^4 = 64/243$.
- $B$ wins just two points in the game. There are ${}^5C_2$ ways for this to happen, namely $BBAAAA$, $BABAAA$, $BAABAA$, $BAAABA$, $ABBAAA$, $ABABAA$, $ABAABA$, $AABBAA$, $AABABA$ and $AAABBA$. Each has probability $(1/3)^2(2/3)^4$, so the probability of one of these outcomes is given by $10(1/3)^2(2/3)^4 = 160/729$.

Therefore, the probability that $A$ wins without a deuce must be the sum of these, namely:

$$\frac{16}{81} + \frac{64}{243} + \frac{160}{729} = \frac{144 + 192 + 160}{729} = \frac{496}{729}.$$

(b) We can mimic the above argument to find the probability that $B$ wins the game without a deuce. That is, the probability of four straight points to $B$ is $(1/3)^4 = 1/81$, the probability that $A$ wins just one point in the game is $4(2/3)(1/3)^4 = 8/243$, and the probability that $A$ wins just two points is $10(2/3)^2(1/3)^4 = 40/729$. So the probability of $B$ winning without a deuce is $1/81 + 8/243 + 40/729 = 73/729$ and so the probability of deuce is $1 - 496/729 - 73/729 = 160/729$.

(c) **Either**: suppose deuce has been called. The probability that $A$ wins the set without further deuces is the probability that the next two points go $AA$ – with probability $(2/3)^2$.

The probability of exactly one further deuce is that the next four points go $ABAA$ or $BAAA$ – with probability $(2/3)^3(1/3) + (2/3)^3(1/3) = (2/3)^4$.

The probability of exactly two further deuces is that the next six points go $ABABAA$, $ABBAAA$, $BAABAA$ or $BABAAA$ – with probability $4(2/3)^4(1/3)^2 = (2/3)^6$.

Continuing this way, the probability that $A$ wins after three further deuces is $(2/3)^8$ and the overall probability that $A$ wins after deuce has been called is $(2/3)^2 + (2/3)^4 + (2/3)^6 + (2/3)^8 + \cdots$.

This is a geometric progression (GP) with first term $a = (2/3)^2$ and common ratio $(2/3)^2$, so the overall probability that $A$ wins after deuce has been called is $a/(1-r)$ (sum to infinity of a GP) which is:

$$\frac{(2/3)^2}{1-(2/3)^2} = \frac{4/9}{5/9} = \frac{4}{5}.$$

**Or (quicker!)**: given a deuce, the next 2 balls can yield the following results. $A$ wins with probability $(2/3)^2$, $B$ wins with probability $(1/3)^2$, and deuce with probability 4/9.

Hence $P(A \text{ wins} \mid \text{deuce}) = (2/3)^2 + (4/9)\, P(A \text{ wins} \mid \text{deuce})$ and solving immediately gives $P(A \text{ wins} \mid \text{deuce}) = 4/5$.

(d) We have:

$$\begin{aligned} P(A \text{ wins the game}) &= P(A \text{ wins without deuce being called}) \\ &\quad + P(\text{deuce is called})\, P(A \text{ wins} \mid \text{deuce is called}) \\ &= \frac{496}{729} + \frac{160}{729} \times \frac{4}{5} \\ &= \frac{496}{729} + \frac{128}{729} \\ &= \frac{624}{729}. \end{aligned}$$

Aside: so the probability of $B$ winning the game is $1 - 624/729 = 105/729$. It follows that $A$ is about six times as likely as $B$ to win the game although the probability of winning any point is only twice that of $B$. Another example of the counterintuitive nature of probability.

## F.3 Chapter 3 – Random variables

1. We require a counterexample. A simple one will suffice – there is no merit in complexity. Let the discrete random variable $X$ assume values 1 and 2 with probabilities 1/3 and 2/3, respectively. (Obviously, there are many other examples we could have chosen.) Therefore:

$$\begin{aligned}\mathrm{E}(X) &= 1 \times \frac{1}{3} + 2 \times \frac{2}{3} = \frac{5}{3} \\ \mathrm{E}(X^2) &= 1 \times \frac{1}{3} + 4 \times \frac{2}{3} = 3 \\ \mathrm{E}(1/X) &= 1 \times \frac{1}{3} + \frac{1}{2} \times \frac{2}{3} = \frac{2}{3}\end{aligned}$$

and, clearly, $\mathrm{E}(X^2) \neq (\mathrm{E}(X))^2$ and $\mathrm{E}(1/X) \neq 1/\mathrm{E}(X)$ in this case. So the result has been shown in general.

2. (a) Recall that $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X)^2)$. Now, working backwards:

$$\begin{aligned}\mathrm{E}(X(X-1)) - \mathrm{E}(X)(\mathrm{E}(X)-1) &= \mathrm{E}(X^2 - X) - (\mathrm{E}(X))^2 + \mathrm{E}(X) \\ &= \mathrm{E}(X^2) - \mathrm{E}(X) - \mathrm{E}(X)^2 + \mathrm{E}(X) \\ \text{(using standard properties of expectation)} &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 \\ &= \mathrm{Var}(X).\end{aligned}$$

(b) We have:

$$\begin{aligned}\mathrm{E}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) &= \frac{\mathrm{E}(X_1 + X_2 + \cdots + X_n)}{n} \\ &= \frac{\mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n)}{n} \\ &= \frac{\mu + \mu + \cdots + \mu}{n} \\ &= \frac{n\mu}{n} \\ &= \mu.\end{aligned}$$

$$
\begin{aligned}
\operatorname{Var}\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) &= \frac{\operatorname{Var}(X_1+X_2+\cdots+X_n)}{n^2} \\
\text{(by independence)} &= \frac{\operatorname{Var}(X_1)+\operatorname{Var}(X_2)+\cdots+\operatorname{Var}(X_n)}{n^2} \\
&= \frac{\sigma^2+\sigma^2+\cdots+\sigma^2}{n^2} \\
&= \frac{n\sigma^2}{n^2} \\
&= \frac{\sigma^2}{n}.
\end{aligned}
$$

3. Suppose $n$ subjects are procured. The probability that a single subject does not have the abnormality is 0.96. Using independence, the probability that none of the subjects has the abnormality is $(0.96)^n$.

   The probability that at least one subject has the abnormality is $1-(0.96)^n$. We require the smallest whole number $n$ for which $1-(0.96)^n > 0.95$, i.e. we have $(0.96)^n < 0.05$.

   We can solve the inequality by 'trial and error', but it is neater to take logs.

   $n\ln(0.96) < \ln(0.05)$, so $n > \ln(0.05)/\ln(0.96)$, or $n > 73.39$. Rounding up, 74 subjects should be procured.

4. (a) For the 'stupid' rat:

$$
\begin{aligned}
P(X=1) &= \frac{1}{4} \\
P(X=2) &= \frac{3}{4}\times\frac{1}{4} \\
&\vdots \\
P(X=r) &= \left(\frac{3}{4}\right)^{r-1}\times\frac{1}{4}.
\end{aligned}
$$

   This is a 'geometric distribution' with $\pi = 1/4$, which gives $\mathrm{E}(X) = 1/\pi = 4$.

   (b) For the 'intelligent' rat:

$$
\begin{aligned}
P(X=1) &= \frac{1}{4} \\
P(X=2) &= \frac{3}{4}\times\frac{1}{3} = \frac{1}{4} \\
P(X=3) &= \frac{3}{4}\times\frac{2}{3}\times\frac{1}{2} = \frac{1}{4} \\
P(X=4) &= \frac{3}{4}\times\frac{2}{3}\times\frac{1}{2}\times 1 = \frac{1}{4}.
\end{aligned}
$$

   Hence $\mathrm{E}(X) = (1+2+3+4)/4 = 10/4 = 2.5$.

(c) For the 'forgetful' rat (short-term, but not long-term, memory):

$$\begin{aligned}
P(X = 1) &= \frac{1}{4}\\
P(X = 2) &= \frac{3}{4} \times \frac{1}{3}\\
P(X = 3) &= \frac{3}{4} \times \frac{2}{3} \times \frac{1}{3}\\
&\vdots\\
P(X = r) &= \frac{3}{4} \times \left(\frac{2}{3}\right)^{r-2} \times \frac{1}{3} \quad \text{(for } r \geq 2).
\end{aligned}$$

Therefore:

$$\begin{aligned}
\mathrm{E}(X) &= \frac{1}{4} + \frac{3}{4} \times \left( \left(2 \times \frac{1}{3}\right) + \left(3 \times \frac{2}{3} \times \frac{1}{3}\right) + \left(4 \times \left(\frac{2}{3}\right)^2 \times \frac{1}{3}\right) + \cdots \right)\\
&= \frac{1}{4} + \frac{1}{4}\left(2 + \left(3 \times \frac{2}{3}\right) + \left(4 \times \left(\frac{2}{3}\right)^2\right) + \cdots \right).
\end{aligned}$$

There is more than one way to evaluate this sum.

$$\begin{aligned}
\mathrm{E}(X) &= \frac{1}{4} + \frac{1}{4} \times \left( \left(1 + \frac{2}{3} + \left(\frac{2}{3}\right)^2 + \cdots \right) + \left(1 + 2 \times \frac{2}{3} + 3 \times \left(\frac{2}{3}\right)^2 + \cdots \right) \right)\\
&= \frac{1}{4} + \frac{1}{4} \times (3 + 9)\\
&= 3.25.
\end{aligned}$$

Note that $2.5 < 3.25 < 4$, so the intelligent rat needs the least trials on average, while the stupid rat needs the most, as we would expect!

## F.4 Chapter 4 – Common distributions of random variables

1. Let $P \sim N(10.42, (0.03)^2)$ for the pistons, and $C \sim N(10.52, (0.04)^2)$ for the cylinders. It follows that $D \sim N(0.1, (0.05)^2)$ for the difference (adding the variances, assuming independence). The piston will fit if $D > 0$. We require:

$$P(D > 0) = P\left(Z > \frac{0 - 0.1}{0.05}\right) = P(Z > -2) = 0.9772$$

so the proportion of $1 - 0.9772 = 0.0228$ will *not* fit.

The number of pistons, $N$, failing to fit out of 100 will be a binomial random variable such that $N \sim \text{Bin}(100, 0.0228)$.

(a) Calculating directly, we have the following.

i. $P(N = 0) = (0.9772)^{100} = 0.0996$.

ii. $P(N \leq 2) =$
$(0.9772)^{100} + 100 \times (0.9772)^{99}(0.0228) + \binom{100}{2}(0.9772)^{98}(0.0228)^2 = 0.6005$.

(b) Using the Poisson approximation with $\lambda = 100 \times 0.0228 = 2.28$, we have the following.

i. $P(N = 0) \approx \mathrm{e}^{-2.28} = 0.1023$.

ii. $P(N \leq 2) \approx \mathrm{e}^{-2.28} + \mathrm{e}^{-2.28} \times 2.28 + \mathrm{e}^{-2.28} \times (2.28)^2/2! = 0.6013$.

The approximations are good (note there will be some rounding error, but the values are close with the two methods). It is not surprising that there is close agreement since $n$ is large, $\pi$ is small and $n\pi < 5$.

2. We have $P(X = 1) = P(X = 2) = \cdots = P(X = k) = 1/k$. Therefore:

$$\begin{aligned} M_X(t) = \mathrm{E}(\mathrm{e}^{Xt}) &= \mathrm{e}^t \times \frac{1}{k} + \mathrm{e}^{2t} \times \frac{1}{k} + \cdots + \mathrm{e}^{kt} \times \frac{1}{k} \\ &= \frac{1}{k}(\mathrm{e}^t + \mathrm{e}^{2t} + \cdots + \mathrm{e}^{kt}). \end{aligned}$$

The bracketed part of this expression is a geometric progression where the first term is $\mathrm{e}^t$ and the common ratio is $\mathrm{e}^t$.

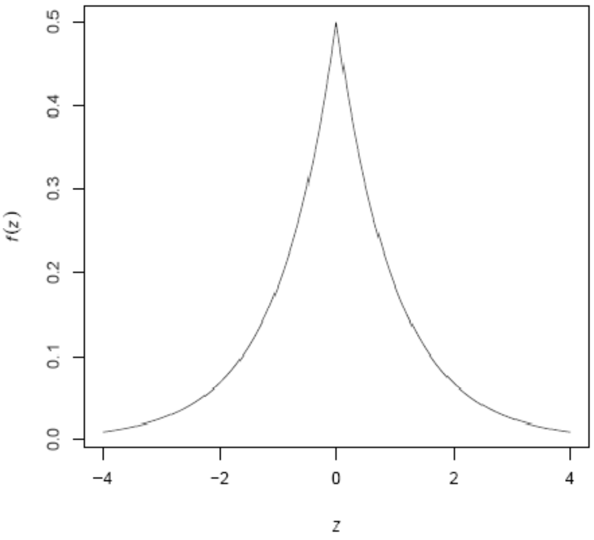
Using the well-known result for the sum of $k$ terms of a geometric progression, we obtain:

$$M_X(t) = \frac{1}{k} \times \frac{\mathrm{e}^t(1 - (\mathrm{e}^t)^k)}{1 - \mathrm{e}^t} = \frac{\mathrm{e}^t(1 - \mathrm{e}^{kt})}{k(1 - \mathrm{e}^t)}.$$

3. (a) For $f(z)$ to serve as a pdf, we require (i.) $f(z) \geq 0$ for all $z$, and (ii.) $\int_{-\infty}^{\infty} f(z)\,\mathrm{d}z = 1$. The first condition certainly holds for $f(z)$. The second also holds since:

$$\begin{aligned} \int_{-\infty}^{\infty} f(z)\,\mathrm{d}z &= \int_{-\infty}^{0} \frac{1}{2}\mathrm{e}^{-|z|}\,\mathrm{d}z + \int_{0}^{\infty} \frac{1}{2}\mathrm{e}^{-|z|}\,\mathrm{d}z \\ &= \int_{-\infty}^{0} \frac{1}{2}\mathrm{e}^{z}\,\mathrm{d}z + \int_{0}^{\infty} \frac{1}{2}\mathrm{e}^{-z}\,\mathrm{d}z \\ &= [\mathrm{e}^z/2]_{-\infty}^{0} - \left[\mathrm{e}^{-z}/2\right]_{0}^{\infty} \\ &= \frac{1}{2} + \frac{1}{2} \\ &= 1. \end{aligned}$$

A sketch of $f(z)$ is shown below.



(b) The moment generating function is:

$$
\begin{aligned}
M_Z(t) = \mathrm{E}(\mathrm{e}^{Zt}) &= \int_{-\infty}^{0} \frac{1}{2}\mathrm{e}^{-|z|}\mathrm{e}^{zt}\,\mathrm{d}z + \int_{0}^{\infty} \frac{1}{2}\mathrm{e}^{-|z|}\mathrm{e}^{zt}\,\mathrm{d}z \\
&= \int_{-\infty}^{0} \frac{1}{2}\mathrm{e}^{z}\mathrm{e}^{zt}\,\mathrm{d}z + \int_{0}^{\infty} \frac{1}{2}\mathrm{e}^{-z}\mathrm{e}^{zt}\,\mathrm{d}z \\
&= \int_{-\infty}^{0} \frac{1}{2}\mathrm{e}^{z(1+t)}\,\mathrm{d}z + \int_{0}^{\infty} \frac{1}{2}\mathrm{e}^{z(t-1)}\,\mathrm{d}z \\
&= \left[\frac{1}{2(1+t)}\mathrm{e}^{z(1+t)}\right]_{-\infty}^{0} + \left[\frac{1}{2(t-1)}\mathrm{e}^{z(t-1)}\right]_{0}^{\infty} \\
&= \frac{1}{2(1+t)} - \frac{1}{2(t-1)} \\
&= (1-t^2)^{-1}
\end{aligned}
$$

where the condition $-1 < t < 1$ ensures the integrands are 0 at the infinite limits.

(c) We can find the various moments by differentiating $M_Z(t)$, but it is simpler to expand it:

$$M_Z(t) = (1-t^2)^{-1} = 1 + t^2 + t^4 + \cdots .$$

Now the coefficient of $t$ is 0, so $\mathrm{E}(Z) = 0$. The coefficient of $t^2$ is 1, so $\mathrm{E}(Z^2)/2 = 1$, and $\mathrm{Var}(Z) = \mathrm{E}(Z^2) - (\mathrm{E}(Z))^2 = 2$.

The coefficient of $t^3$ is 0, so $\mathrm{E}(Z^3) = 0$. The coefficient of $t^4$ is 1, so $\mathrm{E}(Z^4)/4! = 1$, and so $\mathrm{E}(Z^4) = 24$.

Note that the first and third of these results follow directly from the fact (illustrated in the sketch) that the distribution is symmetric about $z = 0$.

4. For $X \sim \mathrm{Bin}(n,\pi)$, $P(X=x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}$. So, for $\mathrm{E}(X)$, we have:

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

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 $0,1,2,\ldots,n-1$ and probability parameter $\pi$.

Similarly:

$$\begin{aligned}
\mathrm{E}(X(X-1)) &= \sum_{x=0}^{n} x(x-1)\binom{n}{x}\pi^x(1-\pi)^{n-x} \\
&= \sum_{x=2}^{n}\frac{x(x-1)n!}{x!\,(n-x)!}\pi^x(1-\pi)^{n-x} \\
&= n(n-1)\pi^2\sum_{x=2}^{n}\frac{(n-2)!}{(x-2)!\,(n-x)!}\pi^{x-2}(1-\pi)^{n-x} \\
&= n(n-1)\pi^2\sum_{y=0}^{n-2}\frac{(n-2)!}{y!\,(n-y-2)!}\pi^{y}(1-\pi)^{n-y-2}
\end{aligned}$$

with $y = x-2$. Now let $m = n-2$, so:

$$\begin{aligned}
\mathrm{E}(X(X-1)) &= n(n-1)\pi^2\sum_{y=0}^{m}\frac{m!}{y!\,(m-y)!}\pi^{y}(1-\pi)^{m-y} \\
&= n(n-1)\pi^2
\end{aligned}$$

since the summation is 1, as before.

Finally, noting Practice question 2 in Chapter 3:

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

5. (a) A rate of 150 cars per hour is a rate of 2.5 per minute. Using a Poisson distribution with $\lambda = 2.5$, $P(\text{none passes}) = \mathrm{e}^{-2.5} \times (2.5)^0/0! = \mathrm{e}^{-2.5} = 0.0821$.

   (b) The expected number of cars passing in two minutes is $2 \times 2.5 = 5$.

   (c) The probability of 5 cars passing in two minutes is $\mathrm{e}^{-5} \times 5^5/5! = 0.1755$.

6. (a) Let $X$ denote the number of fish caught, such that $X \sim \mathrm{Pois}(\lambda)$. $P(X = 0) = \mathrm{e}^{-\lambda}\lambda^x/x!$ where the parameter $\lambda$ is as yet unknown, so $P(X = 0) = \mathrm{e}^{-\lambda}\lambda^0/0! = \mathrm{e}^{-\lambda}$.

   However, we know $P(X = 0) = \pi$. So $\mathrm{e}^{-\lambda} = \pi$ giving $-\lambda = \ln(\pi)$ and $\lambda = \ln(1/\pi)$.

   (b) James will take home the last fish caught if he catches $1, 3, 5, \ldots$ fish. So we require:

$$\begin{aligned}P(X=1) + P(X=3) + P(X=5) + \cdots &= \frac{\mathrm{e}^{-\lambda}\lambda^1}{1!} + \frac{\mathrm{e}^{-\lambda}\lambda^3}{3!} + \frac{\mathrm{e}^{-\lambda}\lambda^5}{5!} + \cdots\\ &= \mathrm{e}^{-\lambda}\left(\frac{\lambda^1}{1!} + \frac{\lambda^3}{3!} + \frac{\lambda^5}{5!} + \cdots\right).\end{aligned}$$

   Now we know:

$$\mathrm{e}^{\lambda} = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots$$

   and:

$$\mathrm{e}^{-\lambda} = 1 - \lambda + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} + \cdots.$$

   Subtracting gives:

$$\mathrm{e}^{\lambda} - \mathrm{e}^{-\lambda} = 2\left(\lambda + \frac{\lambda^3}{3!} + \frac{\lambda^5}{5!} + \cdots\right).$$

   Hence the required probability is:

$$\mathrm{e}^{-\lambda}\left(\frac{\mathrm{e}^{\lambda} - \mathrm{e}^{-\lambda}}{2}\right) = \frac{1 - \mathrm{e}^{-2\lambda}}{2} = \frac{1 - \pi^2}{2}$$

   since $\mathrm{e}^{-\lambda} = \pi$ above gives $\mathrm{e}^{-2\lambda} = \pi^2$.

## F.5 Chapter 5 – Multivariate random variables

1. (a) Recall that for any random variable $U$, we have $\operatorname{Var}(U) = \mathrm{E}(U^2) - (\mathrm{E}(U))^2$, $\mathrm{E}(kU) = k\,\mathrm{E}(U)$, where $k$ is a constant, and for random variables $U$ and $V$, $\mathrm{E}(U+V) = \mathrm{E}(U) + \mathrm{E}(V)$, and also $\operatorname{Cov}(U,V) = \mathrm{E}(UV) - \mathrm{E}(U)\,\mathrm{E}(V)$. We have:

$$\begin{aligned}
\operatorname{Cov}(X+Y, X-Y) &= \mathrm{E}((X+Y)(X-Y)) - \mathrm{E}(X+Y)\,\mathrm{E}(X-Y) \\
&= \mathrm{E}(X^2 - XY + YX - Y^2) - (\mathrm{E}(X) + \mathrm{E}(Y))(\mathrm{E}(X) - \mathrm{E}(Y)) \\
&= \mathrm{E}(X^2) - \mathrm{E}(Y^2) - (\mathrm{E}(X))^2 + \mathrm{E}(X)\,\mathrm{E}(Y) - \mathrm{E}(Y)\,\mathrm{E}(X) + (\mathrm{E}(Y))^2 \\
&= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 - (\mathrm{E}(Y^2) - (\mathrm{E}(Y))^2) \\
&= \operatorname{Var}(X) - \operatorname{Var}(Y)
\end{aligned}$$

as required.

(b) We have:

$$\begin{aligned}
\operatorname{Cov}(a+bX, c+dY) &= \mathrm{E}((a+bX)(c+dY)) - \mathrm{E}(a+bX)\,\mathrm{E}(c+dY) \\
&= \mathrm{E}(ac + adY + bcX + bdXY) - (a + b\,\mathrm{E}(X))(c + d\,\mathrm{E}(Y)) \\
&= ac + ad\,\mathrm{E}(Y) + bc\,\mathrm{E}(X) + bd\,\mathrm{E}(XY) \\
&\quad - ac - ad\,\mathrm{E}(Y) - bc\,\mathrm{E}(X) - bd\,\mathrm{E}(X)\,\mathrm{E}(Y) \\
&= bd\,\mathrm{E}(XY) - bd\,\mathrm{E}(X)\,\mathrm{E}(Y) \\
&= bd(\mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)) \\
&= bd\operatorname{Cov}(X,Y)
\end{aligned}$$

as required.

2. (a) We have:

$$\mathrm{E}\left(\sum_{i=1}^{k} a_i X_i\right) = \sum_{i=1}^{k} \mathrm{E}(a_i X_i) = \sum_{i=1}^{k} a_i\,\mathrm{E}(X_i).$$

(b) We have:

$$\operatorname{Var}\left(\sum_{i=1}^{k} a_i X_i\right) = \mathrm{E}\left(\left(\sum_{i=1}^{k} a_i X_i - \sum_{i=1}^{k} a_i\,\mathrm{E}(X_i)\right)^2\right) = \mathrm{E}\left(\left(\sum_{i=1}^{k} a_i (X_i - \mathrm{E}(X_i))\right)^2\right)$$

$$= \sum_{i=1}^{k} a_i^2 \mathrm{E}((X_i - \mathrm{E}(X_i))^2) + \sum_{1\le i\neq j\le n} a_i a_j\,\mathrm{E}((X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j)))$$

$$= \sum_{i=1}^{k} a_i^2 \operatorname{Var}(X_i) + \sum_{1\le i\neq j\le n} a_i a_j\,\mathrm{E}(X_i - \mathrm{E}(X_i))\,\mathrm{E}(X_j - \mathrm{E}(X_j)) = \sum_{i=1}^{k} a_i^2 \operatorname{Var}(X_i).$$

Additional note: remember there are two ways to compute the variance: $\text{Var}(X) = \text{E}((X-\mu)^2)$ and $\text{Var}(X) = \text{E}(X^2) - (\text{E}(X))^2$. The former is more convenient for analytical derivations/proofs (see above), while the latter should be used to compute variances for common distributions such as Poisson or exponential distributions. Actually it is rather difficult to compute the variance for a Poisson distribution using the formula $\text{Var}(X) = \text{E}((X-\mu)^2)$ directly.

3. (a) The joint distribution table is:

| | | $X = x$ | | |
|---|---|---|---|---|
| | | 0 | 1 | 2 |
| | 0 | 0 | $A$ | $2A$ |
| $Y = y$ | 1 | $A$ | $2A$ | $3A$ |
| | 2 | $2A$ | $3A$ | $4A$ |

Since $\sum\limits_{\forall x}\sum\limits_{\forall y} p_{X,Y}(x,y) = 1$, we have $A = 1/18$.

(b) The marginal distribution of $X$ (similarly of $Y$) is:

| $X = x$ | 0 | 1 | 2 |
|---|---|---|---|
| $P(X = x)$ | $3A = 1/6$ | $6A = 1/3$ | $9A = 1/2$ |

(c) The distribution of $X \mid Y = 1$ is:

| $X = x \mid y = 1$ | 0 | 1 | 2 |
|---|---|---|---|
| $P_{X\mid Y=1}(X = x \mid y = 1)$ | $A/6A = 1/6$ | $2A/6A = 1/3$ | $3A/6A = 1/2$ |

Hence $\text{E}(X \mid Y = 1) = (0 \times 1/6) + (1 \times 1/3) + (2 \times 1/2) = 4/3$.

(d) Even though the distributions of $X$ and $X \mid Y = 1$ are the same, $X$ and $Y$ are not independent. For example, $P(X = 0, Y = 0) = 0$ although $P(X = 0) \neq 0$ and $P(Y = 0) \neq 0$.

**lse.ac.uk/statistics**

Department of Statistics
The London School of Economics
and Political Science
Houghton Street
London WC2A 2AE

**Email: statistics@lse.ac.uk**

**Telephone:** +44 (0)20 7852 3709