Gabriela Czanner, Monika Kováčová, Silvester Czanner

Elements of Risk Analysis with Applications in R

SPEKTRUM
STU

ISBN 978-80-227-5341-8
Published by Slovak University Technology in Bratislava, Slovakia
SPEKTRUM STU Publishing

Contents

Preface **vii**

List of Figures **xiii**

List of Tables **xv**

Symbols **xvii**

I RISK FOUNDATIONS **1**

1 The big picture **3**
Gabriela Czanner and Silvester Czanner
1.1 Motivating examples 4
1.2 What is in the names? 5
1.2.1 Risk 5
1.2.2 Uncertainty 7
1.2.3 Risk management 10
1.3 To think and act like a risk analyst 12
1.3.1 Know how to describe and measure the risk 12
1.3.2 Know how to find an optimal decision 15
1.3.3 Know how to communicate the risks and decisions . . 17
1.4 A look ahead 17
1.4.1 Unit I Risk foundations 18
1.4.2 Unit II Data driven risk analysis 18
1.4.3 Unit III Data and model driven risk analysis 19
1.4.4 Unit IV Decisions under risk 19
1.4.5 Unit V Communication of risk 20
1.5 Summary 20
1.6 Further reading 21

II DATA DRIVEN RISK ANALYSIS **23**

2 Probability **25**
Gabriela Czanner and Silvester Czanner
2.1 The axioms of probability 26
2.2 Assigning probabilities to events: three approaches 27
2.3 Basic rules of probability 29

2.3.1 Joint probability 29
2.3.2 Marginal probability 31
2.3.3 Conditional probability 31
2.3.4 The product rule for joint probabilities 33
2.3.5 Independence of events 33
2.3.6 The law of total probability 34
2.3.7 Bayes' theorem 36
2.4 Tips to think and act like a risk expert 38
2.4.1 Consider using natural frequencies in communication . 39
2.4.2 Risks assessed via medical tests or medical AI 41
2.5 Summary 42
2.6 Further reading 42
2.7 R lab 43
2.8 Exercises 46

III DATA AND MODEL DRIVEN RISK ANALYSIS 51

3 Time series for risk quantification 53
Silvester Czanner and Gabriela Czanner
3.1 Motivation 54
3.1.1 Time series can be used for forecasting of the future . 55
3.1.2 The forecasting steps 58
3.1.3 What do stakeholders want? 61
3.2 Introduction to statistical theory of time series 66
3.2.1 Notation 66
3.2.2 Time series as a realisation of random process 67
3.2.3 Mean, variance, autocovariance and autocorrelation . 67
3.2.4 Stationarity 68
3.3 Exploratory data analysis for time series 69
3.3.1 Goal of exploratory data analysis 70
3.3.2 Trend, seasonal and cyclic components of time series . 70
3.3.3 Estimating the autocorrelation 75
3.4 Time series modelling and forecasting 82
3.4.1 Time series regression models 82
3.4.2 Exponential smoothing models of time series 85
3.4.3 Choosing the best-fitting model 95
3.4.4 Think ZINC! 99
3.4.5 Prediction intervals 103
3.5 Tips to think and act like a risk expert 109
3.5.1 Remember there is no such a thing as a free lunch! . . 109
3.5.2 Be a pro at visualising the risk and uncertainty 111
3.5.3 When relying on a model alone is a wrong idea 112
3.6 Summary 113
3.7 Further reading 114

3.8 R Lab . . . 115
3.9 Exercises . . . 138

4 Markov chains **143**
Gabriela Czanner and Monika Kovacova
4.1 Building a Markov Chain model . . . 144
4.1.1 Terminology . . . 145
4.1.2 The Markov property . . . 147
4.1.3 One-step transition probabilities . . . 149
4.1.4 Initial and one-step probability distributions . . . 150
4.1.5 Multistep transition probabilities . . . 153
4.1.6 Long-run prediction of the state . . . 154
4.2 Further topics on Markov Chains . . . 157
4.2.1 Commuter Cyril example . . . 157
4.2.2 Regular Markov chains . . . 159
4.2.3 Steady-state theorem . . . 159
4.2.4 Interpretation of the long run distribution . . . 160
4.2.5 Irreducible Markov Chains . . . 163
4.2.6 Periodic and aperiodic Markov chains . . . 165
4.3 Tips to think and act like a risk expert . . . 166
4.3.1 Not all sequences are Markov Chains but some can be turned into Markov Chains . . . 166
4.3.2 Sensitivity analysis . . . 168
4.4 Summary . . . 169
4.5 Further reading . . . 170
4.6 R lab . . . 170
4.7 Exercises . . . 177

IV DECISIONS UNDER RISK **183**

5 Decisions under precise risk and under imprecise risk **185**
Gabriela Czanner and Silvester Czanner
5.1 Making decisions under precise risk . . . 187
5.1.1 Measuring the risk with variance . . . 187
5.1.2 What are the preferences of people toward risk? . . . 190
5.1.3 Maximising expected utility of a decision maker . . . 195
5.1.4 Probability premium . . . 200
5.1.5 Monetary premium . . . 202
5.2 Making decisions under imprecise risk . . . 203
5.2.1 General strategy when probabilities are not known . . . 203
5.2.2 Alternative decision criteria . . . 204
5.3 Tips to think and act like a risk expert . . . 208
5.3.1 Be aware of the criticism of the alternative criteria . . . 208
5.4 Summary . . . 210
5.5 Further reading . . . 210

5.6 Exercises . . . 212

6 Decision trees **215**
Monika Kovacova and Gabriela Czanner
6.1 Decision tree . . . 216
6.1.1 Typical decision tree format . . . 216
6.1.2 The five steps of building a decision tree . . . 217
6.2 When probabilities are in a convenient format . . . 218
6.2.1 Example: Manufacturer Mahiro . . . 218
6.2.2 Solution step by step . . . 220
6.3 When probabilities are *not* in a convenient format . . . 225
6.3.1 Example: Investor Iveta . . . 226
6.3.2 Solution step by step . . . 227
6.4 Tips to think and act like a risk expert . . . 236
6.4.1 Sensitivity analysis for decision trees . . . 236
6.4.2 Value of information . . . 237
6.5 Summary . . . 237
6.6 Further reading . . . 238
6.7 Exercises . . . 239

V COMMUNICATION OF RISK **241**

7 Communication of risk **243**
Gabriela Czanner and Silvester Czanner
7.1 Lost in translation: Story 1 . . . 244
7.2 Lost in translation: Story 2 . . . 245
7.3 Objectives of the risk communication . . . 246
7.4 Stakeholders . . . 247
7.5 Brief foundations of risk communication . . . 248
7.5.1 Factors influencing risk perception . . . 248
7.5.2 Cognitive biases and heuristics . . . 249
7.5.3 Communication theories relevant to risk communication 250
7.6 Toward effective risk communication . . . 252
7.6.1 Practical elements of effective risk communication . . 252
7.6.2 Ethical and legal considerations in risk communication 253
7.6.3 Evaluation and improvement of risk communication . 253
7.7 Tips to think and act like a risk expert . . . 254
7.7.1 On risk communication in finance . . . 254
7.7.2 On risk communication in medicine . . . 254
7.7.3 On risk communication in Artificial Intelligence 255
7.8 Summary . . . 256
7.9 Further reading . . . 257

Bibliography **259**

Index **265**

Preface

Imagine facing a complex problem, such as choosing a future career, buying your first house, or investing your money. You may find it very difficult to think clearly about the decision; there are so many different things that you need to consider. You are tempted to go with your "gut" feelings and make your choice entirely based on emotional grounds. Just as you are about to do this, you consider how unfair this may be for your family, friends, or colleagues. What you desperately need at a time like this is a framework and a set of tools which will help you to address the complexities of the problem so that you can consider it dispassionately. You can do a risk analysis. In risk analysis, you think about your objectives, e.g., maximising a salary and being able to travel, and you think about how likely you are to get your dream job or not to get the job; you also think how likely your dream job will enable you to pay a mortgage, and you consider your attitude toward this risk, e.g., are you adventurous and risk loving?

Another example of a risk situation is the Covid-19 pandemic. Our prime minister is deciding whether to impose a lockdown, and he knows there is a risk if he does not impose a lockdown (e.g., virus spread, loss of lives) and there is a risk if he does impose a lockdown (e.g., loss of jobs, the economic instability of country and people). Another example is from medicine, where a medical doctor is deciding whether a patient is in remission, or in engineering, where an engineer is deciding whether an atomic power station is safe.

Hence we study risk because it is an important subject. Risk is omnipresent in every aspect of life. There are many types of risk, depending on the area of life or business: environmental risk; financial risk; health risk; health, safety, and environmental risks; information technology risk; insurance risk; occupational risk; safety risk; and security risk. In engineering, often risk analyses are used to show that an installation conforms to the requirements of a regulator.

Using the theory of probability, we can calculate probabilities of events. Using statistics, we model the data so we get probability estimates in more complex scenarios and so we can study risk factors. Hence, we understand what may affect the risks, and we can build on the communication tools of statistics to communicate the risk to stakeholders.

However, making an optimal decision under risk, or advising someone, involves knowing how to present risk information and communicate risk. It is crucial to know that there are various types of unknowns in our natural and human-built environments (including AI). This makes the risk subject matter complex and exciting to study, practice and teach.

Why teach risk analysis to students of applied informatics, engineering, mathematics, or computer science?

The idea for this book began with a special course at Liverpool John Moores University called "Probability and Risk" for 2nd year of undergraduate students of Applied Mathematics. Risk is a topic that is becoming so important that it needs to be part of the mainstream of applied mathematics and possibly other programs. Students of quantitative disciplines, such as mathematics, will likely have data science posts where they have to do two challenging tasks. Firstly they must quantify risks and decide what information they will use. Secondly, they will need to communicate the risk to the stakeholders so they can make the best decision that suits them. For example, they will communicate to a health official the health risks of a new virus so they can decide on lockdown, or to a clinician so that the clinician can advise a patient about the risks of surgery, or to an investor so she/he can decide about future investment.

However, we realised that no book is suitable to support the mathematics course. There are three types of available risk books. There are risk books that present the material in the context of a particular discipline, such as built environment engineering or cyber security, or epidemiology. Or there are risk books that focus on the communication of risk explaining the human perception of risk while explaining simple mathematical and statistical tools. Or there are risk books that focus on quantifying risk but are highly mathematically advanced and focused on graduate students. They typically provide a small space for the explanation of the risk communication. The tendency is that they explain a small number of advanced risk tools, thus not providing a comprehensive view of the risk discipline. Often, they do not give a sufficient explanation of how risk quantification relates to the newest developments in artificial intelligence.

This book reflects the current need for experts in risk quantification. There are two main needs. Firstly, the field of risk is rapidly evolving, thus requiring risk analysts who are flexible to adapt to new problems arising from growing data complexity and the growing number of new machine learning and artificial intelligence algorithms. This means that they need to have a solid understanding of mathematics, probability, and statistics to know the inner workings of the methods, why they work, and when they fail. So when a new statistical method arrives, they can learn it and judge if it suits the risk problem. Second, risk analysts need to be able to communicate with domain experts from various fields (such as economics, medicine, and cyber security). Communication means: understanding the risk problem in the domain context, understanding the quality of the information, and then communicating back the estimated risk. If the risk is estimated correctly but communicated badly, it can lead to disastrous consequences for the stakeholders. A challenge of risk communication is in human perceptions of the estimated probability

of something good or bad happening and in the fact that we humans have different perceptions of risk - some of us are OK to risk more, others want to stay on the safe side.

This book aims to provide an overall view and understanding of the key concepts of risk quantification and communication. To achieve this, the book starts with a necessary background in mathematics, probability, and statistics. Then the book follows with classical and modern concepts of quantitative risk analysis and making decisions under risk: from the frequentist and Bayesian statistics while providing connections to the recent developments in artificial intelligence. It shows examples and exercises from many life areas, including health, medicine, finance, investment, engineering, cyber security, and consumer behaviour. It provides examples in software R.

The targeted audience are students of undergraduate studies of mathematics and other quantitative discipline such as applied informatics, computer science, engineering, data science, or econometrics. This book is useful for those looking to find jobs as statisticians, data scientists, and risk consultants hence anyone interested in estimating and advising on risks.

Software used in this book

Modern risk analysis requires the use of a calculator or software. Software is needed for two reasons: computation (data exploration, inferential statistics, prediction) and simulation. We will focus on software R. R is a freeware statistical software package maintained by a core user group. The R base package and numerous add-ons are available at http://cran.r-project.org/.

Throughout this book, we will provide `R` code for both computations and simulation. It is not the goal, however, to serve as a primer in `R` language, so some prior knowledge of elementary `R` programming is required. The `R` has extensive help menus and active online user support groups. Readers interested in a more thorough treatment of the R software packages should consult *The R Book* by Michael J. Crawley [16]; alternatively, extra references are provided in relevant chapters.

Structure of the book

This book is aimed at the second year of an undergraduate degree in quantitative disciplines such as applied mathematics, applied informatics, engineering, or computer science. Part I is crucial to learn as it lays the basics of risk assessment. Part II brings the relevant topics from probability that are important

to know; however, they can be skipped if the reader has previous knowledge. Part III is important as it dives deep into specific areas of model-driven risk analysis: times series and Markov chains. We chose these two model-driven approaches as they are highly used in real life while they are the basis for more complex approaches implemented in Artificial Intelligence. Part IV discusses making decisions under precise risk, which is important to learn. Then it provides alternative decision criteria for making decisions under imprecise risk, which may not be relevant to some application domains. Part IV also introduces decision trees, an important topic, especially when making several decisions sequentially. Finally, Part V on risk communication is crucial for any risk analyst, and thus we dare the readers not to skip it.

What makes this book unique

This book has been crafted with many distinctive features. They are:

Inside-chapter feature **Examples** show how the concepts explained within the chapters are used and applied in real life. The examples describe the technical details of topics that are otherwise difficult to comprehend. The examples explain the technical, mathematical, and risk concepts. Using the examples, the reader will be able to develop the skills required to do risk analysis and to communicate with the whole risk assessment team and with stakeholders.

Inside-chapter feature **Caution!** is used to draw attention to important aspects of explained topics of risk analysis, to areas where we witnessed mistakes being made by risk analysts, by risk researchers, and by those who need to make decisions.

End-of-chapter feature **Tips to think and act like a risk expert** brings further tips for risk analysts. We chose them carefully based on our experience from doing research, reading research papers, talking to risk stakeholders such as clinicians, patients, and caregivers, and teaching this subject to undergraduate students of applied mathematics. Often these tips start with stories to introduce a problem, and then there is a discussion about solutions. These tips are here to nudge the reader to see risk analysis from a bird's view: (1) to see how the taught topics connect to other areas of technologies such as AI, (2) to see the advantages and disadvantages of taught approaches, and (3) to see how the taught topics interact with a human mind. Some of the tips are simple and lead to a straight correct answer. Some tips are touching more complex real-life situations where the correct answer depends on the interaction with and between stakeholders. Thus we crafted these tips to bring the risk analyst to the next level: to have holistic thinking about risk.

End-of-chapter feature **Further reading** is used to give the reader further books and papers to read on the topics from the chapter. It is also used to

discuss relevant extensions and more advanced topics - related to data science and artificial intelligence - and to recommend books and papers for the eager reader.

End-of-chapter feature **R lab** is used to give the reader the R code for all solved examples from the chapter, as well as the full output from R. We do not give any R code inside the chapters so that the reader can read the chapter undisturbed by R code explanations. The R lab section also contains further exercises to be solved using R. Their solutions are not provided in the book but can be provided upon request to instructors.

End-of-chapter feature **Exercises** gives the reader an opportunity to practice the topics. The use of a calculator is often needed, but R is not needed. Their solutions are not provided in the book but can be provided upon request to instructors.

Acknowledgements

The authors thank Dr Ivo Siekmann and Mr Vincent Kwasnica for valuable discussions on risk, uncertainty, decisions under risk, and a holistic approach to teaching. The authors would like to thank those students in the second year of the Applied Mathematics program at Liverpool John Moores University who, over five years, provided important feedback on the Risk course material. The authors would like to thank students and colleagues of the Faculty of Informatics and Information Technologies at the Slovak University of Technology, who conducted research with us and inspired several ideas in this book. We thank all research collaborators and colleagues from other universities, as well as patients who let us see how they perceive risk and recommendations from human experts and AI. We take full responsibility for any mistakes found.

In the authors' view, this book represents only one step in what can be done to teach quantification of risk to undergraduate students of quantitative programs. It strives to combine three elements: (1) basics of risk management needed for students of quantitative disciplines, (2) probabilistic and statistical tools for risk analysis with a link to data science and artificial intelligence, and (3) elements of risk communication so that stakeholders can make an informed decision that suits their preferences. This is so that such quantitative graduates can create sound recommendations for actions. We hope to do much more in the years to come, and this book will inspire others to do the same.

Liverpool, UK; Bratislava, Slovakia
Gabriela Czanner, Monika Kováčová, Silvester Czanner May 2023

List of Figures

1.1 Aleatoric and epistemic uncertainty 8
1.2 Risk management 11
1.3 The big picture for risk description 14

2.1 Example Boxes and jewels 30
2.2 Partition of space S into mutually exclusive and exhaustive events 35
2.3 Partition of set B by mutually exclusive and exhaustive events 35
2.4 Bayes' theorem 37

3.1 Overseas visits 1980-1981 55
3.2 Overseas visits 1980-2020 57
3.3 Overseas visits 1980-1981, with four projections 62
3.4 Overseas visits 1980-1981, with prediction intervals 63
3.5 Inflation forecast visualised in a fan chart 65
3.6 Time series examples 71
3.7 Overseas visits 1980-2020, with estimated trend 74
3.8 Overseas visits, with time series decomposition 76
3.9 Overseas visits, with ACF for January 1980 - July 1980 . . . 80
3.10 Overseas visits, with ACF for January 1980 - December 2020 81
3.11 Kings' life span data 83
3.12 Kings life span data, with linear model 84
3.13 Kings life span data, with quadratic model 85
3.14 Daily temperatures time series data 86
3.15 Daily temperatures data, with SES for $\alpha = 0.6$ 90
3.16 Daily temperatures data, with SES at various α values 91
3.17 Daily temperatures data, with SES and SSE at various α values 92
3.18 Daily temperatures data, with SES and Holt smoothing . . . 94
3.19 Overseas visits, with two fitted HW smoothing models 96
3.20 Kings life span data, with quadratic model and goodness of fit 101
3.21 Overseas visits 1980-1981, with a multiplicative HW model and goodness of fit analysis 102
3.22 Overseas visits 1908-1981, with the additive HW model and goodness of fit analysis 103
3.23 Kings life span data, with forecasts 107
3.24 Overseas visits 1980-1981, with HW prediction intervals . . . 109

3.25 Four-time series showing patterns typical of business and economic data 140

4.1 Cereals buyer's state-transition diagram 145
4.2 Illustration of the timeline in Markov Chain 147
4.3 Markov property 148
4.4 Commuter Cyril's one-step transition diagram 158
4.5 Bus driver Lubos's one-step transition diagram 163
4.6 Cereals market shares, initial distribution 0.25 and 0.75 . . . 172
4.7 Cereals market shares, initial distribution 0.0001 and 0.9999 . 173
4.8 Cereals market shares, initial distribution 5/6 and 1/6 176

5.1 Makovnik Bakery weekly profits 188
5.2 Utility functions of risk-averse, loving and neutral people . . . 191
5.3 Diminishing marginal utility 192
5.4 Utility of a person with constant risk aversion 194
5.5 Utility of a person with decreasing risk aversion 194
5.6 A piston 196

6.1 Decision tree format 217
6.2 Decision tree notation 217
6.3 Manufacturer Mahiro's decision tree, after we did steps 1 and 2. 219
6.4 Manufacturer Mahiro's decision tree, after we did steps 1-3 . 220
6.5 Manufacturer Mahiro's decision tree, processing nodes 6 and 9 222
6.6 Manufacturer Mahiro's decision tree, processing node 5 . . . 223
6.7 Manufacturer Mahiro's decision tree, processing node 3 . . . 224
6.8 Manufacturer Mahiro's decision tree, processing node 1 . . . 225
6.9 Investor Iveta's decision tree 227
6.10 Investor Iveta's decision tree after step 1 228
6.11 Investor Iveta's decision tree, after step 1 229
6.12 Investor Iveta's probabilities calculation. 230
6.13 Investor Iveta's decision tree, with Steps 1-2 completed 232
6.14 Investor Iveta's decision tree, with steps 1-3 completed. . . . 233
6.15 Investor Iveta's decision tree, working on step 4, random nodes 234
6.16 Investor Iveta's decision tree, working on step 4, decision nodes 235
6.17 Investor Iveta's decision tree, working on step 4 236

List of Tables

1.1 Kidney stone treatment example data. 4
1.2 You and your friend example. 5
1.3 Risk description table. 15

2.1 Measles new cases 28
2.2 Table of frequencies in boxes and jewels example 30

3.1 Monthly visitors time series for January 1980 - December 1981. 56
3.2 Lagged time series. 77
3.3 Autocorrelation. 78
3.4 Calculations of autocorrelation at lag 2. 78
3.5 Weights for four values of α in simple exponential smoothing. 88
3.6 Daily temperatures data, with SES and SSE. 92
3.7 Kings life span data, summary for two regression models. . . 99
3.8 Overseas visits 1980-1981, summary for HW model. 99

4.1 Cereals buyer Bob's Markov chain realisation 146
4.2 Cereals buyer Bob's initial state, distribution 0.25 and 0.75 . 150
4.3 Cereals buyer Bob's initial state, distribution 0.3325 and 0.6675 151
4.4 Commuter Cyril's one-step transition probabilities 158

5.1 Decisions, states and utilities 195
5.2 Manufacturer Peng's table for utilities and consequences. . . 195
5.3 Manufacturer Peng's utilities, in Piston example 196
5.4 Manufacturer Peng's expected utilities, in Piston example. . . 197
5.5 Stock investment example 198
5.6 Stock investment example's utilities. 199
5.7 Stock investment utilities for investors I and II, when C=10. . 199
5.8 Stock investment utilities for investors I and II, when C=100. 200
5.9 Expected stock investment utilities. 201
5.10 Insurance example. 202
5.11 Insurance example. 202
5.12 Filip's monetary profits in food warehouse example 205
5.13 Filip's utilities in food warehouse example 205
5.14 Filip's utilities in food warehouse example, and Laplace criterion 205

5.15 Filip's utilities in food warehouse example, and Max-min (Wald) criterion 206
5.16 Filip's utilities in food warehouse example, and Max-max criterion 206
5.17 Filip's utilities in food warehouse example, and Hurwicz criterion 207
5.18 Filip's regrets in food warehouse example 208
5.19 Filip's regret in food warehouse example, choosing optimal decision 208
5.20 Criticism of Max-min criteria, when u gets large 209
5.21 Criticism of Min-max criteria, when choosing from two decisions 209
5.22 Criticism of Min-max criteria, when adding another decision . 209

6.1 Manufacturer Mahiro's information on competition 219
6.2 Manufacturer Mahiro's decision tree's terminal node profits . 221
6.3 Investor Iveta's profits 226
6.4 Investor Iveta's expected utilities 226
6.5 Investor Iveta's probabilities 232

7.1 Risk communication, 7 steps 255

Symbols

Symbol Description

PRA	Probabilistic Risk Analysis
PSA	Probabilistic Sensitivity Analysis
QRA	Quantitative Risk Analysis
A	Outcomes (states, events)
C	Consequences
Q	Quantification of uncertainties (measurement or description)
K	Knowledge
SoK	Strength of knowledge
U	Utility, utility function
$u(x)$	Utility function
D, d	Decisions
ML	Machine Learning
AI	Artificial Intelligence
S	Sample space of all outcomes
P	Probability
E	Expected value of a random variable
V	Variance of a random variable
OLS	Ordinary Least Squares
l	likelihood
α	Level of significance
SSE	Sum of squared errors
RMSE	Root mean squared errors
AIC	Akaike Information Criteria
BIC	Bayesian Information Criteria
R^2	R-squared statistic
R^2_{adj0}	R-squared adjusted statistic
ZINC	Four criteria of goodness-of-fit of a model to a data
CI	Confidence interval
PI	Prediction interval
$\hat{y}_{T+h\|T}$	h-step forecast when using all data up to time T
$Q_1, .., Q_4$	Quarters of a year
Cov	Covariance
$\rho_{t,s}$	Correlation between measurements from time t and s
ρ_k	Autocorrelation at lag k.
$\hat{\rho}_k$	Estimate of the correlation between measurements that are k time steps apart
EDA	Exploratory data analysis
ACF	Autocorrelation function
SES	Simple Exponential Smoothing
Holt	Holt's Exponential Smoothing
HW	Holt-Winter's Exponential Smoothing
Y_t, X_t	Random variable measured at time t
s_n	State of a random variable at time t
$p^{(0)}$	Probability distribution at time 0
$\boldsymbol{P}$	One-step probability transition matrix
$p^{(k)}_{i,j}$	Probability of transiting from i to j in k steps
CV	Coeficient of variation
VaR	Value at Risk
EMV	Expected Monetary Value

Part I

RISK FOUNDATIONS

1

The big picture

Gabriela Czanner

Silvester Czanner

CONTENTS

1.1 Motivating examples 3
1.2 What is in the names? 5
1.2.1 Risk 5
1.2.2 Uncertainty 7
1.2.3 Risk management 10
1.3 To think and act like a risk analyst 11
1.3.1 Know how to describe and measure the risk 12
1.3.2 Know how to find an optimal decision 15
1.3.3 Know how to communicate the risks and decisions 17
1.4 A look ahead 17
1.4.1 Unit I Risk foundations 17
1.4.2 Unit II Data driven risk analysis 18
1.4.3 Unit III Data and model driven risk analysis 19
1.4.4 Unit IV Decisions under risk 19
1.4.5 Unit V Communication of risk 20
1.5 Summary 20
1.6 Further reading 21

Learning objectives

1. To explore what is meant by risk and get an overall view of how people and organisations make decisions when facing risk.

2. To explore situations where risk can be calculated wrongly, thus leading to wrong decisions.

3. To learn about the general process of making decisions under risk and uncertainty.

1.1 Motivating examples

Here we discuss two examples and illustrate what can go wrong.

Example. Kidney stone treatment. This comes from a real-life medical study comparing the success rates of two treatments for kidney stones. The table below shows the success rates and numbers of treatments for small and large kidney stones. The term success rate here actually means the success proportion. Treatment A includes open surgical procedures, and Treatment B includes closed surgical procedures. The numbers in parentheses indicate the number of success cases over the total size of the group. Which treatment is more effective?

		Treatment	
		Treatment A	Treatment B
Stone size	Small	Group 1, 93% (81/87)	Group 1, 87% (234/270)
	Large	Group 3, 73% (192/263)	Group 4, 69% (55/80)
	Both	78% (273/350)	83% (289/350)

TABLE 1.1: Kidney stone treatment example data.

The paradoxical conclusion is that treatment A is more effective when used on small stones, and also when used on large stones, yet treatment B appears to be more effective when considering both sizes at the same time. In this example, the "lurking" variable (or confounding variable) causing the paradox is the size of the stones, which was not previously known to researchers to be important until its effects were included.

Based on these effects, the paradoxical result is seen to arise because the effect of the size of the stones overwhelms the benefits of better treatment (A). In short, the less effective treatment B appeared to be more effective because it was applied more frequently to the small stones cases, which were easier to treat. Which treatment is better, A or B?

Note that this is an example of so-called *Simpson's paradox*, which also goes by several other names, is a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined. It is also referred to as *Simpson's reversal*, *Yule–Simpson effect*, *amalgamation paradox*, or *reversal paradox* also called *Ecological fallacy*. Simpson's paradox happens when groups of data show one particular trend; however, this trend is reversed when the groups are combined together. Understanding and identifying this paradox is important for correctly interpreting data.

Example: You and your friend. You and your friend each do problems

on Brilliant, and your friend answers a higher proportion correctly than you on each of the two days. See the Table 1.2. Who is better, you or your friend?

		Person	
		You	Your friend
Day	Saturday	87.5% (7/8)	100% (2/2)
	Sunday	50% (1/2)	62.5% (5/8)
	Total	80% (8/10)	70% (7/10)

TABLE 1.2: Your and your friend example.

On Saturday, you solved 7 out of 8 attempted problems, but your friend solved 2 out of 2. You had solved more problems, but your friend pointed out that he was more successful, since $7/8 < 2/2$. This seems a fair evaluation for Saturday. On Sunday, you only attempted 2 problems and got 1 correct. Your friend got 5 out of 8 problems correct. Your friend says he/she was better on Sunday, since $1/2 < 5/8$. However, the competition is about the one who solved more accurately over the weekend, not on individual days. Overall, you have solved 8 out of 10 problems whereas your friend has solved 7 out of 10 problems. Hence, even though your friend solved a higher proportion of problems each day, you actually won the challenge by solving the higher proportion for the entire weekend. This seemingly unintuitive possibility is again an instance of Simpson's paradox.

1.2 What is in the names?

There is a basic terminology that is used when talking about risk. Some terminology is specific to some fields; some are general. We will start by defining the main terms and provide examples, to illustrate what they mean.

1.2.1 Risk

The term risk does not have a unique worldwide accepted definition. The literature on the subject of risk has grown rapidly in the last 20 years, and the word "risk" is used in many different ways. The purpose of this section is to discuss briefly what we mean by risk and in what way the concept has been established in a mathematical setting. The following are the four typical definitions of risk:

1. Risk is a *situation involving exposure to danger* [3]. Risk is the possibility of something bad happening. When the risk is present, we know all the outcomes (effects, implications) of an activity (or inactivity) with respect to

something that humans value (such as health, well-being, wealth, property, or the environment), and we know the probabilities of such outcomes. In other words, Risk is the situation under which the decision outcomes and their probabilities of occurrences are known to the decision-maker.

2. An old definition of risk is *source of harm or hazard*. This definition comes from Blount's "Glossographia" [11]. Modern equivalents refer to "unwanted events" [4] or "something bad that might happen" [1].

3. Another definition of risk is *chance of harm*. This definition comes from Johnson's "Dictionary of the English Language" [2]. It has also been paraphrased as "possibility of loss" [3] or "probability of unwanted events". Here, the risk is defined as the probability of harm occurring.

4. Another definition of risk is any *uncertain event or set of circumstances* that, should it occur, would affect the ability to meet objectives". Note that here the risk is defined as an event.

Examples of risk:

1. Medical risk: A patient (or his carer) wants to know if he has glaucoma disease in one of his eyes. He/she may ask: Do I have glaucoma? If I have it, then if I do not get treated, when will I lose sight? If I get treated, are there chances of side effects?

2. Natural hazard-triggered technological accidents (known as Natechs) are a subject of increasing concern. They are the industrial accidents resulting from natural hazard events. The growing concern is due to the growing exposure of highly industrialised and urbanised areas to natural hazards [57].

3. Financial risk: A financial institution may like to know about the possibility of the growth of the Gross Domestic Product index going up or down. They may ask: What are the chances that the next year's growth rate will be between 1.5% and 3.5% (stable economy)? What are the chances that the growth will be less than 1.5% in each of the next three quarters? What are the chances that within a year, there will be negative growth (recession)?

4. Spacecraft flight risk: Problem of assessing the risk for a spacecraft with a specific mission. For example, the Apollo and Shuttle projects plan to send astronauts to Mars. When preparing for such flights, risk consideration is crucial. However, the problem is much different from smoking risk or risk of common cancers because there are no relevant available data. Here statistical and mathematical models are used (called *model-based risk assessment*) as experience in the form of observations of the performance of the whole spacecraft is not available in the planning stage [9].

5. Legal: What is the chance that the suspect committed the crime?

6. Safety: What are the chances that the bridge will collapse? What are the chances that a clutch will break in the car?

7. Epidemiology: What are the chances of hospitals being overfilled if we do not roll out a national lockdown for Covid-19?

8. Cyber security: What are the chances of a cyber attack in the next hour?

9. Credit risk: What are the chances that a person will not repay the loan?

10. Artificial Intelligence (AI)-related risk: What are the chances that AI incorrectly identifies an ill person as healthy?

1.2.2 Uncertainty

The Oxford English Dictionary states that "uncertainty is the state or character of being uncertain in mind, a state of doubt, want of assurance or confidence, hesitation, irresolution" [3]. The Cambridge Dictionary states that "uncertainty is a situation in which something is unknown or certain" [1]. These two definitions are similar in the sense that they say something is not known. Note that the Oxford dictionary explicitly says that uncertainty is the state of mind, while Cambridge conveys the same message albeit implicitly.

Two main types of uncertainty are aleatoric and epistemic uncertainty [23] (see Figure 1.1). Next, we look at each type and discuss some examples.

Aleatoric uncertainty occurs when the probabilities are quantifiable to a high degree of certainty, but we do not know the outcomes. In other words, we know the probabilities precisely. However, there is uncertainty about the outcome due to randomness (random chance). Such aleatoric uncertainty is an irreducible uncertainty. Examples are:

1. We flip a coin. We know it is a regular coin, so the probability of a head is 50% and a tail is 50%. The aleatoric uncertainty is not knowing the outcome of the flip (or of the next five flips).

2. The manufacturing line is producing gloves and has a probability of 1 in 200 producing a faulty product. The aleatoric uncertainty is not knowing the quality of the next produced product or a randomly chosen product.

3. We flip a different coin. We know it is an irregular coin, and a trusted source told us that the probability of a head is 72% and a tail is 28%. The trusted source is a keen statistician who previously flipped the coin 300 times and got the probabilities for us, which are quite precise since 300 is a large number. The aleatoric uncertainty is not knowing the outcome of the flip (or of the next five flips).

Epistemic uncertainty occurs when there is a lack of knowledge. The word "epistemic" means relating to knowledge or to the degree of its validation. This uncertainty is reducible. Examples are:

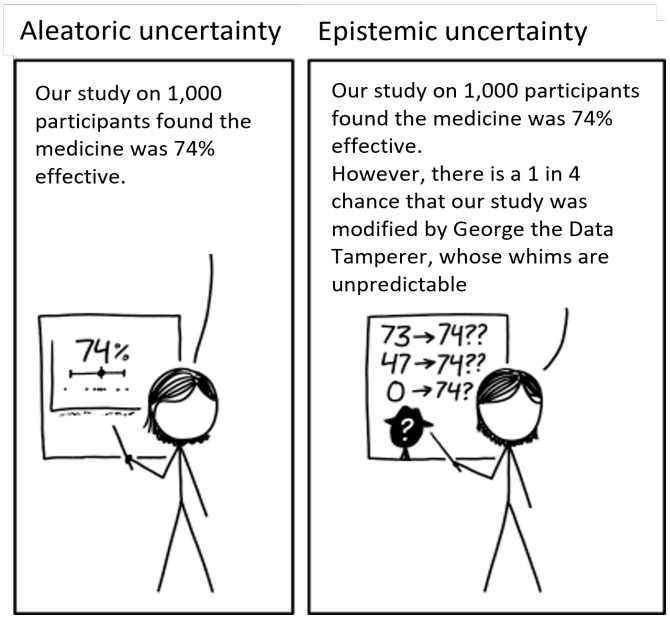


FIGURE 1.1: Aleatoric and epistemic uncertainty. In the left scenario, there is a risk that the medicine does not work for some patients. We (the risk analyst) are certain that the medicine does work for 74% of patients because we had a very large sample and we are confident that the data are of good quality. So there we have an aleatoric uncertainty about for whom the medicine will work, but there is no uncertainty about the proportion 74%. On the right, there is a different scenario. There is still the risk that the medicine does not work for some patients. However, we have an additional uncertainty as we are not certain about the probability of 74% anymore, thus we question our knowledge of the probability. Where there is an epistemic uncertainty too, in our view of this situation.

1. When President Obama's team was deciding whether to raid the hideout of bin Laden, they had information from their intelligence team that the probability of him being in the compound is 51%. Obama's team was facing an epistemic uncertainty on whether bin Laden is in the compound [23]. They had some estimate of how likely he is there, but it was just an estimate so the quality of the estimate was questionable, and the number was not 100% certainty, but rather a smaller number, that is 51%.

2. It is August 2020 and the Bank of England wants to predict future inflation. Past data are considered to be used but there is uncertainty about the suitability of the data to predict the future, as now the country (and the whole wide world) is in the middle of an unprecedented fight with the Covid-19 pandemic which makes product and labour markets unpre-

dictable as well as the government's policies. There is epistemic uncertainty on how to do the prediction.

3. A medical team has data on patients who developed gastric cancer as well as those who did not develop the cancer. The data science team uses the data and develops a risk prediction model. For each patient, the model estimates the risk of developing gastric cancer a year from now. But then the data science team finds out that data have many missing values! For example, men were more likely to not report their smoking and drinking status (important risk factors for this cancer), and those who feel well were more likely to decide not to participate in the study. This may have brought biases into the risk prediction model! There is now epistemic uncertainty about the risk model.

4. Two months ago, company PlanetProtect bought and installed a new AI software to detect early digital attacks on their IT system. They want to know how accurate it is. The seller claims an accuracy of 90%, but the seller says that they did not test it on companies like PlanetProtect and thus the accuracy may change "a bit" (though we may be sceptical about the "a bit" part). There is now epistemic uncertainty about the accuracy of the AI system for PlanetProtect. So PlanetProtect decides to reduce this uncertainty. They collect data from two months. The data are used to estimate the probability of detecting the attack being 76% with a 95% confidence interval (56% and 96%). They believe no one tampered with the data. But the data are too short to have a precise estimate of the probability. The current estimate of the probability is therefore imprecise with the width of the confidence interval of 40%. This is again an epistemic uncertainty about the accuracy of the new AI software to detect attacks early. More data will help to reduce this epistemic uncertainty.

Epistemic uncertainty is personal and temporal [19]. For example, the uncertainty on how accurate the AI software that PlanetProtect bought to detect digital attacks on their AI system. Epistemic uncertainty is personal because it relates to each person's state of mind whether it is based on a cognitive process of an expert assessor, a hunch feeling of a non-expert individual, or based on some statistical calculation (from t-test all the way to a complex AI algorithm), or combination of all. It is temporal because it can and should be updated as a piece of new information becomes available. The task of a risk expert is to express such uncertainty at the time the uncertainty was obtained. Therefore there is no single "true" uncertainty.

Aleatoric uncertainty is a property of the real world, referring to real differences between the members of a population of real-world entities. The term population refers either to biological organisms such as people or to non-biological items such as the activity of the IT system during specific 10 minutes, in the company PlanetProtect. If we chop the time into 10 minutes intervals, then each interval is an entity of the real world and each either was

under digital attack (value 1 or Yes) or not under attack (value 0 or Not). Such differences in values across the time intervals represent the aleatoric uncertainty (also called variability of the real world, or uncertainty caused by variability).

In our research practice we witnessed that dialogue with non-risk experts improved once we agreed at the start "the uncertainty of what we are discussing?". The communication also improved once we started using the terms epistemic and aleatoric uncertainty. Thus in our book, we will use the terms epistemic uncertainty and aleatoric uncertainty. When we say uncertainty, we will mean any of the two types: epistemic or aleatoric.

Caution! When reading reports, books, or papers on uncertainty, it is important to judge if they speak about aleatoric or epistemic uncertainty. Sometimes they say it at the beginning of their report, sometimes they do not say so the reader needs to deduct it from the context. For example, in [19] the authors use the term uncertainty when they talk about its specific type the epistemic uncertainty and they do acknowledge it.

1.2.3 Risk management

Risk management refers to a systematic approach to managing risks, and sometimes to the profession that does this. A general definition is that risk management consists of "coordinated activities to direct and control an organization with regard to risk" [55].

1. *Planning, establishing the scope, context, and criteria.* Problem issue definition. Clarify who the stakeholders are (patient, public, caregivers, business owner, government, neighbouring countries...). Set study objectives. Establish relevant principles and approaches.

2. *Risk assessment: Risk analysis and Risk evaluation. Risk analysis* is recognising and characterising risks, cause analysis, consequence analysis, risk characterisation, deciding on measures of risk, studying and comparing alternative decisions with respect to risk, communicating and consulting the risks. *Risk evaluation* is judging the significance of the risks to support decision-making, ranking alternative actions to support decision-making, communicating and consulting the alternative actions.

3. *Use of risk assessment for decision making* This includes selecting and implementing options for addressing risk, monitoring and reviewing, recording, and reporting.

In general, the aim of risk management is to assist organisations in setting strategies, achieving objectives, and making informed decisions. The outcomes should be scientifically sound, cost-effective, integrated actions that treat risks while taking into account social, cultural, ethical, political, and legal considerations.

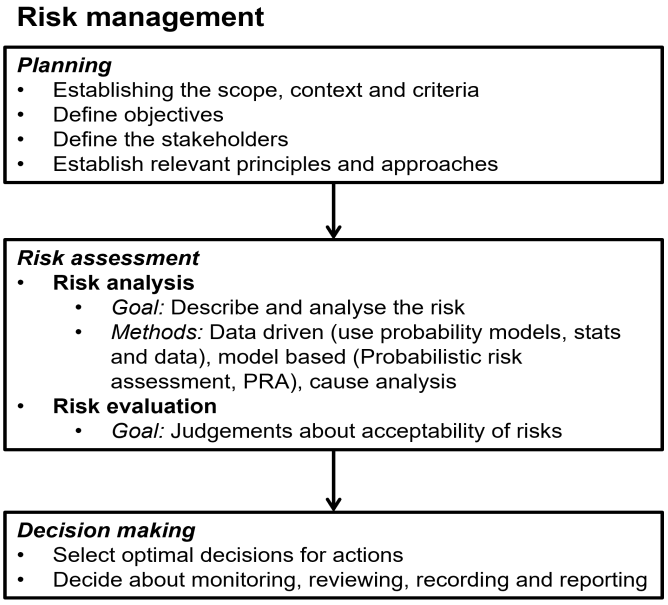


FIGURE 1.2: Risk management.

In contexts where risks are always harmful, risk management aims to reduce or prevent risks. In the safety field, it aims to protect employees, the general public, the environment, and company assets, while avoiding business interruptions (e.g. nuclear power stations can only have harmful risks, also the sight of a person can only get worse).

For organizations whose definition of risk includes upside (some positive outcome) as well as downside (negative outcome) risks, risk management is as much about identifying opportunities as avoiding or mitigating losses. It then involves getting the right balance between innovation and change on the one hand, and avoidance of shocks and crises on the other (e.g. value of a house can quite randomly go up or down).

Risk assessment is a systematic approach consisting of two components: (1) risk analysis i.e. recognising, characterising, and recording risks, this can include identifying the causes and their potential consequences, (2) risk evaluation i.e. evaluating the risk significance, in order to support decisions about how to manage them. In safety contexts, where risk sources are known as hazards, this step is known as hazard identification.

1.3 To think and act like a risk analyst

Next, we discuss three key domains of a risk analyst: (1) Knowing how to describe and measure risk, (2) knowing how to find the optimal decision, and (3) knowing how to communicate the risk and recommend the optimal decision.

1.3.1 Know how to describe and measure the risk

The description and measurement of risk is about developing an understanding of the risk. It can be done qualitatively or quantitatively. The *quantitative risk analysis* (QRA) is also called *probabilistic risk analysis* (PRA). Formally, in a risk analysis, we are attempting to envision how the future will turn out if the decision maker undertakes a certain course of action (or inaction). The risk analysis answers the following four questions:

- **What can happen?** What can go wrong? What can go well? Can we create a list of all outcomes that can happen? What are the positive outcomes and negative outcomes?
- **For each outcome, if it does happen, what are the consequences?** What are the benefits or losses for each stakeholder? Some consequences can be small, some large. For example, an earthquake can have catastrophic consequences albeit with a small probability. Some consequences can be positive, some negative. *How outcomes link to consequences?* This may not be known fully, so a statistical analysis of causes may be done, or experts can be called for their opinion. *Who are the stakeholders, i.e. people affected by outcomes?* In the example of Janette getting a job, she is a stakeholder, additionally, her parents are stakeholders because if she gets a well-paid job, her parents do not need to support her as much or at all. Each stakeholder has some attitude toward each consequence; in future, we will call such attitudes the utilities.
- **How likely is each outcome to happen?** What is the likelihood or probability of the wrong outcome happening? What is the probability of a positive outcome? Some probabilities can be large, some very little. Is probability a useful way to measure uncertainty? There are other measures that we will learn later.
- **What is the strength of knowledge that we have?** What do we use to calculate the probabilities of events? Do we use past data or expert knowledge? Do we trust the expert(s)? Are data relevant? Do we have enough data? Are data noisy?

The four answers to the questions above constitute four elements of risk,

called *risk descriptors* (see Figure 1.3). The first three questions are the triplet of Kaplan and Garrick [37]). The fourth question was added later [9].

From the practical point of view, to describe the risk, it is highly recommended to construct a table that describes all four risk components (see Table 1.3). The i-th row of the table should contain the answers to the above four questions.

$$(A_i, C_i, Q_i, K_i) \text{ for } i = 1, 2, ..., N, \tag{1.1}$$

where

- A_i is a *outcome* where N is the number of all outcomes that can happen. This is also termed event, hazard, threat, opportunity, or risk source. In the example where Janette is looking for a job, one outcome is her getting the new job of her dreams with a high salary.

- C_i is the *consequence* or the evaluation measure of that outcome. In general, the consequences can be positive or negative. In Janette's looking for a job example, if she gets her dream job with a high salary, she will be happy and will buy the house of her dreams; if she gets a job that pays less, she will not be so happy, and if she ends up jobless then as a consequence she will need to live with her parents.

- Q_i is the *quantification (measurement or description) of uncertainty* of the outcome. For example, it can be the probability of getting my dream job, the probability of getting another than a dream job, and the probability of being jobless. There are other ways to measure or describe uncertainty, which we will discuss in later chapters.

- K_i is the *knowledge* that we used to list all outcomes and consequences and used for measurement of uncertainty - as well as the *strength of knowledge* (SoK). For example, in order to estimate the probability we use historical data or data obtained via the experiment (e.g. Latin squares design experiment, or completely randomised design study). Such data help us to estimate the probabilities above (e.g. the probability of Janette landing a dream job with a high salary). Sometimes, we cannot do the experiments (for ethical or practical reasons), and hence we cannot collect data. So we seek expert opinion. For example, it is hard to think of an experiment that will give us data to estimate the probability of Janette landing a dream job, as she only lives once, and like every person, she is unique with unique sets of talents and unique way of performing at a job interview. So, in this situation, we (or Janette) may ask for an expert opinion or even ask several experts, like trusted friends and mentors. In both situations, whether we collect data or ask experts in order to estimate the probabilities, we need to make a note of how we made all the estimates of probability and think if we trust the data source and if we trust the expert. Any uncertainty in

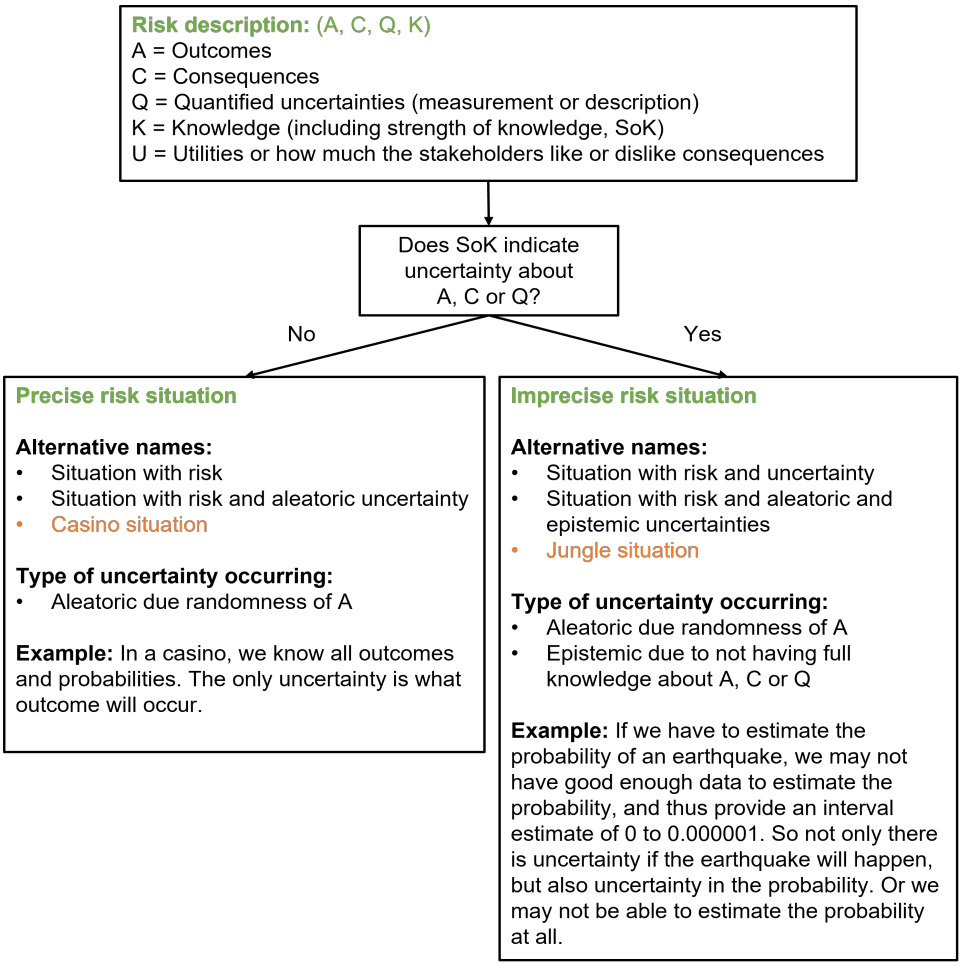


FIGURE 1.3: The big picture for risk description. It addresses four risk questions: A, C, Q, and K. It lists all sources of uncertainty: aleatoric uncertainty due to randomness as well as epistemic uncertainty due to weak knowledge. The strength (or weakness) and source of knowledge must be included in the description under section K.

the estimated scenario or in the estimated probability and consequences needs to be recorded. For example, there may be sources of bias in the data, thus yielding uncertainty in estimated probability.

Outcome	**Consequence**	**Uncertainty**	**Knowledge**
A_1	C_1	Q_1	K_1
A_2	C_2	Q_1	K_2
...	...	...	...
A_N	C_N	Q_1	K_N

TABLE 1.3: Risk description table. It helps to organise all descriptors of risk into a table. Here the experiment or data are assumed to be available to obtain the list of all potential outcomes, the probability estimates, and the consequences estimates. The data may include knowledge elicited from a relevant domain expert, e.g., a clinician.

1.3.2 Know how to find an optimal decision

Imagine that someone asks us to recommend a decision. A decision needs to be made for a particular situation. A situation may involve risk. How shall we proceed in advising? If there is a risk involved, it means we need to recommend a decision under risk. Assume we already described the risk 1.3. Next, we need to ask the following questions:

1. What are each stakeholder's *preferences* about each of the consequences? This is about utilities. E.g. how much Janette likes the idea of having a dream job with a high salary? How much Janette is scared of being unemployed? Is the weight of being unemployed far more heavy than the joy of a dream job?

2. What *decisions* (actions) are available to take? E.g. is Janette able to find resources to pay for the degree that is relevant for that dream job, buy relevant clothes for the interview, or to get mentoring to get her prepared for the interview?

A decision analysis under (precise) risk considers a set of decisions (actions) in the face of risk while considering *utilities* of the consequences. This means the risk analyst needs to add a fifth element to the risk description: the utilities (see Eq. 1.2):

$$(A_i, C_i, Q_i, K_i, U_i) \text{ for } i = 1, 2, ..., N, \tag{1.2}$$

where

- A_i is the outcome (see above in section Description of risk)
- C_i is the consequence (see above in section Description of risk)
- Q_i is the uncertainty (see above in section Description of risk)
- K_i is the knowledge (see above in section Description of risk)

- U_i is the *utility*, preferences or loss function of the consequence. For example, Janette can be very unhappy if she does not get her dream job, but she can be five times more unhappy if she ends up unemployed. Sometimes, the decision maker's preferences are described formally by a *utility function*. A decision maker can be risk appetite or averse. The appetite or aversion to risk can be specified using a utility function or loss function. The evaluation of the utility takes into account the costs of actions and the costs/benefits of the consequences of the specific action, which may be monetary or of other forms. We will explain them fully in Chapter5 Decisions under risk and uncertainty.

Now it is important to say more about the knowledge (K_i):

1. We need data so we can estimate uncertainty Q_i (e.g. probability of a microwave exploding). Sometimes we cannot do experiments, and we cannot collect data. So we seek expert opinion. In either case, we need to write down how we made all the estimates of probability: Was it by using data? Do we trust the data? Was it from an expert? Do we trust the expert?

2. Typically, experiments are done to obtain information about each Q_i. Here we need to consider experimental design, sample size.

3. Sometimes, we do not need to experiment because there is an existing dataset that has been collected in the past. Sometimes the data have been collected over time repeatedly – every second, day, month, or year. Sometimes the data collection happened at a one-time point (e.g., Latin square design collected data). If data are collected at a one-time point (i.e., so-called cross-sectional data), then we use relevant statistical methods to estimate probabilities (e.g., General Linear Model and ANOVA analysis in Latin Squares Design).

4. There can be values of a continuous variable collected repeatedly over time, thus creating *Time Series*. For example daily new Covid-19 cases, hourly price of shares. Many decision problems come from events over time. We can use time series data to answer questions like: What is a stock price going to be in five days? Having the protection of the stock price, should I buy or sell the stock? What temperature will the nuclear reactor core be in 10 seconds? Having the prediction of the temperature, should the control rods be inserted now? We learn about time series in Chapter 3.

5. People collect timing of events datasets. E.g., the timing of hurricanes, the timing of switching to a new cereal brand. Then we can use such past data to ask the following questions: Is a hurricane likely to develop in the next 48 hours? Should an evacuation be ordered? This is tackled by methods called Markov Chains (see Chapter 4).

6. Data can be collected in a clinical trial. E.g., a trial was conducted to compare treatments A and B for high blood pressure, 100 patients took

medicine A and another 100 patients B took medicine B. Adverse events were collected as required by ethics. It was observed that 5 and 8 people had been vomiting after taking the medicine A and B, respectively. How would you use such data to estimate the probability of this adverse event?

7. The data may include knowledge elicited from a relevant domain expert, e.g., a clinician. Such process of obtaining the knowledge is called the *expert knowledge elicitation*.

There are two principled types of *decisions theories*:

1. *Non-statistical decision theory.* It does not use data, i.e. it does not run experiments or collect data. It is purely based on probability theory and utility theory.

2. *Statistical decision theory utilises data.* It is based on statistics, probability theory, and utility theory. It is based on knowing the random variables and their expectations. It is also based on statistical Inference (hypothesis testing and confidence intervals), and it can also use Bayesian statistics.

When making a decision, we as the decision maker may face the following:

- One-stage decision problem. E.g., we decide if we buy a share, and then we face the outcome, we decide to which school we put our child.

- Multi-stage decision problem. E.g. we decide if we pay for the advice of a financial advisor, and then after his advice, we decide if we buy the share or not, hence we are making our second decision. Here we can use Decision Trees (see Chapter 6).

1.3.3 Know how to communicate the risks and decisions

This is a crucial quality of a risk analyst: to communicate the risks and optimal decisions. Communication is not just a mere statement of risks and decisions, it is a dialogue. We devote Chapter 7 to this topic, where we look into examples.

1.4 A look ahead

Through this book, you will learn how to bring your *deciding under risk* and thinking to life. The structure of this exploration is outlined here. The chapters are divided into three broader units, each having a unique theme. However, there is a common thread throughout the book: building and analysing statistical models for the behaviour of some variable Y so that we can quantify the risk, and we can recommend the best actions under risk.

1.4.1 Unit I Risk foundations

Motivating question. Let us assume we have a stakeholder who is deciding if and how to invest her money into a stock to maximise her profit, Y. We know that risk and uncertainty are present. At our disposal, we have some simple data that we can use. How can we incorporate our risk and uncertainty thinking into a formal model of the variable Y? In other words, how can we estimate those probabilities p_i in Equation 1.1?

Throughout the examples, we showed that a risk analyst needs the following three kinds of skills: technical skills, skills in applying knowledge about human cognition, and communication skills. In the next, we discuss technical skills.

We need *probability*, which is a mathematical sub-discipline devoted to understanding random phenomena. It is one of the pillars of statistical (classical or Bayesian statistics) and non-statistical decision-making tools. We use probability throughout this book.

We need *statistics*, which is a discipline on extracting knowledge from data and from the provided information. There are three pillars: study design, understanding the properties of statistical methods, and data analysis (inference or prediction, i.e., classification, discrimination). There are two main streams of statistics: frequentist statistics and Bayesian statistics. And there are further related areas: data science Machine Learning, and artificial intelligence. Often there may be two or more statistical methods appearing suitable for the problem at hand; however, some will be useful and some incorrect to use. Hence we need to develop a *critical eye*: we keep learning the new statistical methods, keep up to date on what new methods are developed, what is the current understanding of the limitations of the classical methods we always check assumptions of the statistical method.

We need *computing skills*. We need computing skills for several reasons. Firstly, we need to process the data, so we use statistical packages like R, STATA, SPSS, Minitab, etc. Second, there is no statistical theory to help us to do estimation, we need to resort to *computer simulations*. Third, following a quantification of risk we would want to perform *sensitivity analyses*. The risk quantification evaluates the degree of knowledge or confidence in the calculated numerical risk results. The sensitivity analysis indicates what input changes are the analysis results most sensitive to. Monte Carlo computer simulation methods are generally used to perform sensitivity analysis.

1.4.2 Unit II Data driven risk analysis

Let us assume we have a stakeholder deciding if and how to invest her money into a stock to maximise her profit, Y. There is a risk involved and the stock market situation is simple. How to advise her?

If there are relevant data that we can use and if the scenario is simple, we

may be able to use some simple model from probability theory. For example, we may use a model where we assume independence of events and Normal distribution of the outcomes. We can then estimate the needed probabilities p_i for the risk description (Equation 1.1). We discuss the key probability concepts in Chapter 2.

1.4.3 Unit III Data and model driven risk analysis

Let us assume again that we have a stakeholder deciding if and how to invest her money into a stock to maximise her profit, Y. There is a risk involved and the stock market situation is complex with several factors to consider (such as technological changes in the relevant industry). How to advise her?

If there are relevant data that we can use and if the scenario is complex, we may be able to use some more complex model from statistical modelling. For example, we may use a Time Series model or Markov Chain. After fitting those models, we can then estimate the needed probabilities p_i for the risk description (Equation 1.1). We discuss the key probability concepts in Chapter 3 and 4.

This book is only focusing on Time Series and Markov Chains for three reasons. First, they are very commonly used methods for risk analysis. Second, we intended this book to be an intermediate-level book and we consider Time Series and Markov Chains suitable, while we assume no prior knowledge of them and we only show the intermediate-level modelling details. Third, it would not be possible to put all models in one book. So these two modelling approaches serve as examples of how to use statistical models for risk estimation, interpretation, and communication.

Machine learning and Artificial Intelligence methods can also be used to estimate the risks. In fact, many researchers call the Time Series and Markov Chains examples of machine learning methods. There is extensive research on the development and application of statistical and machine learning methods for risk estimation. In our chapters, we will point at several key examples.

1.4.4 Unit IV Decisions under risk

Let us assume again we have a stakeholder deciding if and how to invest her money into a stock to maximise her profit, Y. We know that risk and uncertainty are present. At our disposal, we have some data e_i that we can use. We know how to use the data to estimate the risk. We know how to recognise sources of uncertainty. Next, how should we advise the stakeholder? What action appears to be the best to take for the stakeholder? In other words, how can we recommend the best action a_i in Equation 1.2?

We first need to learn how much the stakeholders value certain outcomes, i.e. their utilities and we also need to learn how to find the best action if there is epistemic uncertainty (Chapter 5. We will also learn how to advise if the decision maker needs to do two or more decisions in sequence, in Chapter 6.

The skill of advising the best decision relates to fields of decision science, human cognition, psychology, philosophy, and economics.

1.4.5 Unit V Communication of risk

We need to be able to *communicate with stakeholders* to find out what the problem is and what they want. What properties a solution must satisfy? This is also essential so we to ask for the right data, right information. We need to be able to *communicate with domain experts*. This is again essential so we ask for the right data, so we get a good understanding of the problem they want to solve, and so we communicate our results usefully and correctly. We need to be able to *communicate with colleagues risk investigators*: mathematicians or data scientists, computer scientists. We need to be able to *communicate with stakeholders* the results of risk analysis and to navigate them so they can make a decision that aligns with their values and is the best informed too.

1.5 Summary

We learned in this chapter:

1. We learned that risk as a science discipline started evolving 40 years ago and is still developing. There are various definitions of risk; however, some became obsolete when new research evidence came to light. The risk research is multidisciplinary, from psychology (cognition, perception), and economics (consumer behaviour, utility, decision theory) to STEM disciplines (including the theory of probability and statistics). To stay up to date, it is important to follow risk guidelines of scientific societies such as *Society for Risk Analysis* (SRA), https://www.sra.org/.

2. We learned that risk and uncertainty are two different concepts. Risk is a situation that can lead to several outcomes; some of the outcomes have negative consequences. Uncertainty is a state of mind of the stakeholder (or of the risk analyst) when something is not known or certain. Probabilities can be unknown either totally, or they can be imprecisely quantified (such as "from 10 to 20%), or there can be uncertainty in the quality of the data from which we quantified the probabilities. Or outcomes and consequences may be unknown, either all of them or some of them (such as the Covid-19 consequences to the economy of countries were unknown in March 2020).

3. We learned the key components of risk description: outcome (A), consequences (C), quantification of uncertainties (Q), and knowledge strength (K).

4. We learned the main principles of risk management and that risk analysis is one of its components.

5. We learned that doing (or advising) on an optimal decision depends on the situation. There are two situations. The first situation is when the stakeholder is making a *decision under precise risk* which is also called as *decision under risk and aleatoric uncertainty* or *decision under risk*. The second situation is when the stakeholder is making a *decision under imprecise risk* which is also called *decision under risk and aleatoric as well as epistemic uncertainty* or *decision under uncertainty*. Later chapters of this book will show some strategies on how to advise a stakeholder in each of the two situations.

1.6 Further reading

The chapter was based on our research experience as well as the experience of others and on several books, which we list here:

1. To understand more about the risk topics we highly recommend the book *Risk Science* by Aven and Thekdi [9]. We used their notation and definitions of risk and uncertainty. They give more examples and further details on risk management and communication.

2. The book *Chance Rules: An Informal Guide to Probability, Risk, and Statistics*, by Brian Everitt [20] is a good reading material on introduction to risk.

3. The book *Probabilistic Risk Analysis. Foundations and Methods* by Tim Bedford and Roger Cooke [10] is an excellent advanced-level mathematical book. It was written from materials for a master-level module on mathematical foundations of risk. It has a good introduction to uncertainty, with a concise summary of relevant probability and statistics; and with chapters on system analysis, fault trees expert opinion, human reliability, project risk management, and uncertainty analysis.

4. The book by Carlton and Devore *Probability with Applications in Engineering, Science, and Technology* [13] is an intermediate-level textbook, highly suitable for undergraduate studies of applied mathematics.

5. The book *The Essentials of Risk Management* by Michel Crouhy, Dan Galai, and Robert Mark [17] is an intermediate-level book focused on risk in business, banks, finance, interest rates, credit risk, equity price risk, market risk, commodity price risk, foreign exchange risk, operational risk,

liquidity risk, corporate risk. It explains how to effectively implement an enterprise-wide risk management program, allocate capital.

6. The book *Risk Assessment and Decision Analysis with Bayesian Networks* by Fenton and Neil [22], is an intermediate to advanced level book on risk.

7. *The Book of Why* by Judea Pearl & Dana MacKenzie [46] is written for a general audience, and it is not on risk assessment and management, but it is a book on causality. They describe the history of famous paradoxes: how mathematicians and policymakers reacted. The given story of Simpson's paradox, among other paradoxes, and how this connects with artificial intelligence.

8. For discussion on the role of *artificial intelligence* in risk we recommend starting by reading the following research papers: Ale [8], Choi and Lambert [14], Guikema [27], [28], Nateghi and Aven [41], and Thekdi, Tatar, Santos and Chatterjee [58].

9. Uncertainty is an active area of research. More can be found in these resources: [9], [30], [61], [59], [43].

Another relevant term is *Risk Intelligence*. It is defined as "the organisational ability to think holistically about risk and uncertainty, speak a common risk language, and effectively use forward-looking risk concepts and tools in making better decisions, alleviating threats, capitalising on opportunities, and creating lasting value" (Columbia University professor Leo Tilman).

Additionally, we recommend the following journals on risk:

1. The journal that best covers the engineering areas of risk as discussed in this module is: *Risk Engineering and System Safety*.

2. Interdisciplinary journals covering methodology research in risk analysis and interdisciplinary applications: *Risk: Health, Safety, and Environment Risk Analysis*, *Risk, Decision, and Policy* and *The Journal of Risk Research*, *Journal of Approximate Reasoning*.

Part II

DATA DRIVEN RISK ANALYSIS

2

Probability

Gabriela Czanner

Silvester Czanner

CONTENTS

2.1 The axioms of probability 26
2.2 Assigning probabilities to events: three approaches 26
2.3 Basic rules of probability 29
2.3.1 Joint probability 29
2.3.2 Marginal probability 31
2.3.3 Conditional probability 31
2.3.4 The product rule for joint probabilities 33
2.3.5 Independence of events 33
2.3.6 The law of total probability 34
2.3.7 Bayes' theorem 36
2.4 Tips to think and act like a risk expert 38
2.4.1 Consider using natural frequencies in communication .. 38
2.4.2 Risks assessed via medical tests or medical AI 41
2.5 Summary 42
2.6 Further reading 42
2.7 `R` lab 43
2.8 Exercises 46

In this chapter, we introduce the basic principles of probability. We will use probability to tell how likely individual outcomes or scenarios are. We will discuss examples where probability quantifies the uncertainties Q in risk analysis, thus contributing to the big picture of risk description (Figure 1.3).

This chapter is the building block of *data-driven risk analysis*. By risk analysis, we mean the quantification of the uncertainty (e.g. how likely there will be a flood). By data-driven risk analysis we mean the analysis where we use data, which can be from an experiment (e.g., rolling dice) or from an expert

e.g., a broker revealing to us his belief on whether a particular investment will be successful.

Learning objectives

1. Learn about three definitions of probability. Learn how this leads to three ways we assign a probability to an event (outcome or scenario).
2. Learn about the fundamental rules of probability (including Bayes's rule).
3. Simulate random events in `R` and calculate probabilities of events in `R`.
4. Learn the names of probabilities used in medical tests and, more broadly, in Artificial Intelligence.

2.1 The axioms of probability

The set of all possible *outcomes* of an experiment is called the *sample space*. We will denote the outcomes by $O_1, O_2, \ldots$, and the sample space by S. Thus, in set-theory notation,

$$S = \{O_1, O_2, \ldots\}$$

All outcomes in S must be *mutually exclusive* and *exhaustive*. We are uncertain about which outcome will occur, but we know that one of them will occur. Sometimes we may be interested in the occurrence of a collection of outcomes, e.g. if an even number will roll on a die. Such occurrence is called an *event*.

Example. Roll of a die. One example of a sample space is $S = \{1, 2, 3, 4, 5, 6\}$ for one roll of a die. Such sample has six outcomes $O_1 = 1$, $O_2 = 2$, $O_3 = 3$, $O_4 = 4$, $O_5 = 5$ and $O_6 = 6$, and many events e.g. "A = an even number", or "B = a number less than 5". The outcomes $O_1 ... Q_6$ are also called simple events.

Next, we list the three axioms of probability. The axioms state that the probabilities must satisfy the following properties:

a) The probability of any event must be nonnegative, e.g., $P(O_i) \geq 0$ for each i.

b) The probability of the entire sample space must be 1, i.e., $P(S) = 1$.

c) For two disjoint events A and B, the probability of the union of A and B is equal to the sum of the probabilities of A and B, i.e., $P(A \cup B) = P(A) + P(B)$.

These axioms are important to ensure that mathematics is consistent with our everyday notions of probability (likelihood, chance).

2.2 Assigning probabilities to events: three approaches

We have a sample space $S = \{O_1, O_2, \ldots\}$ of all outcomes and we now want to assign the probabilities to them. There are three approaches to assigning probabilities to outcomes: the classical approach, the relative-frequency approach, and the subjective approach.

Classical approach. If an experiment has n simple outcomes, this method would assign a probability of 1/n to each outcome. In other words, each outcome is assumed to have an equal probability of occurrence. This method is also called the *axiomatic approach.*

Example. Roll of a die. Assume a die is rolled once. So the simple events are numbers from 1 to 6, and thus the sample space is $S = \{1, 2, \ldots, 6\}$. We can use the classical approach to assign probabilities to each possible outcome. According to such an approach, each simple event has a 1/6 chance of occurring. However, this is correct under the assumption that the die is perfect.

Example. Two rolls of a die. Next, we assume two rolls of a die. So the sample space of all simple events is $S = \{(1,1), (1,2), \ldots, (6,6)\}$. Assume that the die is perfect and assume that the two rolls are "independent." We can use the classical approach to assign the probabilities. Each simple event has a $1/6 \times 1/6 = 1/36$ chance of occurring.

Relative-frequency approach. In some situations, the probabilities are assigned on the basis of experimentation or historical data. Formally, let A be an event of interest, and assume that you have performed the same experiment n times so that n is the number of times A could have occurred. Further, let n_A be the number of times that A did occur. Now, consider the relative frequency n_A/n. Then, in this method, we "attempt" to define $P(A)$ as:

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

The above can only be seen as an attempt because it is not physically feasible to repeat an experiment an infinite number of times. Another issue with this definition is that if we repeat it two times, i.e. we run n experiments two times, we get two sets of n experiments which will typically result in two different ratios. However, we expect the discrepancy to converge to 0 for large n. Hence, for large n, the ratio $\frac{n_A}{n}$ may be taken as a reasonable approximation for the probability $P(A)$.

Example. Roll of a die. When we roll a die, the sample space is $S = \{1, 2, \ldots, 6\}$. We can use the relative frequency approach to assess the

probabilities. We can roll the given die 100 times (say) and suppose the number of times the outcome one is observed is 17. Thus, $A = 1$, $n_A = 17$, and $n = 100$. Therefore, we say that $P(A)$ is approximately equal to $17/100 = 0.17$.

Example. Measles weekly new cases. Liverpool health authority tracks the monthly new cases of Measles in the past 20 weeks. The resulting data is Table 2.1. We can use the relative-frequency approach to estimate

New cases	Number of weeks
0-1000	2
101-200	10
201-300	8

TABLE 2.1: Measles new cases.

probabilities of the event "a future week number of cases being 101-200". Since $10/20 = 0.5$, we have that there is a 50% chance that Liverpool will have 101-200 new cases of measles on any given future week.

Subjective approach. In the subjective approach, we define probability as the *degree of belief* that we hold in the occurrence of an event. Therefore, we use judgment as the basis for assigning probabilities.

Caution. Notice that the classical approach of assigning equal probabilities to simple events is, in fact, based on data as well as on judgment. This is because we may have to use our judgment to decide if the equal probabilities assumption is valid. What is somewhat different here is that the use of the subjective approach is usually used in experiments that cannot be repeated (e.g., see the Horse race example below).

Example. Horse race. Consider a horse race with six horses running. What is the probability for a particular horse to win? Is it reasonable to assume that the probability is 1/6? To answer the question, we note that we can not apply the relative-frequency approach. People regularly place bets on the outcomes of such "one-time" experiments based on their judgment as to how likely it is for a particular horse to win. Indeed, having different judgments is what makes betting possible.

Example: Stock price. What is the probability for a particular stock to go up tomorrow? Can we apply a relative-frequency approach to assess the probability? To answer these questions, we note again this stock "experiment" can not be repeated. We can not apply the relative-frequency approach. What we need is a sophisticated data-driven model - while incorporating expert judgment too. Sophisticated models, such as statistical or machine learning models, rely on past data and do forecasts for the future. Expert judgments

are needed too: we can ask for expert opinion about probability and compare it against the data-driven approach probability, and we can ask for expert opinion about the probability when we feel uncertainty, e.g. ask an expert if the unmeasured factors that drove the data are going to change. Note that the financial crisis in 2007 was not predicted by any sophisticated model. Hence, the combination of the two approaches is needed, as blindly following data only is dangerous, and blindly following ill-founded judgments is often also dangerous.

When we present the assigned (estimated) probabilities to stakeholders, we have a responsibility to *interpret the probabilities* in an understandable way. This is a matter of clear communication of probabilities. The way we interpret the probability depends on how we assign (estimate) its value. Here are some interpretations:

1. In the Two rolls of a die example, the interpretation is: "In many experiments where we roll a die twice, we observe 1 and 1 $1/36 \times 100\%$ times".
2. In the Measles example, we interpret the probability as: "50% of future weeks will have 101-200 new cases".
3. In the Horse race example, we wanted the probability of the particular horse winning. Let us assume we asked Anna to say what she believes the probability is. Anna can be an expert, or non-expert or a half-expert. Let us assume that Anna feels that there is a 90% chance of the particular horse winning. Then such probability interpretation is: "Anna believes that there is a 90% chance that the horse wins, or that the odds are 9:1 of the horse winning".

2.3 Basic rules of probability

In what follows we present the basic rules of probability.

2.3.1 Joint probability

So far, we have been dealing with the probability of single events (e.g. of getting a number 1 in one roll of a die). What if we want the probability that involves multiple events? In other words, what if we are interested to know the probability of two events happening at the same time? For example, what if we want to know the probability of Pedro going to the cinema and of raining outside? Such probability is called a joint probability since it expresses the likelihood of two events happening jointly. We will explain the joint probabil-

ity via an example.

Example. Boxes and jewels. A girl in India called Swati is choosing the next jewel for her choodamani. She has two boxes full of jewels: an oval box and a square-shaped box. Each box contains some red rubies and some white pearls - in Figure 2.1. We want to know the probability of Swati picking a particular box and a particular jewel. Let us start with a notation:

- Let B be the event of picking a box. The values of B are o for oval and s for square shape box.
- Let J be the event of picking a jewel. The values of J are r for red ruby and w for white pearl.

We are told that Swati picks a box proportionally to the number of jewels in the box. Then she picks a jewel from the chosen box. What is the probability of choosing the oval box and choosing a red ruby? To answer the question, it helps to organise all outcomes from Figure 2.1 above into a Table 2.2. In how many ways can we get an oval box and a red ruby? The answer is 2.

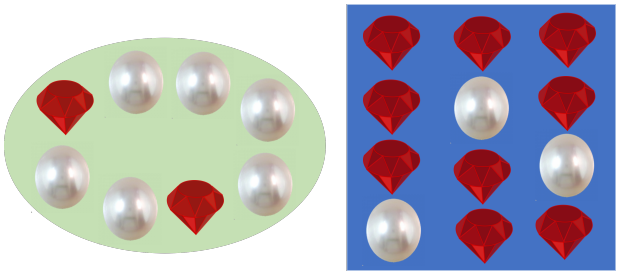


FIGURE 2.1: Example Boxes and jewels. Oval and square boxes with white pearls and red rubies.

		Jewel		
		red ruby (r)	white pearl (w)	Total
Box	oval (o)	2	6	8
	square (s)	9	3	12
	Total	11	9	20

TABLE 2.2: Table of frequencies in Boxes and jewels example.

We can now answer questions about joint probabilities:

- $P(B = o \cap J = r) = \frac{2}{20} = 0.1$ We use $\cap$ here to indicate joint appearance of two events. In other places, also in books, we use "," or "AND", thus $P(B = o, J = r) = 0.1$, or $P(B = o \, AND \, J = r) = 0.1$

- $P(B = o \cap J = w) = \frac{6}{20} = 0.3$
- $P(B = s \cap J = r) = \frac{9}{20} = 0.45$
- $P(B = s \cap J = w) = \frac{3}{20} = 0.15$

So, for example, the probability of choosing the oval box and ruby is 0.1. This is a joint probability, as it says how likely the two events will happen jointly, even though Swati picks the box first and then she picks the jewel, hence sequentially.

2.3.2 Marginal probability

In the Example of Boxes and Jewels, we can ask about the probability of choosing a particular box. Or we can ask about the probability of choosing a particular jewel. All probabilities are:

- $P(B = o) = 8/20 = 0.4 = P(B = o \cap J = r) + P(B = o \cap J = w)$
- $P(B = s) = 12/20 = 0.6 = P(B = s \cap J = r) + P(B = s \cap J = w)$
- $P(J = r) = 11/20 = 0.55 = P(B = o \cap J = r) + P(B = s \cap J = w)$
- $P(J = w) = 9/20 = 0.45 = P(B = o \cap J = w) + P(B = s \cap J = w)$

So, e.g. the probability of choosing the oval box is 0.4. This is called the marginal probability because it concerns just one event, the box, i.e. it concerns the oval box "margin" of the Table 2.2.

2.3.3 Conditional probability

We are going to expand our probability notation a bit more. We want to be able to specify the probability of an event, given that another event has occurred. In other words, we want to be able to specify the probability of an event, under the condition that another event has occurred. For this, we use conditional probability.

Definition of the conditional probability. The probability of event A given that B has occurred will be denoted as $P(A \mid B)$ and defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{2.1}$$

The conditional probability is answering the following question: under the condition that event B occurs (did occur, or will occur), what is the probability that event also A occurs (occurred, or will occur)? If I reveal to you that event B occurred, what do you think is the probability of event A occurring too? The answer is $P(A \mid B)$. If I do not reveal to you if event B occurred, you can still think in hypothetical terms: if hypothetically B occurred, then what is the probability of event A occurring too? The answer is $P(A \mid B)$.

Example. Boxes and jewels. (continues). Consider the following scenario: Swati randomly chooses a box and a jewel. We are curious. We ask Swati what she chose. Let us assume two situations:

1) Firstly, Swati does not want to reveal to us what she chose. She does not want to reveal the box shape. She does not want to reveal the jewel type, either. What is the probability that she chose red ruby? In other words, what is our belief about choosing a red ruby by Swati?

2) Then after some persuasion, Swati tells us that she chose the oval box, but she still does not want to reveal what jewel she chose. Now we have some information: we know she chose the oval box. What is the probability of choosing a red ruby if we know she chose the oval box? In other words, what is the probability of choosing a red ruby, among all jewels that that come from the oval box?

Solution. First situation is concerting the probability of choosing a red ruby and it is not concerning the shape of the box. So we need to find the marginal probability $P(J = r)$. We know $P(J = r) = \frac{11}{20} = 0.55$ (as we calculated before).

The second situation is different. We are told that the oval box was chosen i.e. we are provided with a new piece of information that "Swati chose the oval box". Knowing that the oval box was open, what is our belief that the chosen jewel is a red ruby? So we now realise that what we are asked to calculate is the conditional probability: $P(J = r \mid B = o)$. Using the Table 2.2 we get the probability is

$$P(J = r \mid B = o) = \frac{2}{8} = 0.25$$

Alternatively, we can use the definition of the conditional probability

$$P(J = r \mid B = o) = \frac{P(J = r \cap B = o)}{P(B = o)} = \frac{0.1}{\frac{8}{20}} = \frac{2}{8} = 0.25$$

Note that the 0.55 probability is often interpreted as a prior probability. It is our initial belief of choosing a red ruby before we knew what box was chosen. Then the 0.25 probability is our posterior probability. It is our updated belief of a red ruby is being chosen. Notice that once we learned that the chosen box is oval the probability decreased to 0.25, in this example.

Example. Boxes and jewels. (continues). What is the probability of choosing the oval box given that we know that a red ruby was chosen?

Solution: We realise that what we are asked is to calculate another conditional probability:

$$P(B = o \mid J = r)$$

Using Table 2.2 we see that the probability is

$$P(B = o \mid J = r) = \frac{2}{11} = 0.1818$$

An alternative calculation is to via using the definition of the conditional probability

$$P(B = o \mid J = r) = \frac{P(B = o \cap J = r)}{P(J = r)} = \frac{0.1}{\frac{11}{20}} = \frac{2}{11} = 0.1818$$

2.3.4 The product rule for joint probabilities

We already saw how to calculate the joint probabilities from a table (e.g. Table 2.2). If we are not provided such a table, we can use the product rule (or multiplication rule) for probabilities. The following is always true for any events A and B:

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A) \tag{2.2}$$

Example. Boxes and jewels. (continues) What is the probability that Swati randomly selects a red ruby and the oval box?

Solution. We are told that the probability of choosing the oval box is $\frac{8}{20}$, and we are told that the probability of choosing a red ruby from the oval box is $\frac{2}{8}$.

We realise we are asked to find the joint probability of two events occurring

$$P(J = r, B = o)$$

which can also be written as

$$P(J = r \cap B = o)$$

We can use the product rule for the joint probabilities:

$$P(J = r \cap B = o) = P(J = r \mid B = o)P(B = o) = \frac{8}{20}\frac{2}{8} = 0.1$$

2.3.5 Independence of events

Intuitively, when two events are independent of each other, they do not affect each other. Formally, two events A and B are independent if and only if they satisfy the condition

$$P(A \mid B) = P(A), \tag{2.3}$$

which is equivalent to satisfying the condition

$$P(B \mid A) = P(B), \tag{2.4}$$

and that is equivalent to satisfying the condition

$$P(A \cap B) = P(B)P(A). \tag{2.5}$$

So to check the independence, we need to check if one of the above conditions is true. The equivalence of the three conditions can be proved by using the product rule of probabilities.

Example. Boxes and jewels. (continues) Are choosing a red ruby and the oval box independent events, in our example?

Solution. There are three ways to solve this.

Solution 1. What is the probability of choosing a red ruby if we know that the oval box was chosen? It is the conditional probability $P(J = r \mid B = o) = 0.25$. What is the probability of choosing a red ruby? It is the marginal probability, $P(J = r) = 0.55$. So we get

$$0.25 = P(J = r \mid B = o) \neq P(J = r) = 0.55$$

These two probabilities are not the same, hence events "choosing a red ruby" and "choosing the oval box" are not independent.

Solution 2. What is the probability of choosing the oval box if we know that a red ruby was chosen? It is the conditional probability $P(B = o \mid J = r) = 0.1818$. What is the probability of choosing an oval box? It is the marginal probability, $P(B = o) = 0.4$. So we get

$$P(B = o \mid J = r) = 0.1818 \neq 0.4 = P(B = o)$$

These two probabilities are not the same, hence events are not independent.

Solution 3: The probability of choosing a red ruby is the marginal probability $P(J = r) = 0.55$. The probability of choosing the oval box is the marginal probability $P(B = o) = 0.4$. The probability of choosing them both is the joint probability $P(B = o, J = r) = 0.1$. Hence we get

$$0.1 = P(B = s \cap J = r) \neq P(J = r)P(B = o) = 0.55 \times 0.4$$

so the two events "choosing a ruby" and "choosing the oval box" are not independent.

2.3.6 The law of total probability

We start with a definition of *events, mutually exclusive* and *events, exhaustive*. Events $A_1, \ldots, A_k$ are mutually exclusive if no two have any common outcomes. Hence, no two events can occur together. The events $A_1, \ldots, A_k$ are exhaustive if $A_1 \cup A_2 \cup \ldots \cup A_k = S$. Hence, all events exhaust the whole space S, and this means that at least one of events A_i occurs. We are interested in events $A_1, \ldots, A_k$ that have both properties: they are *mutually exclusive*

and exhaustive events, as illustrated in Figure 2.2. Hence, exactly one of the events $A_1, \ldots, A_k$ occurs. We will need probabilities of such events in later sections when we will be calculating conditional probabilities.

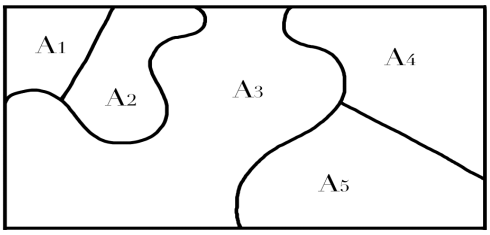


FIGURE 2.2: Partition of space S into mutually exclusive and exhaustive events A_1,...,A_5.

The law of total probability. Let $A_1, \ldots, A_k$ be mutually exclusive and exhaustive events. Then for any other event B, the following holds

$$P(B) = P(B \mid A_1)P(A_1) + \ldots + P(B \mid A_k)P(A_k) = \sum_{i=1}^{k} P(B \mid A_i)P(A_i)$$

This can be illustrated in the following Figure 2.3, where the events $B, A_1, \ldots, A_k$ are represented via sets B.

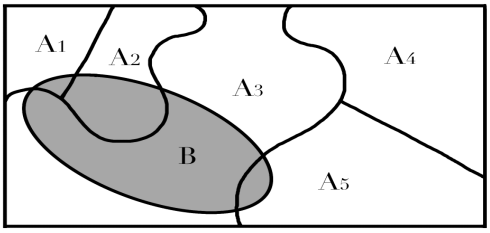


FIGURE 2.3: Partition of set B by mutually exclusive and exhaustive events A_1, ..., A_5

Example. University students. At a certain university, 4% of men are over 6 feet tall and 1% of women are over 6 feet tall. The total student population is divided in the ratio of 3:2 in favour of women. If a student is selected at random from all students, what is the probability that the student is over 6 feet?
Solution. We start by setting a notation:

- M = "Student is Male"

- F = "Student is Female"
- T = "Student is over 6 feet tall"

We recognise that M and F partition the space of students. This is important for the total law of probability. We need to calculate $P(T)$.

We know the following probabilities

- $P(F) = 3/5 = 0.6$
- $P(M) = 1 - 0.6 = 0.4$
- $P(T \mid M) = 0.04$
- $P(T \mid F) = 0.01$

Finally, we use the law of total probability

$$P(T) = P(T \mid F)P(F) + P(T \mid M)P(M) = 0.01 \times 0.6 + 0.04 \times 0.4 = 0.022$$

Answer: The probability that a randomly selected person is over 6 feet tall is 0.022.

2.3.7 Bayes' theorem

In probability theory and statistics, (alternatively or) describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if the probability that someone has cancer is related to their age, using Bayes' theorem the age can be used to more accurately assess the probability of cancer than can be done without knowledge of the age. Using Bayes theorem we find $P(A_j|B)$ if we know $P(B|A_j)$.

The central idea of Bayes' theorem—that we start with a prior belief about the probability of an unknown hypothesis and revise our belief about it once we see evidence—is also a central concept of the law.

One of the many applications of Bayes' theorem is Bayesian inference, a particular approach to statistical inference. With Bayesian probability interpretation, the theorem expresses how a degree of belief, expressed as a probability, should rationally change to account for the availability of related evidence. Bayesian inference is fundamental to Bayesian statistics.

The Bayes' rule is defined and interpreted as an update rule that changes a prior probability into a posterior probability by incorporating a data-based likelihood. It is calculated as follows:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{2.6}$$

where

- $P(A \mid B)$ is the *posterior probability*. This is the updated probability that we aim to calculate.

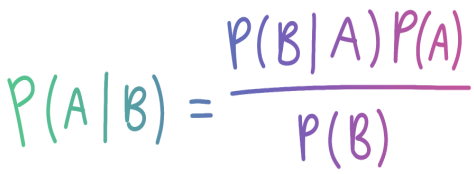


FIGURE 2.4: Bayes' theorem. This is also known as Bayes's rule or Bayes's law.

- $P(A)$ is the *prior probability*. This is the prior belief we have about the event A before we know if B occurred or not.
- $P(B)$ is a marginal probability of event B occurring. Sometimes, this can be tedious to calculate. We calculate it as

$$P(B) = \sum_{all\ values\ of\ a} P(B \mid A = a)P(A = a)$$

  by the law of total probability, where a are all possible outcomes of A e.g. all types of jewels.
- $P(B \mid A)$ is called the *likelihood*, also called *data-based likelihood*.

Bayes' theorem is named after Reverend Thomas Bayes (read as /beiz/; lived in 1701?–1761). In short, the theorem is visualised in Figure 2.4.

Proof of Bayes's rule. By definition, we have

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

hence

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

By the same argument, due to symmetry, we have

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

So, we have

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

hence

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

Example. University students. (continues.) If a student is selected at random from among all those over six feet tall, what is the probability that the student is a woman?

Solution. We start by setting a notation:

- M = "Student is Male",
- F = "Student is Female". (Note that M and F partition the space of students.)
- T = "Student is over 6 feet tall"

We recognise that what we need to calculate is a conditional probability:

$$P(F \mid T)$$

We know this:

- $P(F) = \frac{3}{5} = 0.6$
- $P(M) = 1 - 0.4 = 0.6$
- $P(T \mid M) = 0.04$
- $P(T \mid F) = 0.01$
- $P(T) = 0.022$ This last probability is what we calculated earlier using the law of total probability.

The fact that we need to calculate the quantity $P(F \mid T)$ while we know $P(T \mid F)$ should ring a bell: "The Bayes theorem may work here!" By using Bayes's theorem we get

$$P(F \mid T) = \frac{P(T \mid F)P(F)}{P(T)} = \frac{0.01 \times 0.6}{0.022} = 0.272727$$

Answer: If a student is selected at random among all those over six feet tall, the probability that the student is a woman is 27.27%. In other words: among the students that are over 6 feet tall, there are 27.27% of women.

2.4 Tips to think and act like a risk expert

Next, we look again into the problem of calculating probabilities. We will illustrate the ideas in other real world examples (i.e. use cases).

2.4.1 Consider using natural frequencies in communication

Here we will calculate again the conditional probabilities in one real-world example. However, we will consider two situations: first, we will be presented with all data via probabilities (as we did in the previous section), and then we will be presented with all information via natural frequencies. In both situations, we will use Bayes's rule to find the required conditional probability.

Example. Colorectal cancer. First, we introduce a problem assigned via probabilities. To diagnose colorectal cancer, the hemoccult test – among others – is conducted to detect occult blood in the stool. This test is used from a particular age on, but also in routine screening for early detection of colorectal cancer. Imagine you conduct a screening using the hemoccult test in Pennsylvania. For symptom-free people over 50 years old who participate in screening using the hemoccult test, the following information is available for this region:

- The probability that a person has colorectal cancer is 0.3%.
- If a person has colorectal cancer, the probability is 50% that he will have a positive hemoccult test.
- If a person does not have colorectal cancer, the probability is 3% that he will still have a positive hemoccult test.

Imagine a (randomly chosen) person with age over 50, no symptoms from Pennsylvania who has a positive hemoccult test. What is the probability that this person actually has colorectal cancer? Can you answer this without using a pen and paper, just by doing the calculations in your head? What is your best guess? It turns out that the majority of people (almost everyone) cannot do this mentally.

Solution. Next, we use pen and paper to do the calculations. We start with notation and with what we know:

- Let A = Person has cancer (where by a person we mean a human being who is symptom-free, over 50 years old and resides in Pennsylvania)
- Let B = Person has positive test

Then we list the provided information

- P(A)=0.3% = 0.003
- $P(B \mid A) = 50\% = 0.5$
- $P(B \mid notA) = 3\% = 0.03$

What we are asked to calculate is $P(A \mid B)$ while we are provided with $P(B \mid A)$, so we recognise we have to use Bayes's rule:

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

We get

$$P(A \mid B) = \frac{0.5 \times 0.003}{0.5 \times 0.003 + 0.03 \times (1 - 0.003)} = \frac{0.0015}{0.03141} = 0.048$$

where in the denominator we used the Rule of total probability.

Answer: So, among all people who are symptom-free, over 50 years old, residing in Pennsylvania, and with a positive test, only 4.8% actually have cancer.

Next, we discuss the same problem, but now assigned via **natural frequencies.**

Example. Colorectal cancer. (continues) To diagnose colorectal cancer, the hemoccult test – among others – is conducted to detect occult blood in the stool. This test is used from a particular age on, but also in routine screening for early detection of colorectal cancer. Imagine you conduct a screening using the hemoccult test in Pennsylvania. For symptom-free people over 50 years old who participate in screening using the hemoccult test, the following information is available for this region:

- Thirty out of every 10,000 people have colorectal cancer.
- Of these 30 people with colorectal cancer, 15 will have a positive hemoccult test.
- Of the remaining 9,970 people without colorectal cancer, 300 will still have a positive hemoccult test.

Imagine a (randomly chosen) person with age over 50, with no symptoms, from Pennsylvania who has a positive hemoccult test. What is the probability that this person actually has colorectal cancer?

Can you answer this without using a pen, i.e. mentally? It turns out that much more people are able to correctly calculate this mentally. Next, we use pen and paper and do the calculations.

Solution. Among the 10,000 people there are 15 and 300 who have positive tests, where those 15 do have cancer. So the probability of having cancer, if the test is positive, is, thanks to Bayes's rule:

$$15/(15 + 300) = 0.048$$

Answer: So, among all people who are symptom-free, over 50 years old, residing in Pennsylvania, and with a positive test, only 4.8% actually have cancer. This answer is identical to the one we got when using the conditional probabilities.

Caution! In the next, we reflect on the calculations above. What did just happen? We tried to solve the same problem, but each time the information was provided in different ways: first time via conditional probabilities and then via natural frequencies.

a) Which information did you find easier to use to find the answers?

b) Which information would you find easier when explaining to a non-mathematician?

c) How is typically information provided in newspapers? Conditional probabilities or natural frequencies?

2.4.2 Risks assessed via medical tests or medical AI

The probabilities in Section 2.4.1 are important for daily life. We hear them when we are in a doctor's office when the doctor tells us how accurate a blood test is to detect a vitamin C deficiency, we hear them in the news when a public health professional tells us how accurate the home test kit for Covid-19, or we hear them in the news when a journalist tells us the accuracy of a newly developed AI algorithm to detect glaucoma.

Such probabilities are key to understanding the risks and making informed decisions. The risks estimated via medical tests or from medical AI are expressed via probabilities. Such probabilities have special names. In the Colorectal cancer example, the risk expert should communicate the following way:

1. "The probability that one of the people (asymptomatic, over 50, from Pennsylvania) has colorectal cancer is 0.3%." This is the so-called *prevalence of disease*.

2. "Among those with colorectal cancer, the chance that the test is positive is 50%." This is called the *sensitivity of the test*.

3. "Among those without colorectal cancer, the chance that the test is negative is 100-3=97%". This is called the *specificity of the test*.

4. "Among those whose test is positive, the chance that the person has cancer is 4.8%". This is called *positive predictive value of the test*. This is the value that a patient wants to know.

5. "Among those whose test is negative, the chance that the person does not

have cancer is $\frac{9670}{15+9670} = 0.998 = 99.8\%$". This is called *negative predictive value of the test*. This is the value that a patient wants to know.

Caution! A good risk expert (and journalist or health professional, too!) never communicates sensitivity and positive predictive values only. All four numbers must be communicated. Why? Because most tests are based on using a threshold (such as a threshold on the probability of glaucoma using an AI algorithm); where by decreasing the threshold, we improve sensitivity while decreasing the specificity. In other words, a given test's sensitivity and specificity are negatively associated. Then the selection of a positive test threshold involves an inherent balancing act (see, e.g., [51]). Additionally, we list further important elements of risk communication in our later Chapter 7.

2.5 Summary

We have learned in this chapter:

1. We discussed some simple examples of quantifying uncertainty using probability. We learned Bayes' theorem, how to recognise when to use it, and how to use it.

2. We learned how to simulate random events from a prescribed distribution and find required probabilities using `R`.

3. We learned that using data in natural frequencies provides the easiest way to calculate and communicate risks.

4. We learned terminology used in medical tests: sensitivity, specificity, negative predictive value, and positive predictive value. Such terminology, however, goes beyond medical tests, and it is more and more used in Artificial Intelligence, where the goal is to automatically detect or predict a presence or absence of an event (such as a threat to the digital system or fraud at band account, or disease of a plant, presence of a default of a construction etc).

2.6 Further reading

Our chapter was based on our research and the research of others, as well as on our comprehension of several monographs. The main sources of our inspiration were:

1. For an intermediate-level reading on probability, we recommend the book *Probability and its applications* by Carlton and Devore [13].

2. For those starting with `R` or who like to refresh their practice, we recommend the online book by Avril Coghlan [15]. Alternatively, for a gentle introduction to `R`, we recommend the online book *Introduction to Econometrics with R* by Christoph Hanck, Martin Arnold, Alexander Gerber and Martin Schmelzer M. [29]. For a comprehensive reference on R, we recommend *The R Book* by Michael J. Crowley [16].

3. For an in-depth understanding of the communication of risks, we recommend the work of psychologist Gerd Gigerenzer, especially his 2002 book named *Reckoning with risk. Learning to live with uncertainty* [25] and his 2014 book named *Risk savvy. How to make good decisions* [26]. Several sections and examples in our book were inspired by his book.

4. For a further understanding of Bayesian statistics, we recommend the book *Causal Inference: What If* by Miguel Hernán and James Robins [31].

5. Regarding conditional probabilities, one of the most famous examples *Monty Hall Problem*. As it turned this problem is quite intriguing, and it confused several famous statisticians and journalists. A fascinating discussion about this problem and its history is in *The Book of Why. The new science of cause and effect* by Judea Pearl and Dana Mackenzie [46].

2.7 `R` lab

Here, we give several questions to solve in `R`, and we provide solutions with detailed explanations. Then we give further questions without solutions.

1. **[Purpose: to practice the simulation of random events and finding the probabilities, in `R`].** Here we will use computer simulations to find the probability of an event. We will use software `R` (and we will use `RStudio` interface). As probability models in engineering and science have grown in complexity, many problems have arisen that are too difficult to solve analytically, i.e., using mathematical tools like Bayes' theorem. Instead, computer simulations provide us with an effective way to estimate probabilities of very complicated events and of random phenomena. Here, in this tutorial, you will be introduced to the principles of probability simulation, and demonstrate a few examples in R.

   We begin with an example in which we know the exact probability solution analytically so that we will be able to compare it with the solution from the simulation. Suppose we have two independent devices, which function with probabilities 0.6 and 0.7, respectively.

a) What is the probability that both devices function? Do simulations in R to find this probability.

b) What is the probability that at least one device functions? Do simulations in R to find this probability.

Solution. Let D1 and D2 denote the events that the first and second device functions, respectively. We know

$$P(D1) = 0.6, P(D2) = 0.7$$

We recognise that in (a) we are asked to find the joint probability $P(D1 \cap D2)$ and we are asked to do this in R via computer simulation. If we were not asked to do simulations, we would simply use the knowledge of independence of $D1$ and $D2$, and this would lead us to

$$P(D1 \cap D2) = P(D1)P(D2) = 0.42.$$

We also recognise that in (b) we are asked to find $P(D1 \cup D2) = 1 - P(notD1 \cup D2)$ where we use independence again, so this is equal to

$$1 - P(notD1)P(notD2) = 1 - 0.4 \times 0.3 = 1 - 0.12 = 0.88$$

Before we write an R code to solve (a) and (b), it helps to first write a pseudo-code where we sketch the basic ideas and flows in the R code. Here is a pseudo-code for our example in several steps:

1) We define a counter A. We set it to 0. Counter A will store number of times (out of 10000) where D1=1 and D2=1, i.e. the number of times when both devices worked.

2) We decide that $D1$ is a variable that has a value of 1 if the first device works, and zero otherwise. We need to decide how we do such a simulation. We can check if R is having a function to simulate a dichotomous or binary variable (both D1 and D2 are dichotomous or binary discrete variables). If R does not have a simulation of dichotomous or binary variables then we use this trick: to simulate a binary variable u1 with P(D1=1)=0.6 we simulate a random number from a uniform distribution on [0,1] interval, let call this simulated number c. Then if c¡=0.6, we set D1=1 (i.e. we say the first device functions), and if c¿0.6 then we set u1=0 (i.e. we say that the first device does not work). Can you see why this trick work?

3) We decide that D2 is a variable that has a value of 1 if the second device works, and zero otherwise. To simulate values of D2 we do the same trick as above, but with 0.7 (instead of 0.6). Can you see that this simulates D1 and D2 independently?

4) We decide to simulate D1 and D2 10,000 times. Obviously, the more simulations the higher precision of our estimate will be, and the longer we need to wait for the computer to do all the simulations.

5) In a loop, we simulate the two events D1 and D2 independently 10000 times (conveniently, we will use a built-in function `R` to do the loop in `R`).

6) For each of 10,000 simulations calculate this: if D1=1 and D2=1, then update the counter i.e. increase the value of A by 1; otherwise do not update the counter A.

7) Finally, after all, 10,000 simulations, calculate the estimated probability as A/10,000.

8) Since we use computer simulations, our answer in the previous bullet point may not be exactly equal to 0.42, but we should be reasonably close if we do enough simulations. How do we know that we have enough simulations i.e. that 10,000 is enough here?

From the pseudo code above we can write the following `R`-code:

```
# ------------------------------------------------------------
# a)  What is the probability that both devices function?
# Do simulations in R to find this probability.
# ------------------------------------------------------------
  # We will simulate two devices D1 and D2.
  # We will count how many times they both work.
  # Then we will store the count in a structure called A.
  A<-0 #Initialisee the counter at number zero (i.e. zero
    events so far)
  # Next, do 10,000 simulations of an event
  for (i in 1:10000){
    # Next, generate 1 random number from uniform distri-
    # bution on intervals 0 and 1.
    D1=runif(1)
    # Next, generate another random number from unif
    distribution on (0,1)
    # This way D1 and D2 are simulated independently of each
    other
    D2<-runif(1)
    # Next, simulates the joint event when both devices work.
    This is
    # accomplished by the following if condition.
    if(D1<0.6 && D2<0.7) {
      # If this condition above is true, then we increase the
    counter
      # i.e. this increases the count if both devices work
    jointly.
      A<-A+1
    }
  }
  #Nextt, give the estimated probability, for (A)
[1] 0.425
# ------------------------------------------------------------
# b) Simulation to find the probability of at least one device
    functioning.
# ------------------------------------------------------------
  B<-0
  for (i in 1:10000){
```

```
32      D1=runif(1) # generate a random number from the uniform
        distribution
33      D2<-runif(1)
34      if(D1<0.6 || D2<0.7) {
35        B<-B+1
36      }
37    }
38    B/10000 #This gives an estimate for theh probability that at
        least onedevice functions.
39  [1] 0.8835
40  # NEXT - Do try to change the seed into a different number and
        see how the estimates change!
```

Answer: Our estimate of the probability of both devices to function is 0.425. Our estimate of the probability of at least one device functioning is 0.8835.

2. [**Purpose: to practice finding the probabilities using computer simulations, in R.**] Consider the following game. You will flip a coin 25 times, winning £1 each time it lands heads (H) and losing £1 each time it lands tails (T). Unfortunately for you, the coin is biased in such a way that P(H)=0.4 and P(T)=0.6. What is the probability you come out ahead, i.e., you have more money at the end of the game than you had at the beginning? You will use simulation to find out this probability.

   Here is a hint that should give you a start: Since winning by flipping a coin 25 times is a random event, you need to simulate it in many runs. Hence your simulation will consist of many runs. In each run, you will simulate the 25 flips of the coin. In each run, you will need to keep track of much money you have won or lost at the end of the 25 tosses. Define an R structure A as the following: A = "We come out ahead".

   a) First, write your pseudo code.

   b) Using your pseudo code, write the R code that will estimate the probability.

2.8 Exercises

Solve the following exercises by using a pen, paper, and calculator.

1. [**Purpose: to practice marginal, conditional and joint probabilities.**] The population of a particular country consists of three ethnic groups. Each individual belongs to one of the four major blood groups. The accompanying joint probability table gives the proportions of individuals in the various ethnic group-blood combinations. Suppose that an individual is randomly selected from the population.

	Blood type			
Ethnic group	**O**	**A**	**B**	**AB**
1	0.082	0.106	0.008	0.004
2	0.135	0.141	0.018	0.006
3	0.215	0.200	0.065	0.020

a) Calculate $P(Blood = A)$, $P(Ethnic = 3)$ and $P(Blood = A \cap Ethnic = 3)$.

b) Calculate both $P(Blood = A \mid Ethnic = 3)$ and $P(Ethnic = 3 \mid Blood = A)$ and explain in context what each of these probabilities represents.

c) If the selected individual does not have type B blood, what is the probability that he or she is from ethnic group 1?

2. **[Purpose: to discuss conditional probabilities.]** Suppose an individual is randomly selected from the population of all adult males living in the USA. Let A be the event that the selected individual is over 6ft in height, and let B be the event that the selected individual is a professional basketball player. Which do you think is larger $P(A \mid B)$ or $P(B \mid A)$? Why?

3. **[Purpose: to practice marginal, conditional and joint probabilities, and independence of events.]** The accompanying table gives information on the type of coffee selected by someone purchasing a single cup at a particular airport kiosk. Consider purchasers arriving at the kiosk randomly i.e. independently of each other. Consider the next customer (hence, we are considering a randomly selected customer).

	Cup size purchased		
Coffee type	**Small**	**Medium**	**Large**
Regular	14%	20%	26%
Decaf	20%	10%	10%

a) What is the probability that the individual purchased a small cup? A cup of decaf coffee?

b) If we learn that the next customer purchased a small cup, what is now the probability that he/she chose decaf coffee? And how do you interpret this probability?

c) If we learn that the next customer purchased decaf, what is now the probability that a small cup was selected, and how does it compare to the corresponding unconditional probability from a)?

d) If we learn that the next customer purchased decaf, what is now the probability that a small cup was selected, and how does it compare to the corresponding unconditional probability from a)?

e) Is choosing cup size and coffee type independent? Explain.

4. [**Purpose: to practice the law of total probability and Bayes's theorem.**] Suppose that we have three coloured boxes (red), (blue), and (green). Box r contains 3 apples, 4 oranges, and 3 limes, box b contains 1 apple, 1 orange, and 0 limes, and box g contains 3 apples, 3 oranges, and 4 limes. A box is chosen at random with probabilities P(r)=0.2, P(b)=0.2, P(g)=0.6, and then a piece of fruit is selected from the box (with equal probability of selecting any of the items in the box).

   a) What is the probability of selecting an apple?

   b) If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

5. [**Purpose: to practice the law of total probability and Bayes's theorem.**] A factory production line is manufacturing bolts using three machines, A, B, and C. Of the total output, machine A is responsible for 25%, machine B for 35%, and machine C for the rest. It is known from previous experience with the machines that 5% of the output from machine A is defective, 4% from machine B, and 2% from machine C. A bolt is chosen at random from the production line and found to be defective.

   a) What is the probability that it came from machine A?

   b) What is the probability that it came from machine B?

   c) What is the probability that it came from machine C?

6. [**Purpose: to practice the law of total probability and Bayes's theorem.**] An engineering company advertises a job in three newspapers, A, B and C. It is known that these papers attract undergraduate engineering readerships in the proportions 2:3:1. The probabilities that an engineering undergraduate sees and replies to the job advertisement in these papers are 0.002, 0.001, and 0.005 respectively. Assume that the undergraduate reads only one newspaper and then either decides whether to apply for the job, hence there is either zero or one job application from each undergraduate.

   a) If the engineering company receives only one reply to its advertisements, calculate the probability that the applicant has seen the job advertised in place A.

   b) If the company receives two replies, what is the probability that both applicants saw the job advertised in paper A? Assume that the two readers are independent (e.g. do not live in the same household).

7. [**Purpose: to practice the probabilities, law of total probabilities and Bayes's theorem.**] 1% of people have a certain genetic defect. 90% of the tests for the gene detects the defect (true positives). 9.6% of the tests are false positives. If a person gets a positive test result, what are the odds they actually have the genetic defect?

8. [**Purpose: to practice Bayes's rule, as well as the communication of the results.**] This example was created about Lateral flow (LDF) test for Covid-19. In 2021, Dylan Mistry and his collaborators compared several LFDs (published in BMC Infectious Diseases, 2021, 21:828). One of the manufacturers of LFD is Bioeasy. Mistry investigated how accurate is the LDF Bioeasy test. For that purpose, they randomly selected n=856 people from London, with no or mild symptoms. For these people who participate in the study, the following information was available: The probability that a (randomly chosen) person has Covid-19 is 5.22% (i.e., 5.22% is the estimated prevalence for London at that time). If a person has Covid-19, the probability is 82.05% that he/she will have a positive LDF Bioeasy test. If a person does not have Covid-19, the probability is 8.73% that he will still have a positive LDF Bioeasy test. Answer the following questions:

   a) What was the prevalence of Covid-19 in 2021, among Londoners with no or mild symptoms?

   b) Imagine a randomly chosen person (from London, with no or mild Covid-19 symptoms) from London who has a positive LFD Bioeasy test. What is the probability that this person actually has Covid-19? Answer this question by using probabilities.

   c) Next answer the question b) by using natural frequencies. Hint: you do not need to use n=856, you can use any n you wish, just do not use small values.

   d) Imagine a random sample of 500 people from London with no or mild Covid-19 symptoms who have positive LDF Bioeasy. How many of these people actually have Covid-19?

Part III

DATA AND MODEL DRIVEN RISK ANALYSIS

3

Time series for risk quantification

Silvester Czanner

Gabriela Czanner

CONTENTS

3.1 Motivation 54
3.1.1 Time series can be used for forecasting of the future ... 55
3.1.2 The forecasting steps 58
3.1.3 What do stakeholders want? 60
3.2 Introduction to statistical theory of time series 65
3.2.1 Notation 66
3.2.2 Time series as a realisation of random process 67
3.2.3 Mean, variance, autocovariance and autocorrelation 67
3.2.4 Stationarity 68
3.3 Exploratory data analysis for time series 69
3.3.1 Goal of exploratory data analysis 69
3.3.2 Trend, seasonal and cyclic components of time series ... 70
3.3.3 Estimating the autocorrelation 75
3.4 Time series modelling and forecasting 82
3.4.1 Time series regression models 82
3.4.2 Exponential smoothing models of time series 85
3.4.3 Choosing the best-fitting model 95
3.4.4 Think ZINC! 99
3.4.5 Prediction intervals 103
3.5 Tips to think and act like a risk expert 108
3.5.1 Remember there is no such a thing as a free lunch! 109
3.5.2 Be a pro at visualising the risk and uncertainty 111
3.5.3 When relying on a model alone is a wrong idea 112
3.6 Summary 113
3.7 Further reading 113
3.8 R Lab 114
3.9 Exercises 138

Time series is a set of data points collected over time, typically at regular intervals. In a time series, the order of the data points is important, as they represent a sequence of observations recorded over time. The time series data are typically measured on a continuous scale (such as the weight of a person) or as counts (such as the number of visitors to a museum). Time series data is often used in various fields, including finance, economics, engineering, and natural sciences, to analyse and forecast trends and patterns. Examples of time series data include stock prices, temperature measurements, and monthly sales figures. In financial risk analysis, time series models can be used to analyse historical data on stock prices, interest rates, or other financial indicators to predict future market trends and fluctuations.

Time series data can be analysed using various statistical techniques, including trend analysis, seasonal analysis, and forecasting. The goal of time series analysis is to identify patterns, trends, and other relationships in the data that can be used to make predictions about future values.

Time series data values are almost always correlated with each other, e.g. if my weight is above average today, it was likely above the average two days ago. Such correlation is the main reason why time series data need their own types of data analytic methods, so the correlation is suitably accounted for. This gave rise to the development of a field called Time Series Analysis.

Learning objectives

- Learn what time series data are, learn several time series modelling methods, learn how to find a suitable model for forecasting, learn how to fit the models and use the models to forecast the future. Learn how we express the uncertainty of such a forecast.
- Practice modelling and forecasting in R.
- Discuss how to communicate the calculated uncertainties to stakeholders.

3.1 Motivation

First, we look at the motivation of how time series help to quantify risk. Time series can be used for three types of goals:

- **Forecasting.** We can estimate what may happen in the future.
- **Control.** We can estimate if anything is going wrong. For example, if someones eye is unhealthy (now, not in the future), if a production process is out of control (such as producing defects).

- **Understanding the features of the data.** This includes seasonality, cycle, trend, and any change points. The degree of seasonality in agricultural prices may indicate the degree of development. This also includes understanding what causes our time series data (such as the effect of government policy on the number of visitors to the UK).

3.1.1 Time series can be used for forecasting of the future

Here we illustrate what we mean by forecasting. We start by giving several examples, we define terminology, and we show what types of forecasts we want to accomplish. Then we discuss the process of building such forecasts in later Sections.

Example. Overseas visits. We are going to use data on the number of overseas visitors to the UK from December 1981 till December 1981 (Table 3.1) obtained from Office of National Statistics [42]. The data are visualised on a time series plot, Figure 3.1. What patterns can be seen in the figure?

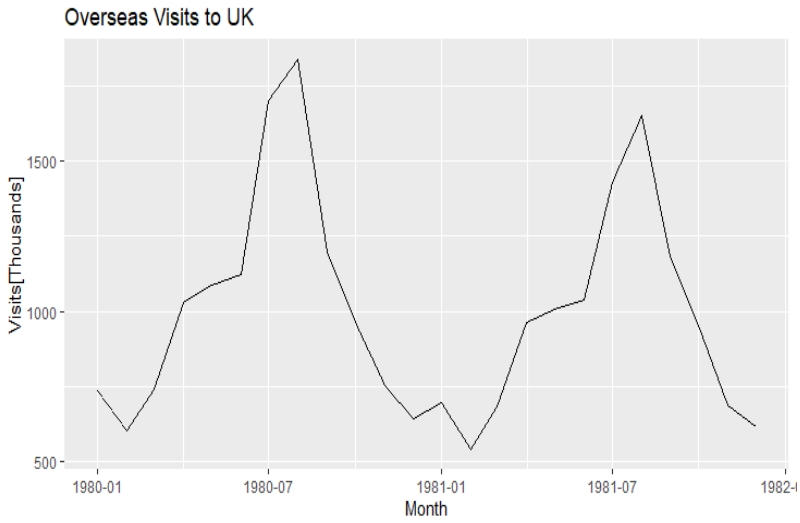


FIGURE 3.1: Overseas monthly visits to the UK January 1980 to December 1981. The visits are measured in thousands of visitors. The values are in Table 3.1.

Solution. The time series plot (Figure 3.1) immediately reveals some interesting features: During December and January months, the visits are the lowest; during the summer months, the visits are the highest; and the year 1980 seemed to have slightly higher visits than the year 1981.

t	Month	Year	Visits
1	JAN	1980	739
2	FEB	1980	602
3	MAR	1980	740
4	APR	1980	1028
5	MAY	1980	1088
6	JUN	1980	1124
7	JUL	1980	1699
8	AUG	1980	1839
9	SEP	1980	1200
10	OCT	1980	963
11	NOV	1980	755
12	DEC	1980	642
13	JAN	1981	695
14	FEB	1981	540
15	MAR	1981	685
16	APR	1981	962
17	MAY	1981	1007
18	JUN	1981	1039
19	JUL	1981	1430
20	AUG	1981	1650
21	SEP	1981	1181
22	OCT	1981	954
23	NOV	1981	689
24	DEC	1981	619

TABLE 3.1: Monthly visitors time series for January 1980 - December 1981.

Example. Overseas visits. (continues) In the next, we will consider the same monthly time series, but now from January 1980 till December 2020, hence 41 years. The Figure 3.2 reveals some interesting features:

- During December and January months, the visits are the lowest; during August, the visits are the highest; this is called a seasonal pattern.
- The extent of these seasonal differences (August visits minus January visits) is roughly the same, except it seems that these differences are somewhat larger in the last three years.
- There is a general increasing trend except for the decreasing trend from January 2000 to December 2001. Why could this be?
- There is a drop in visits from January 2008 to January 2009; this coincides with the financial worldwide crisis. There is a constant trend from January 2009 to December 2012, possibly still due financial crisis. Then there is a constant trend from January 2016 to December 2019. This is then followed

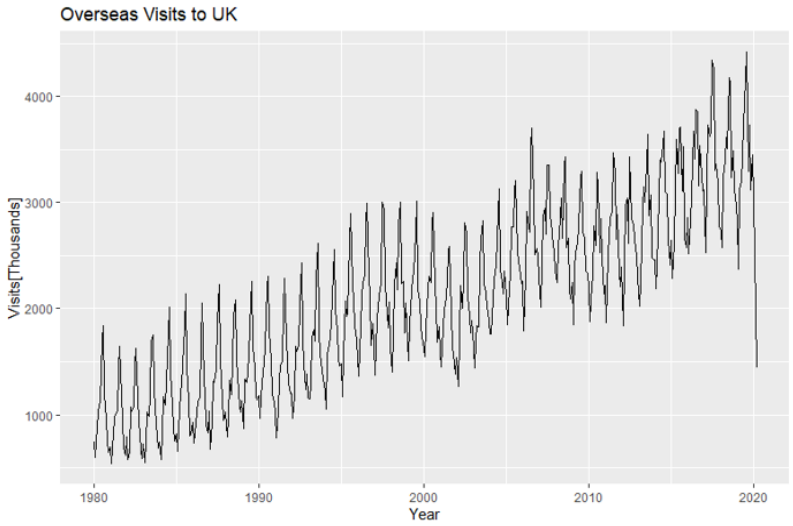


FIGURE 3.2: Overseas monthly visits to UK Jan 1980 to Dec 2020. The visits are in the thousands.

by a drop from December 2019 to March 2020, which coincides with the Covid-19 pandemic and the closing of flights.

- We also see some random fluctuations.
- Hence, when preparing the forecasts from this series, we would need to take into account the seasonal pattern, the trend, as well as the random fluctuations.

What would the tourism institutions like to know from these visits' time series? They may be interested in having answers to the following questions:

1. What is the expected number of visits each month from April-December 2020?
2. What are the chances that from April to December 2020, the total number of visits will be between 2500 and 3500 thousand?
3. What are the chances that in the next three months, there will be at least 3000 visits per month?

Next, assume it is December 1981, and we have the visit data from January 1980 - December 1981. In which of the two following scenarios should we feel more confident to do a forecast?

- Scenario 1: We are asked to use the data to predict the number of visits for January 1982.

- Scenario 2: We are asked to use the data to predict the number of visits for April 1982.

Answer: We should feel more confident to predict January 1982 because this is one month ahead prediction. We are less confident to predict April 1982 as it is four months ahead of the prediction. This means that we will need to have a way to quantify such confidence numerically. We will do it in future sections.

Next, we need to introduce several definitions. *Smoothing* time series is a technique where we are trying to remove 'noise' from data. We can do this by building a statistical model of the data, such as a linear model with additive noise. The smoothed data would then be on the line, and we will also call them the fitted or predicted values. For smoothing, we don't need an explicit model. One example of that will be a so-called *moving average* operation, which also smooths data and does not assume an explicit model. Other examples of smoothing operations are Kernel smoothing, smoothing splines, linear regression, loess regression, and exponential smoothing (see Sections 3.4.2). *Forecasting* is a process of making predictions into the future beyond the time scale of available data. We want to be able to predict something that has not happened yet. In essence, we want to see into the future. Forecasting is part of a bigger scheme of statistical inference. The idea is to use available data to infer about the future. We can do this by first estimating the mean or trend of the data, then we see what properties such estimates have, and we use them to predict the future.

3.1.2 The forecasting steps

When someone gives us **time series dataset**, how do we start creating the forecast? We start by clarifying what the ultimate goal of the stakeholder or decision-maker is. This is the most challenging and the most important part. If we do not understand the goal, we create a useless forecast. Our role as a mathematician / statistician / data analyst is not just to listen to the decision-maker and make notes. Our role should not be passive. It should be active. We should ask clarifying questions about the goal. Then we should make suggestions: what is possible or impossible to do with data and forecasting. E.g. "It is now the year 2023, and I will help you to make predictions for the year 2030, but be aware that those predictions may not be trusted as they are too far from the data we have now. I will help you to make predictions for the whole year of 2024, but if you are planning to roll out a new immigration policy in January 2024, my predictions will not be valid anymore. So if you want me to predict what happens after the rollout of the immigration plan, you need to give me past time series data of Monthly Visits that covers a period when a similar rollout happened."

What is a time series dataset? It is a set of data obtained on the same variable repeatedly over time, while the time intervals are equidistant. In time

series, the unit of analysis is a day (or a month, in Figure 3.1), depending on how regularly you collect the data). A time series dataset is always organised into a table; usually, the variables are in columns (in Table 3.1, the columns are Time and Visits), and the units are in rows.

What makes the time series data special? Time series data is different from data collected at one-time point. Data collected at one-time point are called cross-sectional studies, such as randomised block design, Latin squares design or cohort studies or surveys. The time series data are collected over time and hence are usually correlated; hence they are not independent. So, we see that the most basic assumption of the classical regression model is violated in time series data: in regression models, we assume that data are uncorrelated. This calls for more flexible statistical methods.

What does it mean that the data in the time series are usually correlated? It means that e.g. if on January 2021, the number of new COVID-19 active cases is higher than the overall trend, then in February 2021, the number of new cases will more likely be also above the overall trend. That is an example of a positive correlation.

In what follows is a set of five steps that we should follow when forecasting from time series data:

- **Step 1. Figuring out what the problem is.** Finding out what the goal of the decision maker is and reframing it in statistical language. As Hyndman and Athanasopoulos [34] point out, this is the most difficult part of forecasting. Defining the problem carefully requires an understanding of the way the forecasts will be used, who requires the forecasts, and how the forecasting function fits within the organisation requiring the forecasts. A forecaster needs to spend time talking to everyone who will be involved in collecting data, maintaining databases, and using the forecasts for future planning. To be successful in this first step, it helps to ask the following questions to those who will use this forecast: "What is your goal? What do you want to achieve? E.g. do you want to know how much ice cream will be sold by each vendor? Or do you want to know how much ice cream all your vendors will sell on average? Do you want to forecast if the patient will be in remission a year from now? Or do you want to know the proportion of patients a year from now?"

- **Step 2. Gathering information.** There are always at least two kinds of information required: (a) statistical data and (b) the accumulated expertise of the people who collect the data and use the forecasts. Often, it won't be easy to obtain enough historical data to be able to fit a good statistical model. In that case, the judgmental forecasting methods can be used (see Chapter 4 of Hyndman and Athanasopoulos book [34]). Occasionally, old data will be less useful due to structural changes in the system being forecast; then, we may choose to use only the most recent data. However,

remember that good statistical models will handle evolutionary changes in the system; don't throw away good data unnecessarily.

- **Step 3. Exploratory time series analysis (EDA).** This is a preliminary stage of analysis of data. We start by graphing the data. Then we look at the data, and we ask ourselves the following questions: "Are there consistent patterns? Is there a significant trend? Is seasonality important? Is there evidence of the presence of business cycles? Are there any outliers in the data that need to be explained by those with expert knowledge? How strong are the relationships among the variables available for analysis?" Various tools have been developed to help with this analysis. These are Graphical Tools for Time Series and a method called Time Series Decomposition. During EDA analysis, we graph the time series and we speak to the application domain experts (e.g. experts on the stock exchange if we want to forecast prices of shares, and pandemic experts if we want to predict Covid-19 new cases). The result of such a discussion is a list of potential candidate models. We learn the basics of EDA in Section 3.3 after we first introduce notation and basic definitions in Section 3.2.

- **Step 4. Fitting several candidate models and choosing the best-fitting and well-fitting model.** We fit all candidate models and compare them. We aim to choose the best-fitting models (one or more equally good models). The choice of the best-fitting model depends on the availability of historical data, the strength of relationships between the forecast variable and any explanatory variables, and how the forecasts are to be used. Each model is itself an artificial construct that is based on a set of assumptions (explicit and implicit) and usually involves one or more parameters which must be estimated using the known historical data. Some of the commonly used model types are Time Series Regression models (see Section 3.4.1), exponential smoothing models (see Section 3.4.2), Box-Jenkins ARIMA models, dynamic regression models, hierarchical forecasting, neural networks and vector autoregression. To choose the best fitting model, we will use several suitable criteria in Section 3.4.3. Once we find the **best fitting model**, we will do goodness-of-fit checks of such a model to see if it is also a **well-fitting model**(Section 3.4.3).

- **Step 5. Evaluating the forecasting model and making recommendations to a decision maker.** Once a model has been selected as appropriate for forecasting, next its parameters are estimated, and the model is used to make forecasts. The model's performance can only be properly evaluated after the data for the forecast period have become available. Several methods have been developed to help in assessing the accuracy of forecasts. When using a forecasting model in practice, numerous practical issues arise, such as how to handle missing values and outliers or how to deal with short time series. When communicating the forecast to a decision maker, there are several points we will discuss in Section 3.5.

3.1.3 What do stakeholders want?

The reason why we study time series data is that they help us to forecast the future world, i.e. the future outcomes and their probabilities. Knowing future outcomes (scenarios) and quantifying uncertainties (via, e.g. probabilities) is the key to assessing the risks and making decisions as we discussed in Section 1.3.1: see A for outcomes and Q for quantified uncertainties, in Figure 1.3.

From the view of statistics, the number we are trying to forecast is unknown. What we are trying to do is to estimate the unknown quantity as best as we can. For example, the total visits for next month could take a range of possible values, and until we add up the actual visits at the end of the month, we don't know what the value will be. So until we know the visits for next month, it is an unknown quantity.

We should be more certain in our forecast of the time series value for the next month than for 13 months ahead. As one month it is relatively close, we usually have a good idea of what the likely visit values could be. On the other hand, if we are forecasting the visits for the same month next year, the range of possible values can be much wider. In most forecasting situations, the range of possible forecasted values will be wider as we forecast into a more distant future. In other words, the further ahead we forecast, the more uncertain we are, and hence we will need a method to quantify such uncertainty too.

How do time series help to estimate future scenarios? We will learn one typical pipeline: we fit a suitable time series model, and then we do extrapolation into the future (see section 3.1.2).

Example: Overseas visits. (continues) Imagine it is December 1981, and we have the overseas visits data for January 1980 - December 1981. We want to forecast the overseas visits for the next 10 years. How do we present such forecasts to the stakeholders so that it is simple to communicate?

Solution. Here, we discuss the way to present a forecast to stakeholders. The mechanics of calculating the forecast are explained in future sections.

The first possibility to present a forecast to stakeholders is by showing several possible future scenarios that are likely to happen. Plotted in black in Figure 3.3 is the total overseas visitors to the UK from Jan 1980 to Dec 1981. Also shown are four individual possible futures into January 1982 – December 1991.

In Figure 3.3, as we try to predict a more distant future (e.g. January 1990), the variability of the four individual futures for January 1990 is about three times larger than the variability of the four individual futures for January 1983, i.e. we are less certain in simulating possibly futures for January 1983 then for January 1990.

The second possibility to present a forecast to stakeholders is by showing the centre of the range of possible values the future can take as well as by

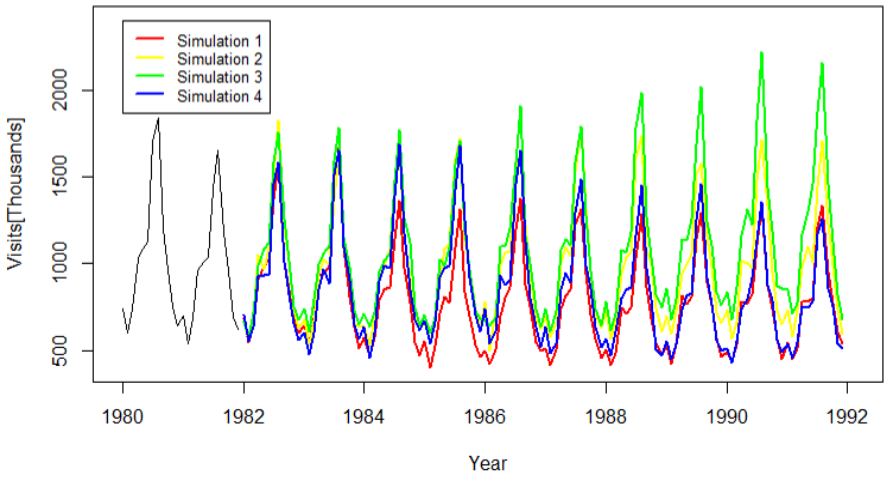


FIGURE 3.3: Total monthly overseas visitors to the UK (January 1980 - December 1981, see the black curve) along four individual possible futures for January 1982 – December 1991.

showing the prediction intervals as in Figure 3.4. The Figure shows 80% and 95% prediction intervals for future UK overseas visitors.

When we are creating a forecast (Section 3.4), we usually start by calculating the centre of the range of possible values the future can take. Since, for each time, the middle is just one number (i.e. one point on the plot), we call it a *point forecast*. When we connect all point forecasts, we get the whole path which is also called the *central path of projection* [47]. Then we make sure that our point forecast (central path) is accompanied by a *prediction interval* giving a range of values the random variable could take with relatively high probability. For example, a 95% prediction interval contains a range of values which should include the actual future value with probability 95%. The choice of the percentage depends on stakeholder needs. For example, in high-stake situations such as predicting the number of new Covid-19 cases, very high coverage of prediction intervals is needed, such as 99%.

The width of the prediction intervals (Figure 3.4) indicate the amount of our *uncertainty*: the wider the prediction intervals are the more uncertain we are. As we forecast a more distant future (e.g. January 1990) the variability of the forecast for January 1990 (as expressed by the width of the prediction intervals) is about three times larger than the variability of January 1982; i.e. there is more uncertainty in predicting the visits for January 1990 than for January 1982. We will learn how to calculate such intervals in Section 3.4.5.

Two goals of the stakeholders when they want to predict the fu-

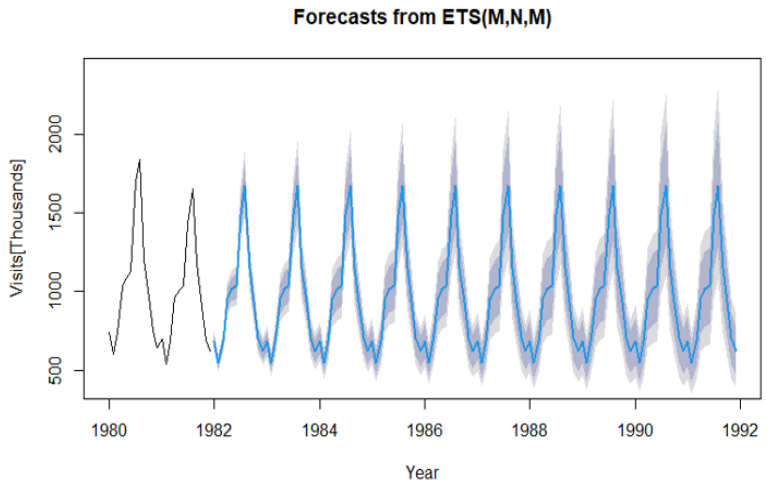


FIGURE 3.4: Total overseas visits to the UK (January 1980 - December 1981, see the black curve) along four individual possible futures for January 1982 – December 1991. The blue line is the average of the possible future values, which we call the point forecasts. The dark grey region is the 80% prediction interval, the light grey region is the 95% prediction interval.

ture. When we use time series to forecast the future and communicate the risks to a stakeholder, we need to understand the stakeholder's goals. There are two types of goals (see also Step 1 in Section 3.1.2). As data analysts, we should be able to figure out which of the two goals we are facing (or both). In the next, we discuss each of the two goals via illustrative examples:

When the goal is to forecast the expected (i.e. mean) outcome. We discuss this via two examples:

1. You are a data analyst and are approached by an ice cream maker. She produces ice cream and sells it via 30 vendors, and she wants to know what sale she can expect on 30 July from her 30 vendors. She has data from 2 vendors, the other 18 vendors are new. We can use the data to forecast the expected sales: i.e. of the mean sale. Then we just multiply by 30 to get the sales estimate for 30 vendors.

2. You are a data analyst and approached by an investor. She wants to invest now in a set (i.e. portfolio) of 100 shares, and she wants to see what her profit (or loss) will be in a month from now. She has time series data on one of these shares. You recognise that here it is useful to try to estimate the expected (hence mean, average) profit a month from now.

The two examples above have one thing in common: they require estimating the expected value of mean future sales/profits. We can use time series data to estimate the expected future scenario (i.e. the mean future scenario). Note, that in statistics, the term the expected value is the same as the term the mean value. So, for example, we can use data from January 1980 - December 1992 to estimate the mean number of visits in January 2006. In other words, we can estimate the expected number of visits in the month of January 2006. This forecasted mean number of visits (i.e. forecasted expected number of visits) for the month of January 2006 is just a single number, representing one point on the line of real numbers, and hence it is called the *point estimate* of the mean (see the blue line in Figure 3.4). For each month, we can also calculate the confidence interval for the mean, which we can also call the *interval estimate of the mean*. Note that the intervals in Figure 3.4 are NOT the confidence intervals for the mean; they are prediction intervals for individual outcomes - which we discuss in the next.

When the goal is to forecast the individual outcome/future. Again, we discuss two examples:

1. What if we only have one ice cream vendor that sells our ice cream and we want to know how many ingredients to put into the truck in 7 days from now on 30th July? What if we have 30 vendors that sell our ice cream, and we want to predict how much each individual vendor needs to stock up on the ice cream in 7 days from now on 30th July? This means we need to forecast the ice cream demand of individual vendors on 30th July.

2. What if we want to invest and we only have money to buy one asset? Then there is not useful to only predict the average return. It is more useful to forecast the return by calculating the future average return as well as the prediction interval. This will give us an idea of what all possible returns are and their probabilities, this will help us in deciding about one single asset investment. We recognise we are in a riskier situation here as we cannot diversify the risks by buying several assets. We put all our money into one asset.

In the two examples above, we recognise that it is needed to estimate all the future scenarios (outcomes, sales, profits and losses) and their likelihoods. (Note, once we know all future scenarios, we can use them to estimate the mean.) What we do in these situations is that we use time series to estimate all future scenarios that are likely to happen i.e. that are supported by the past time series data. We then estimate the probability (i.e. likelihood) for each of these scenarios. Some scenarios will be more likely to happen, some less. For example, we will forecast all future numbers of visits that can possibly occur, as well as their probabilities. Some forecasted numbers will be more supported by the data, and this will be expressed by having high probabilities. Some forecasted numbers will be less supported by the data, and this

will be expressed by having low probabilities. We will express these ranges by constructing a prediction interval. Such intervals were showed in as grey intervals in Figure 3.5. A further motivating example follows in the next paragraph.

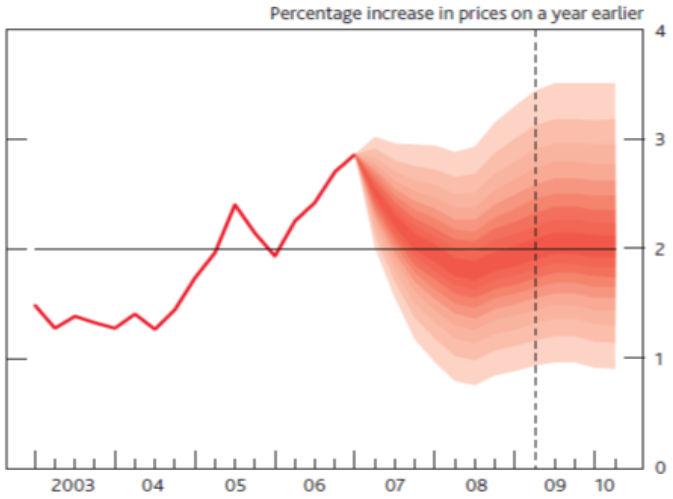


FIGURE 3.5: Inflation forecast visualised via a fan chart thus showing the uncertainty of the future outcomes. This forecast is based on market interest rate expectations and based on the collective judgement of the monetary policy committee. From Bank of England Inflation Report, May 2007.

Example. Inflation. A so-called *fan chart* can be used to convey uncertainty about future outcomes. Below is the fan chart depicting the forecast of inflation. It shows the probability of various outcomes for inflation in the future. The interpretation is as follows: If economic circumstances identical to today's were to prevail on 100 occasions, the monetary policy committee's best collective judgement is that inflation over the subsequent three years would lie within the darkest central band on only 10 of those occasions. The fan chart is constructed so that outturns of inflation are also expected to lie within each pair of the lighter red areas on 10 occasions. Consequently, inflation is expected to lie somewhere within the entire fan chart on 90 out of 100 occasions. The bands widen as the time horizon is extended, indicating the increasing uncertainty about outcomes. The dashed line is drawn at the two-year point. In Section 3.4.5 we will learn how to create such fan plots. Here, we discussed why they are needed.

3.2 Introduction to statistical theory of time series

Next, we give theoretical background for time series. This will be then used in later sections when we bring forecasting models and tools.

3.2.1 Notation

Next, we introduce the notation for time series. We will use the capital letter Y to denote the variable measured repeatedly over time (e.g., the monthly number of visitors) and use the small letter y to denote the actual value (i.e., the value that was observed and recorded). We use subscript t for time, so y_t denotes the observation at time t. We use the capital letter T to denote the number of time points for which we have data; hence it is the last time point. Hence, we denote all our observed time series data as

$$y_1, \ldots, y_T$$

We want to predict the future, i.e. the values of time series beyond the time T. We will call such predictions the *forecasts*. Each forecast will be based on some information. A useful information is the actual collected data: $y_1, \ldots, y_T$. Then we can calculate the forecast at time $T+h$ given the data $y_1, \ldots, y_T$. Such forecast will be called a *h-step forecast* taking into account all observations up to time T, and we denote it as

$$\hat{y}_{T+h|T}$$

where $h \geq 1$ is the "horizon" in the future which we like to forecast. The "hat" indicates that the quantity is an estimate. The subscript $T + h|T$ indicates that we are estimating the time series value at time $T + h$ while conditioning our forecast on the data $y_1, ..., y_T$.

In principle, when we do forecasting, we can use any relevant information available to us. For example, if we want to predict the number of visitors to the UK in the year 2024, we can use the past number of visitors time series until and including 2023, but we can also use other information such as changes in travel policies, prices of hotels in the UK until and including 2023. In this chapter, we only show how to use the past time series of the number of visitors to do the forecast.

In principle, we can also think about predicting values $y_{t+1|t}$, where $t < T$, which is predicting the value y_{t+1} where we take into account all previous observations $y_1, y_2, \ldots, y_t$. We call such predicted values as *fitted values*

$$\hat{y}_{t+1|t}$$

And we reserve the word forecast for a scenario where we predict into the future, i.e. beyond time T.

It is important to note that the value $\hat{y}_{T+h|T}$ is random. In other words, there is a random chance involved in it due to random variations in $y_1, ..., y_T$, hence $\hat{y}_{T+h|T}$ is a realisation of the random variable $\hat{Y}_{T+h|T}$. The set of values that this random variable could take, along with their relative probabilities, is known as the probability distribution of $\hat{Y}_{T+h|T}$. In forecasting, we call this the *forecast distribution*, and we use it to construct the interval forecast. We use $\hat{y}_{T+h|T}$ as the point forecast. If the forecast distribution is symmetric, then the forecast is the mean of the forecast distribution.

3.2.2 Time series as a realisation of random process

Time series is a time-ordered collection of observations. From a probability point of view, time series is a realisation of a collection of T random variables, $\{Y_1, Y_2, \ldots, Y_T.\}$ Time series may be needed to answer questions such as:

- What is the predicted number of visits in January 2004?
- What is the expected number of visits in February 2004 given that in January 2004 it was 4,100?
- Are the values of past time series independent of each other? Do the observations carry information about the next observations?

The first two questions are about the moment (i.e. the expected value) of the time series at some time point, and the last question is about the correlation of the individual components of the time series at various time points.

We use small letters to denote the realisations of $Y_1, Y_2, \ldots, Y_T$, i.e. we write $y_1, y_2, \ldots, y_T$. They are also called observations, measured values or recorded values.

3.2.3 Mean, variance, autocovariance and autocorrelation

For time series $\{Y_1, Y_2, ...\}$ the *mean function* is defined by

$$\mu_t = E(Y_t)\, for\, t = 1, 2, ...$$

That is, μ_t is just the expected value of the process at time t. In general, μ_t can be different at each time t.

The *autocovariance function*, $\gamma_{t,s}$, is defined as

$$\gamma_{t,s} = Cov(Y_t, Y_s)\, for\, t, s = 1, 2, ...$$

where $Cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E(Y_t Y_s) - \mu_t \mu_s$

Since the time series $\{Y_1, Y_2, \ldots, Y_T\}$ has T components, we can have T variances (i.e. variance of each Y_t), and we can have $T(T-1)/2$ covariances

(i.e. covariance of each component with each other component).

The *autocorrelation function*, $\rho_{t,s}$, is given by

$$\rho_{t,s} = Corr(Y_t, Y_s) = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}} = \frac{Cov(Y_t, Y_s)}{\sqrt{Var(Y_t)Var(Y_s)}} \tag{3.1}$$

which is the correlation between the t-th and s-th time of Y. In the mumerator, we can normalise so that the autocorrelation is in the range $[0, 1]$. Recall that both covariance and correlation are measures of the linear dependence between random variables. The following important properties follow from our definitions:

$$\gamma_{t,t} = Var(Y_t)$$
$$\gamma_{s,t} = \gamma_{t,s}$$
$$\mid \gamma_{t,s} \mid \leq \sqrt{\gamma_{t,t}\gamma_{s,s}}$$
$$\rho_{t,t} = 1$$
$$\rho_{t,s} = \rho_{s,t}$$
$$\mid \rho_{t,s} \mid \leq 1$$

The correlation $\rho_{t,s}$ is unitless and somewhat easier to interpret: it gives the strength of the dependence between Y_t and Y_s. Values of $\rho_{t,s}$ close to ± 1 indicate strong linear dependence, whereas values near zero indicate weak linear dependence. If $\rho_{t,s} = 0$, we say that Y_t and Y_s are *uncorrelated*.

3.2.4 Stationarity

To make statistical inferences about the time series on the basis of observed data, we must usually make some simplifying while reasonable assumptions about the data-generating mechanism. The most important such assumption is that of *stationarity*. The basic idea of stationarity is that the probability laws that govern the behaviour of the time series do not change over time.

A time series $\{Y_1, Y_2, ...\}$ is said to be *strictly stationary* if the joint distribution of $\{Y_{t_1}, ..., Y_{t_n}\}$ is the same as the joint distribution of $\{Y_{t_1-k}, ..., Y_{t_n-k}\}$ for all choices of time points $t_1, t_2, \ldots t_n$ and all choices of time lag k.

So, for $n = 1$, this means that Y_{t_1} and Y_{t_1-k} have the same statistical distribution, i.e. same means, same variances:

$$E(Y_{t_1}) = E(Y_{t_1-k}) = \mu,\ Var(Y_{t_1}) = Var(Y_{t_1-k}) \tag{3.2}$$

And, for $n = 2$, this means that $\{Y_t, Y_s\}$ has same distribution as $\{Y_{t-k}, Y_{s-k}\}$, so for example the covariances must be the same:

$$Cov(Y_t, Y_s) = Cov(Y_{t-k}, Y_{s-k}) \tag{3.3}$$

for all t, s and k. Putting $k = s$ and then $k = s$, we get

$$\gamma_{t,s} = Cov(Y_{t-s}, Y_0) = Cov(Y_0, Y_{t-s}) = Cov(Y_0, Y_{|t-s|}) = \gamma_{0,|t-s|} \quad (3.4)$$

This means that the covariance between Y_t and Y_s depends on time only through the time difference $\mid t - s \mid$ and not otherwise on the actual time points t and s. Thus a strictly stationary time series have mean, variance and covariance functions of simple forms:

$$\mu = E(Y_t),\ \gamma_k = Cov(Y_t, Y_{t-k}),\ \rho_k = Corr(Y_t, Y_{t-k}) \quad (3.5)$$

Simply said, if a time series is strictly stationary with finite variance, then the covariance and correlation depend on lag k only.

A time series $\{Y_1, Y_2, ...\}$ is said to be *weakly stationary* if

1. The mean function is constant over time, i.e. $E(Y_t) = \mu$, for all $t = 1, 2, ...,$ and
2. the covariance function depends on time via the time lag ($\gamma_{t,t-k} = \gamma_{0,t-k}$, for all t and any lag k).

Strict stationarity of time series means that the data-generating mechanism is not changing over time. In other words, the joint distributions for the time series are the same. If the joint distributions for the process are all multivariate normal distributions, then the two definitions (weakly and strongly stationary) coincide.

In future sections, we will assume that the time series data are a sum of trend, seasonal and cyclic components, as well as a random component. Thus the mean will not be stationary, and we will need a more complex model than $E(Y_t) = \mu$. We will discuss some statistical models that have a mean that is not a constant. We will assume that the mean can be characterised by a model with a finite number of parameters (remember simple linear regression has two parameters: intercept and slope). We will also assume that the error term component is weak and stationary with zero mean, constant variance and Normal distribution.

3.3 Exploratory data analysis for time series

As we mentioned in Section 3.1.2, an exploratory data analysis (EDA) is what we need to do before we do any modelling or forecast.

3.3.1 Goal of exploratory data analysis

Exploratory data analysis for time series is a preliminary stage of analysis of time data where the goals are

- to visually check for any outliers,
- to visually inspect any expected or unexpected patterns (trends, seasonal changes, cyclic changes),
- to visually inspect for the time of any sudden change in the patterns (such a point in time is called a "change point"),
- to check for any data quality problems (such as missing data, outliers),
- to check for patterns in the variance of data (is it constant or increasing or decreasing with changes in overall trend),
- to get some initial statistical properties of the time series data, such as an estimate of the autocorrelation and any relevant statistical significance tests.

Various tools have been developed for EDA of time series. One family of tools are called *Graphical Tools for Time Series* with the main visualisation tool being the **time series plot**, i.e. the plot of the data series versus time, such as in Figure 3.6. EDA also uses a method called *Time Series Decomposition* which decomposes the time series data into the three components: trend, seasonal and random (see Section 3.3.2).

After we do the EDA analysis, we show its results to the experts in the field. For example, we show our EDA results to the experts on the stock exchange if we want to forecast the prices of shares. Such discussion then leads to a list of potential models. Such models are then fitted and evaluated numerically to see which can be used for further analysis and for forecasting.

3.3.2 Trend, seasonal and cyclic components of time series

Trend component. A trend exists when there are long-term changes in the data. The trend does not have to be linear. It can be quadratic, polynomial, or any curve, really. Sometimes e.g. government may introduce a travel ban for six months, and that causes a change in the trend of the number of overseas visitors time series. A change in trend can be temporary, for six months, or a long term change.

Seasonal component. A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. Seasonality is always of a fixed and known frequency. In Figure 3.6 we give four examples of time series. The monthly sales of new one-family houses in the USA show seasonality with the smallest sales at Christmas time. The duration of these seasonal fluctuations is therefore 1 year. Another example of seasonal

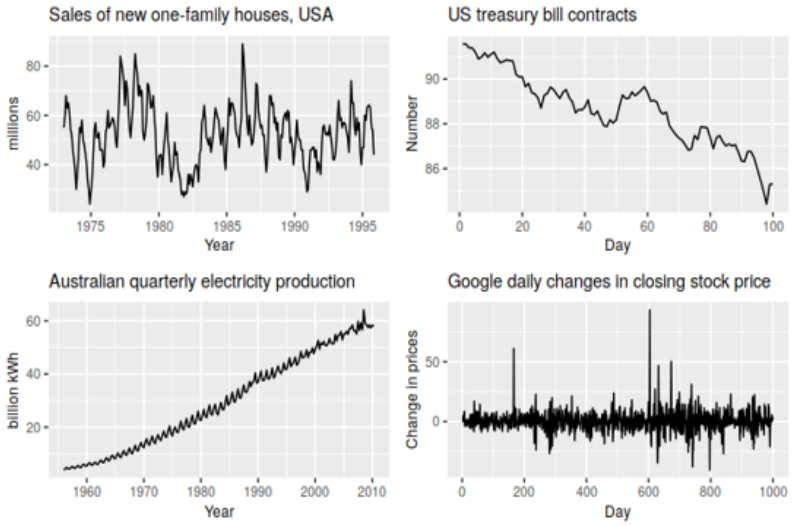


FIGURE 3.6: Time series examples. Top left: Monthly housing sales in the USA. Top right: Hourly US treasury bill contracts. Bottom left: Australian quarterly electricity production. Bottom right: Google daily changes in closing stock price. Source: [34].

changes is the Australian quarterly electricity production. The Google daily changes in closing stock price do not seem to show any fluctuations that would be 1 year long, at least it is not obvious to our eyes.

Naturally, not all time series are suitable to study seasonality such as The Hourly US treasury bill contracts in Figure 3.6. We need time series to be measured frequently enough (at least 4 times a year) and to be measured for a long enough time (for at least one year), to be able to judge if there is an effect of a season.

Cyclic component. A cycle occurs when the data exhibit rises and falls that are not of a fixed frequency. These fluctuations are usually due to economic conditions, and are often related to the "business cycle". The duration of these cyclic fluctuations is usually at least 2 years.

Noise component. When we see a variation in the time series that cannot be explained by overall trend, seasonality, by cycle then such variation is due to noise, also called random or residual variation. All previous components (trend, seasonal and cyclic) are non-random components (also called deterministic components). Noise is a random component of the time series. Random variation cannot be predicted but must be acknowledged and incorporated in the time series analysis, as more noise increases our uncertainty about the deterministic components.

What about a situation when the government issues a travel restriction? Such as in early 2020 due to Covid-19. Such travel restrictions caused a large

long-term change in the trend of time series (see Figure 3.2). Such change is not a random error, rather it is a change in trend caused by an external force, in this case, the government intervention. Such a change can be modelled by adding suitable covariates into the time series model.

Example. Four examples of time series. Using the time series in Figure 3.6 describe the trend, seasonality and cyclic components.

Solution. Possible answers:

- The monthly housing sales (top left) show strong seasonality within each year. There is no apparent overall increase or decrease in the data. There is some strong cyclic behaviour: the first full cycle lasts 7 years (1975-1962), and another full cycle is seen to last 9 years (1962-1991).
- The US treasury bill contracts (top right) show results from the Chicago market for 100 consecutive trading days in 1981. Here there is no seasonality, but an obvious downward trend. Possibly, if we had a much longer series, we would see that this downward trend is actually part of a long cycle, but when viewed over only 100 days it appears to be a trend.
- The Australian quarterly electricity production (bottom left) shows a strong increasing trend, with strong seasonality. There is no evidence of any cyclic behaviour here. Also, we see that the size of the seasonal fluctuations increases with the increasing trend.
- The daily change in the Google closing stock price (bottom right) has no trend, seasonality or cyclic behaviour. There are random fluctuations which do not appear to be very predictable and no strong patterns that would help with developing a forecasting model.

How do we estimate the components of the time series? There are various tools to quantitatively estimate and evaluate the components of time series, commonly called: *Time Series Decomposition tools*. The general idea is this: first, we will estimate the trend, m_t, then we will estimate the seasonal component, s_t, and then the remainder (i.e. the errors, or residuals). Note here, that the cyclic and trend component are estimated as one component.

Estimating the trend component. A simple method to estimate the trend m_t at time t is to take the average of Y_t and of its neighbouring values, a so-called *moving average*. This then poses a question of how many neighbours do we use, and if we use a simple average or a weighted average. Another way to estimate a trend is by using a linear or nonlinear regression model.

Example. Overseas visits. (continues) We will be estimating the trend via the moving average in the Overseas visits data.

Solution. For each Y_t (number of visits) we may try to calculate the following average:

$$Z_t = \frac{Y_{t-1} + Y_t + Y_{t+1}}{3} \tag{3.6}$$

which is called a moving average of order 3. In general, for an odd number m the moving average of order m is

$$Z_t = \frac{Y_{t-k} + Y_{t-k+1} + \ldots Y_t + \ldots Y_{t+k-1} + Y_{t+k}}{m}, \tag{3.7}$$

where $k = \frac{m-1}{2}$. If m is even, then the value from two moving averages (where k is rounded up and down respectively) are averaged, centring the moving average. It can be shown, that this is then:

$$Z_t = \frac{Y_{t-k} + Y_{t-k+1} + \ldots Y_t + \ldots Y_{t+k-1} + Y_{t+k}}{m}, \tag{3.8}$$

where $k = \frac{m}{2}$.

What order, m, should we choose for the moving averages? In Figure 3.7 we show several moving averages for the Overseas visits data. We should choose the moving average of order 12 as the best estimate of the trend in the overseas data. The reason for such a decision is that the seasonality pattern repeats every 12 observations i.e. we have monthly data. The `R` code to produce these plots is in Section R Lab, Question (1).

Estimating the seasonal component of time series. Now that we have the trend estimated, we can try and estimate the seasonal component. We will proceed in steps:

1. We start by removing the trend from the time series. This will give just the data that contain seasonal component and the noise, hence we will call it *detrended series* and we will use the notation D_t hence we have:

$$D_t = Y_t - Z_t \tag{3.9}$$

2. Then from this detrended series, we estimate the *raw seasonal factors*. Each seasonal factor is the mean over all seasonal factors for that season. In our example, we have 12 seasonal factors, as the season repeats every 12 observations (Why? Because we observe 12 values in each year). So in our example for January, we have the following estimated raw seasonal factor value:

$$F_j = \frac{\sum_{all\, t\, belonging\, to\, month\, j} D_t}{n_{years}} \tag{3.10}$$

where for January we have $j = 1$,... December has $j = 12$, n_{years} is the number of years for which we have data.

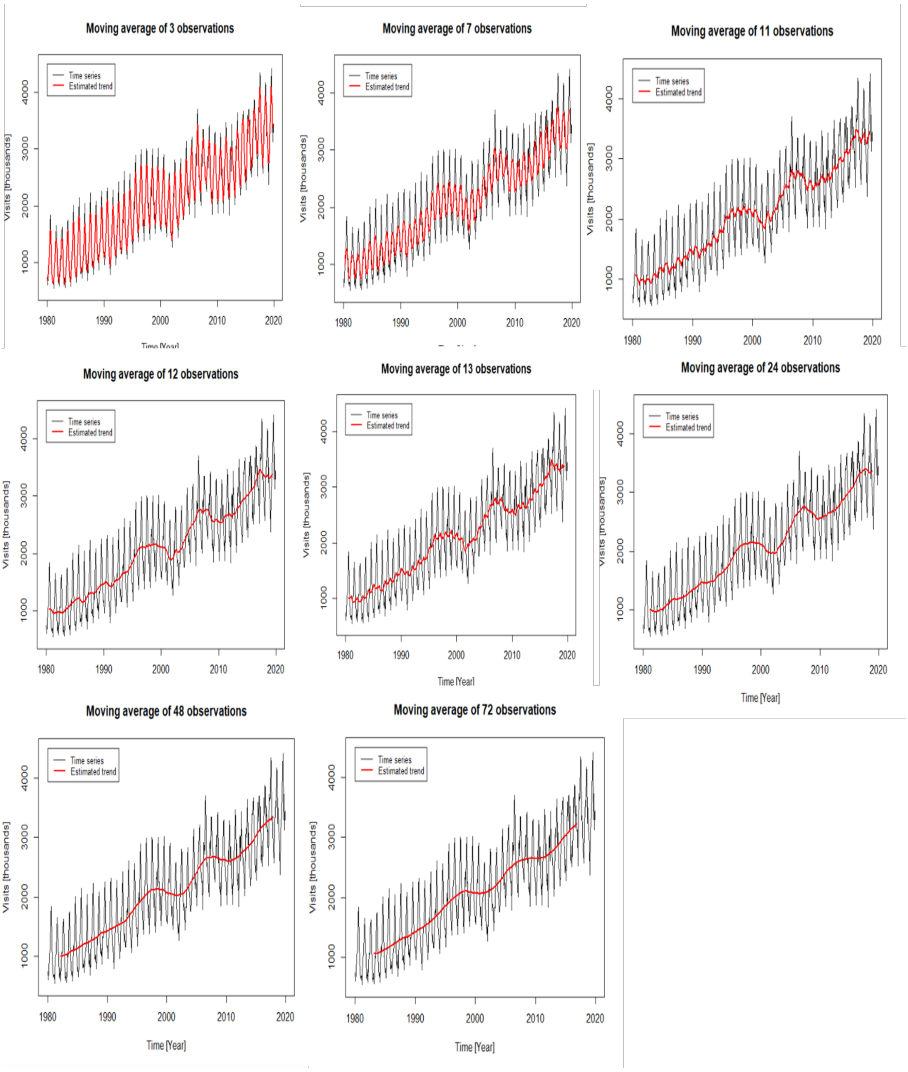


FIGURE 3.7: Overseas monthly visits to UK January 1980 to December 2019 together with estimated trend via moving averages technique.

3. Then we use the raw seasonal factors to get the values of the *corrected seasonal factors:*

$$s_j = F_j - \bar{F} \tag{3.11}$$

where $\bar{F}$ is the mean over all F_j, $j = 1, \ldots, J$. Note that $J = 4$ for quarterly data, and J=12 for monthly data.

Estimating the noise (the remainder) component. Once we have

the estimate of trend and seasonality we are able to estimate the noise. We do it simply by taking this difference:

$$w_t = D_t - s_t \tag{3.12}$$

Next, we can create a seasonally adjusted time series. Seasonally adjusted time series are defined as series that contain a trend and noise only. We obtain it by taking the difference:

$$U_t = Y_t - s_t \tag{3.13}$$

Note that sometimes when people (organisations) post time series data on the internet, they often post seasonally adjusted time series.

Decomposition of time series, in a nutshell. When we are asked to do time series decomposition, we need to provide four plots: original observed series, estimated trend, estimated seasonal component, and estimated noise. This can be conveniently done in `R` using function `decompose`, which is part of library `forecast` and which gives the decomposition of Overseas visits as one Figure as we see in Figure 3.8. For `R` details, see the Question 1, in Section R Lab.

In the Overseas visits example, the decomposition (Figure 3.8) showed that the trend is roughly linear with potentially some cyclic fluctuations, and there is a strong seasonal component. The seasonal component does not appear to increase or decrease with trend. All this should help us to choose model candidates in the next Section 3.4.

3.3.3 Estimating the autocorrelation

We defined the autocorrelation function for time series data in Equation 3.1 of Section 3.2.3. Here we show how to estimate it, interpret it, and how to decide if the time series are autocorrelated.

The autocorrelation (Eq. 3.1) can be estimated from a sample of time series data $y_1, \ldots, y_T$ via the following formulae

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2} \tag{3.14}$$

where T is the length of the time series. Some books denote it as r_k. The parameter k is called a *lag*. The term "auto" in autocorrelation comes from the fact that it is a correlation of time series with itself.

In `R` there is a function `Acf` (also `acf`) that automatically calculates the autocorrelation for all the lags that the user specified. Then the values of autocorrelation for several lags is called the *autocorrelation function* or *autocorrelogram*.

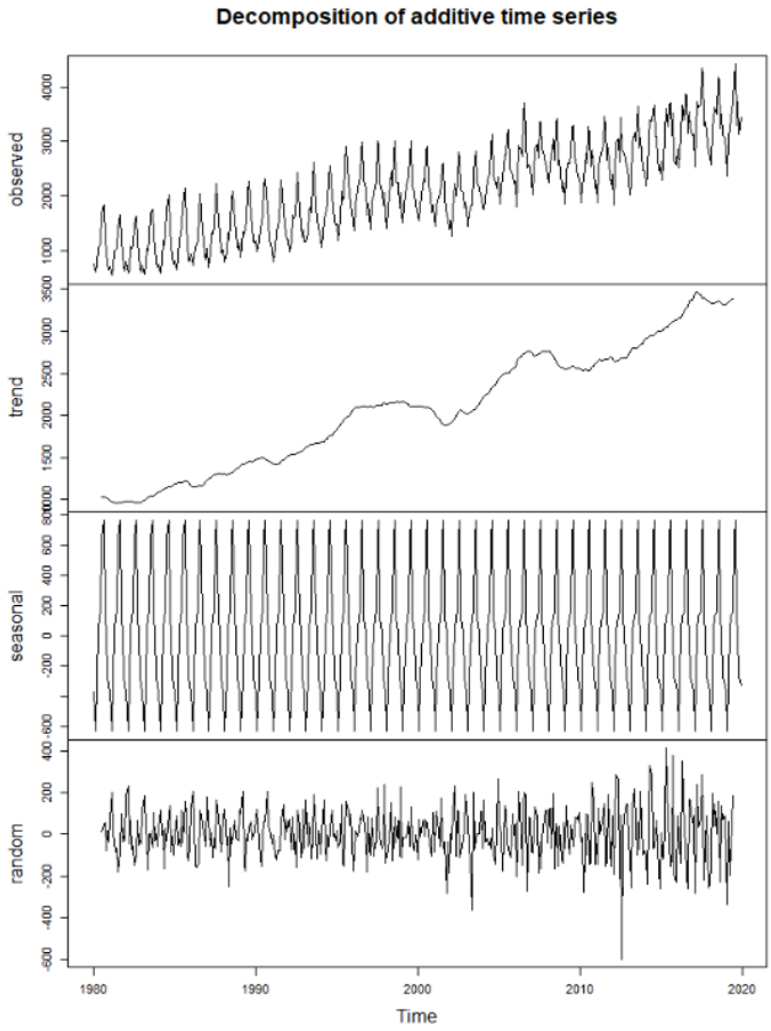


FIGURE 3.8: Time series decomposition of overseas monthly visits to the UK. The data are used for January 1980 to December 2019). This decomposition was done by using built-in function `decompose`.

Interpretation of autocorrelation of a time series data is similar to Pearson correlation of data on two variables. In the Overseas visits example, a positive autocorrelation between Y_t and Y_{t-1} means the following: if the number of visits at time t is large (above the overall trend) then the number of visits at time $t-1$ is also large (above the overall trend), roughly speaking (most of the time). We call this a *positive autocorrelation at lag 1*. It is called

lag 1, because the difference between the times t and $t-1$ is equal to 1. We will be looking at any lag e.g. $k = 1, 2$ or more. We explain the autocorrelation further on time series data in the next examples.

Example: Overseas visits. (continues) Here, we use data from Jan 1980-July 1980 only i.e. the first seven time points. We calculate autocorrelation manually at lag 2 (so $k = 2$), using the formulae (3.14). To avoid too long calculations, for the educational purpose, we will now pretend that we only have seven values of Overseas visits monthly time series data:

$$\mathbf{y} = (739, 602, 740, 1028, 1088, 1124, 1699)$$

Solution. We know that autocorrelation is the correlation of the time series with its lagged values. So for our calculations, we could easily create a table with the original and lagged series in separate columns and then use the formulae 3.14. We start with generating the Table 3.2 of the lagged series.

Time	y_t	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-4}	y_{t-5}	y_{t-6}
1	739	NA	NA	NA	NA	NA	NA
2	302	739	NA	NA	NA	NA	NA
3	740	302	739	NA	NA	NA	NA
4	1,028	740	302	739	NA	NA	NA
5	1,088	1,028	740	302	739	NA	NA
6	1,124	1,088	1,028	740	302	739	NA
7	1,699	1,124	1,088	1,028	740	302	739

TABLE 3.2: Lagged time series.

Next, we use the Table 3.2 to create another Table 3.4, which contains the necessary calculations for the autocorrelation at lag $k = 2$. One of the first calculations we need to do is the sample mean

$$\overline{y} = 7020/7 = 1002.8571$$

then for $k = 2$ we have the covariance estimate

$$\sum (y_t - \bar{y})(y_{t-2} - \bar{y}) = 91,672.24$$

and variance estimate

$$\sum_{t=1}^{T} (y_t - \bar{y})^2 = 806,572.86$$

so the autocorrelation estimate at lag $k = 2$ is

$$\hat{\rho}_2 = \frac{91,672.24}{806,572.86} = 0.1230$$

Time	y_t	y_{t-2}	$y_t - \overline{y}$	$y_{t-2} - \bar{y}$	$(y_t - \bar{y})(y_{t-2} - \bar{y})$	$(y_t - \bar{y})^2$
1	739	NA	-263.86	NA	NA	69,620.59
2	602	NA	-400.86	NA	NA	160,686.45
3	740	739	-262.86	-263.86	69,356.73	69,093.88
4	1,028	602	25.14	-700.86	-17,621.55	632.16
5	1,088	740	85.14	-262.86	-22,380.41	7,249.31
6	1,124	1,028	121.14	25.14	3,045.88	14,675.59
7	1,699	1,088	696.14	85.14	59,271.59	484,614.88
Sum	7,020	3,897	0.00	-117.29	91,672.24	806,572.86

TABLE 3.4: Calculations for autocorrelation at lag 2.

So the autocorrelation at lag 2 was estimated to be 0.1230. This estimate is calculated from the first seven values of the Overseas visits data, it is different from zero and positive, suggesting a positive correlation in this set of seven time series values. Due to random variation in time series, there is a possibility that the value 0.123 could have happened by chance, so at the moment we cannot generalise to time series beyond our seven data values, and thus we cannot conclude that there is a correlation at lag 2. In order to see if we can generalise our finding of 0.123, we need to do a statistical significance test which will establish if the size 0.123 could have happened by chance or not - which we do next.

Test of significance of the autocorrelation values. Is the autocorrelation present or not in Overseas visits data? In other words, is the estimated value of autocorrelation large enough to be not caused by random chance, but to be likely caused by the actual correlation? This is impossible to judge from the estimated values alone (see the value 0.123 above). Why? Owing to random fluctuations and finite T time series, the autocorrelations are never going to be exactly zero. But we still want to say if an autocorrelation is present or not (i.e. the autocorrelation when T is infinite or very large). What we need to do is to test if autocorrelation is far enough away from zero to say that it is present. This means we need a test of significance. A test of significance for autocorrelation can be done in steps as follows:

1. Let the autocorrelation estimate be denoted as $\hat{\rho}_k$. Then $\hat{\rho}_k \approx N(0, 1/T)$ under the null hypothesis $H_0 : \rho = 0$. In other words, if truly there is no correlation (i.e., $\rho_k = 0$), the estimate $\hat{\rho}_k$ has an approximately normal distribution, with mean zero and variance equal to $1/T$. So if the k -lagged population correlation ρ_k is zero, then the correlation estimate should be small, and its variance is equal to $1/T$.

2. Since we know the $\hat{\rho}_k$ distribution under H_0 we can calculate p-values, but this is not normally done for autocorrelation. Instead, what tends to be done is to show confidence intervals. So for example, the 95% confidence

bands for autocorrelation are calculated as $0 \pm 1.96 \times$ *standard error of the estimate* $\hat{\rho}_k$ is equal to $0 \pm 1.96\sqrt{\frac{1}{T}}$. Hence the 95% confidence bands for autocorrelation are

$$\left(-\frac{1.96}{\sqrt{T}}, \frac{1.96}{\sqrt{T}}\right) \tag{3.15}$$

Then such confidence bands are plotted together with the estimated autocorrelation values $\hat{\rho}_k$. Note, that the 95% confidence band does not depend on the lag k so it is the same for all lags (the same can be said for e.g. 99% confidence etc.). The 95% confidence band was calculated under the assumption of zero autocorrelation. So if there is no autocorrelation, then 95% of estimated autocorrelations must lie within the 95% confidence bands (see the blue vertical dashed lines in the next examples (Figures 3.9, 3.10).

3. The last step is to interpret the confidence band from equation 3.15, i.e. to make a conclusion of the test of significance of the autocorrelation values. Assume we calculated autocorrelation for lags $1, \ldots, k$. The interpretation of the test of significance of autocorrelation is as follows: If at least 95% of the estimated autocorrelations lie inside of these 95% confidence bands, then we conclude no significant autocorrelation in the time series for lags 1 to k, i.e. we conclude that the values are not autocorrelated. If more then 5% of the estimated autocorrelations lie outside of these bands, then we conclude significant autocorrelation, i.e. we conclude that the time series are autocorrelated. These vertical lines are also called *confidence bands for autocorrelation function*, as they illustrate how much the estimated autocorrelation can vary due to random chance: more specifically the 95% bands show where at least 95% of all estimated autocorrelation values should be if there is no autocorrelation.

Example: Overseas visits. (continues) Again, for illustration purpose, we consider only the first seven time points from January 1980 to July 1981. Are the overseas visits time series autocorrelated?

Solution. We calculate the 95% confidence bands for the autocorrelations:

$$\left(-\frac{1.96}{\sqrt{T}}, \frac{1.96}{\sqrt{T}}\right) = \left(-\frac{1.96}{\sqrt{7}}, \frac{1.96}{\sqrt{7}}\right) = (-0.74, 0.74)$$

All six estimated autocorrelations are (not done here, but you should try to calculate them all by hand) are

$$(\hat{\rho}_1, ..., \hat{\rho}_6) = (0.374, 0.123, -0.068, -0.315, -0.386, -0.228)$$

As we see, all estimated autocorrelations are less than 0.74, in absolute value, so none of them is considered significantly different from zero. In other words, we do not have enough evidence to say that the observed values suggest the presence of autocorrelation in the time series. Consequently, this means that

the values $0.374, 0.123, -0.068, -0.315, -0.386, -0.228$ should not be interpreted i.e. we cannot proceed to say that 0.374 is a positive and weak correlation, because we just found that this could have been caused by a random chance!

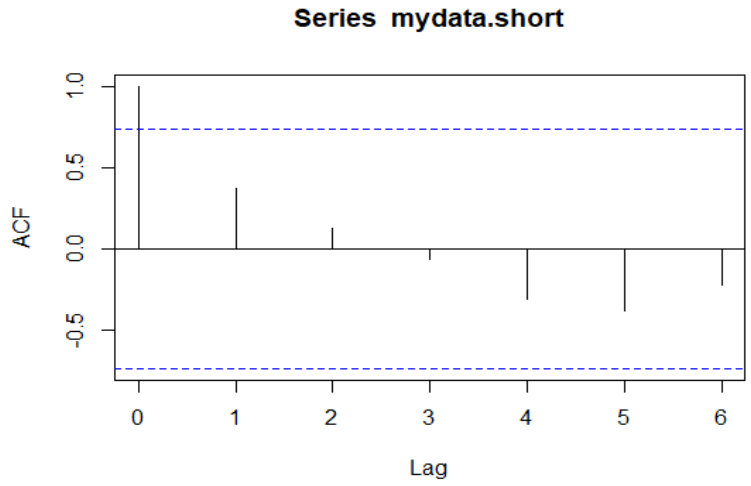


FIGURE 3.9: ACF function for monthly overseas visits data for January 1980- July 1980, hence 7 months only. The 95% confidence bands for autocorrelation are also showed (see the two blue dashed vertical lines). This plot is also called *autocorrelogram*.

In `R` the above calculations can be done via a built-in function called `acf` (see the Section R Lab, at the end of this chapter, for more details). This creates the following Figure 3.9. Note that the plot also shows the autocorrelation for time $t = 0$, and this must be equal to 1, which will always be out of the confidence band, but this will be ignored in our interpretation, as such autocorrelation is not being tested. Note that all other 6 values of autocorrelation are within the 95% confidence band. Thus we conclude that there is no evidence of autocorrelation in the first 7 values of overseas visits data.

Example. Overseas visits. (continues) Next, we consider data from January 1980 to December 1981, and also data from January 1980 till March 2020. We now have more data and we are asked the same question: are data autocorrelated?

Solution. Now, we will use all the data to plot the autocorrelation function. The following Figure 3.10 was done in `R`, of which details are in Section 1. Are there any significant correlations? Are there any patterns in the auto-correlation function?

If we use all 40 years of Overseas visits data (Figure 3.10, the plot on the

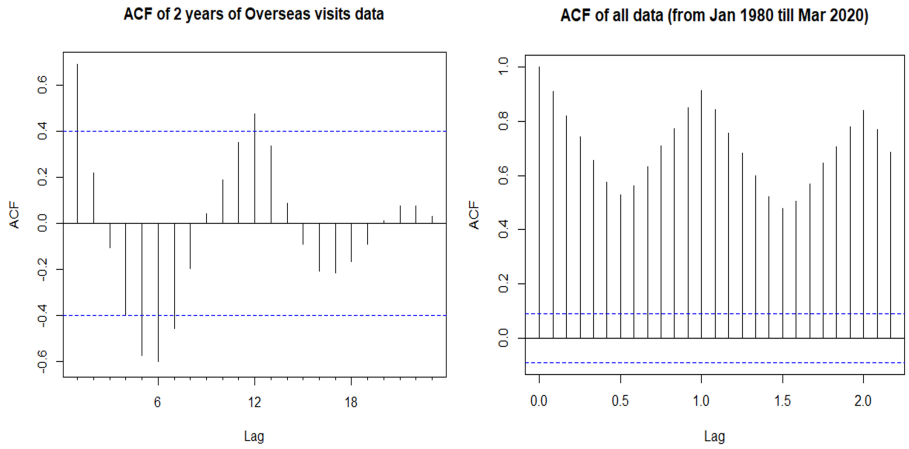


FIGURE 3.10: ACF function for overseas monthly visits to UK data. In the plot on the left, we only used the first two years of data to calculate ACF. On the right, we used all 40 years of data.

right), then all autocorrelations are above the 95% confidence band, hence more than 95% of the autocorrelations are above the 95% confidence band. This means that the autocorrelation is present in the time series and it is positive. We also see a pattern in the autocorrelations: they are the highest at lags k =12 and 24, this means that if we look at all Januaries, they have the strongest correlation, also all Februaries have the strongest correlation etc. The autocorrelation is the smallest for time lags 6 and 18. This means that January and July have the smallest correlation, as well as February and August etc.

If we use only the first two years of Overseas visits data (Figure 3.10, the plot on the right), then not all autocorrelations are above the 95% confidence band. In fact, 4 autocorrelations of 19 are outside of the 95% band, which is more than 5%. Hence there is evidence of autocorrelation in the first two years of overseas data.

We also note that the confidence band is smaller as we use more data i.e. as T is larger. This is consistent with the fact that the confidence bands do have T in the denominator (see Equation 3.15). This means that, if data are really correlated, then with more data we have stronger evidence. When we used only 7 data points, we had no evidence of correlation.

Why do we need the autocorrelation plots and the test of the autocorrelation? We have seen that the autocorrelation plots are used for checking the correlation of values time series dataset. It is not surprising if

we find significant autocorrelations in our time series because we measure the same thing (e.g. monthly number of visitors) repeatedly over time.

If we however do not find the time series data to be autocorrelated, then we may try to describe the time series data with linear regression models (e.g., linear or quadratic or polynomial in Section 3.4.1). This is because one of the assumptions of the linear regression models was that the data are not correlated (in the case of normality: uncorrelated means independent). More often than not the time series data will be autocorrelated and thus not independent. The time series discipline was developed to deal with correlated data (e.g., exponential smoothing models in Section 3.4.2). If we however find time series data uncorrelated, then we have another battery of tools at our disposal: time series regression models, in addition to time series tools.

In future sections, we will see that autocorrelation plots are crucial for the model's goodness-of-fit checking where we will construct an autocorrelation plot for the residuals of the model. As always, a residual is defined as the difference between the actual value minus the fitted value of the time series. This will be shown in later sections of this chapter.

3.4 Time series modelling and forecasting

In this section we introduce two tools for time series modelling: time series regression models and time series exponential smoothing models. In each, we show how the model is specified, how it is fitted to the data, how we choose the best fitting model, how we do a goodness-of-fit analysis of the best fitting model, how the forecasting is done and how we present and communicate the forecast to the stakeholders.

3.4.1 Time series regression models

When time series data are not autocorrelated we can use regression models to explain the patterns in the data. Then we can use the framework of regression models to construct prediction intervals which will serve as a forecast.

The basic concept is that we forecast the output or dependent time series of interest Y_t $(t = 1, 2, ...)$ assuming that it has a linear relationship with a collection of other independent data series (also called inputs) $X_{t1}, X_{t2}, ..., X_{tq}$. For example, we might wish to forecast monthly sales Y using total advertising spend X as a predictor. Or we might forecast daily electricity demand Y using temperature X_{t1} and the days X_{t2} as predictors. We can explain the relationship via a *multiple linear regression model*

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + ... + \beta_q X_{tq} + W_t \tag{3.16}$$

where $\beta_0, \beta_1, ..., \beta_q$ are unknown fixed regression coefficients. The component

W_t is a random error or noise process. We assume that all values w_t are independent, identically distributed from a Normal distribution with zero mean, and variance σ_w^2. Such noise is called the *white noise*. For time series regression, it is rarely the case that the noise is white, although it does happen sometimes.

Example. Kings' life span. We have data on how long each of the 42 kings lived, in England. The ages values are written in the order in which the kings sat on the throne:
60, 43, 67, 50, 56, 42, 50, 65, 68, 43,
65, 34, 47, 34, 49, 41, 13, 35, 53, 56,
16, 43, 69, 59, 48, 59, 86, 55, 68, 51,
33, 49, 67, 77, 81, 67, 71, 81, 68, 70,
77, 56.

The time series plot is in Figure 3.11. Next we will fit linear and quadratic regression models of the Kings' life span data.

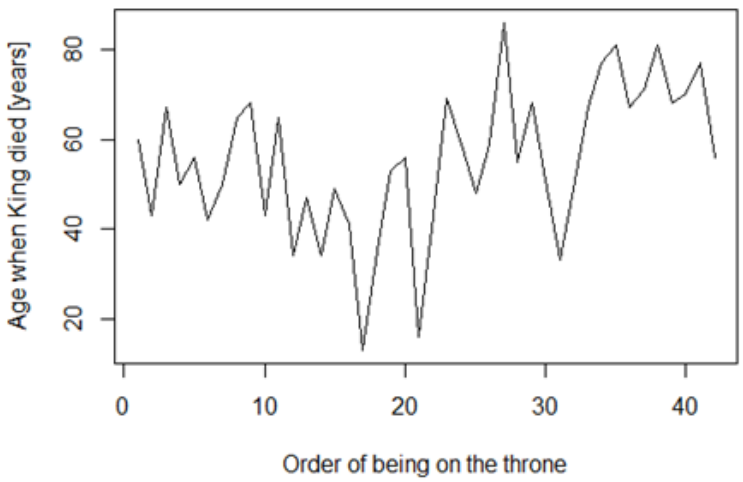


FIGURE 3.11: Kings' life span data.

Solution. The best-fitting linear model is estimated by using times of death as the dependent variable. We use the order in which the king is on the throne as an explanatory variable (values $1, \ldots, 42$), and we will call it $Time$.

The `R` code and output are provided at the end of this chapter, in R Lab Question 2 (in Section 3.8). The best-fitting linear regression model is:

$$y_t = 43.5679 + 0.5450 \times Time + w_t \tag{3.17}$$

The model has highly significant coefficients for the intercept and the independent predictor variable Time (both $p - values < 0.01$). Hence the whole

model has a significant p-value ($0.00805 < 0.01$), which means that Time explains a significant amount of variability in Y. However it is only 14% of the variance of Y that is explained by a linear relationship with Time.

Next, we plot the time series again, but now we overlay the plot with the fitted values in Figure 3.12. We see the linear model is not fitting the data well. The fitted line (the blue solid line in the figure) is underestimating the ages for the first 9 kings (except for 2) then is overestimating for the kings 10 to 25, and then again is underestimating for the kings 35 to 42. We see this also on the plot of residuals vs time where we define the residuals as:

$$w_t = y_t - \hat{y}_t \tag{3.18}$$

For the first 9 kings, the residuals are mostly positive (values 1 to 9) thus yielding a positive non-zero mean, then for the kings 10 to 25 the residuals are mostly negative thus yielding a negative non-zero mean, and for kings 26 to 42 the residuals are again mostly positive. Hence, this suggests a quadratic pattern in residuals i.e. there is a pattern in the Kings data that we did not describe. So next we add a quadratic term into a statistical model for Kings' life span data. In other words, we will fit a quadratic model to the King's life span data.

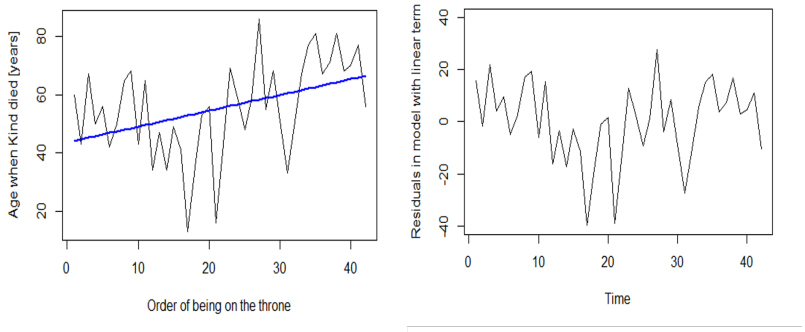


FIGURE 3.12: Kings life span data together with a linear model (left) and residuals vs time plot (right).

Example. King's life span. (continues) In next, we fit a quadratic model to the Kings' life span time series data.

Solution. A quadratic model is a special case of the general linear model. This is because a quadratic model is linear in all beta parameters. The estimation was done in `R` of which details are in R Lab Question 2 in Section 3.8. The estimated quadratic model for Kings' lifespan data is

$$Y_t = 57.61934 - 1.37109 \times Time + 0.04456 \times Time^2 + W_t \tag{3.19}$$

By visual inspection of Figure 3.13 we conclude that however good the linear model was, a quadratic model performs even better, as it appears to explain the additional pattern. It is not enough to do a visual inspection as it can be subjective. So later we will do a quantitative comparison of the two models (Section 3.4.3).

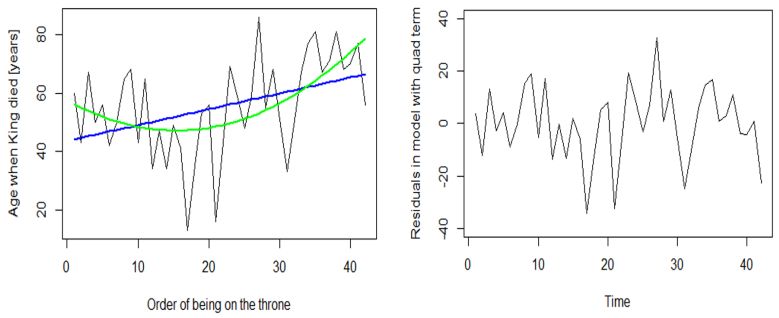


FIGURE 3.13: Kings life span data together with linear (blue line) and quadratic model (green) are shown on the left. Residuals from the quadratic model are shown on the right.

3.4.2 Exponential smoothing models of time series

Exponential smoothing was developed in the late 1950s (see [12], [32], [63]). It has been a motivation for some of the most successful forecasting methods. The main principle is that the exponential smoothing forecasts are weighted averages of past observations. The recent observations get the higher weights, and the older observations get the lower weights. The weights decrease exponentially as the observations get older. This framework is known to generate reliable forecasts quickly, however, the reliability is established after we find all assumptions to be satisfied. The exponential smoothing modelling framework is suitable for a wide range of time series, which is a great advantage and of major importance to applications in economics, industry, society and health.

In this section, we present the mechanics of the most important exponential smoothing models: simple exponential smoothing, Holts' smoothing and Winter Holts' smoothing model.

The goal of prediction (i.e. forecast) is trying to see what the value of a series will be in the future. Each of the exponential smoothing models in this section can generate point predictions (forecasts) as well as prediction intervals. This does rely on assumptions as to how the data were generated.

Simple exponential smoothing (SES). Simple exponential smoothing (sometimes called single exponential smoothing) is a very simple time series method that can be useful for prediction. It is used when time series data do not appear to have a trend or cyclic or seasonal component. For example, the data in Figure 3.4 do not show any clear trending behaviour or any seasonality. They do suggest some increase and decrease in the second half of the data, which might suggest a quadratic trend. We will consider whether a trended method would be better for this series later in this section.

Example: Daily temperatures. We will introduce a simple exponential model on real data. We have data on 31 consecutive days, in central England, in 2004:

$$\begin{aligned}(y_1, \ldots, y_T) =& (17.3, 17.9, 17.3, 15.4, 15.0, 17.6, 18.2, 17.2, 16.6, 15.7, \\ & 15.1, 16.8, 17.2, 18.7, 19.4, 18.3, 17.9, 18.5, 20.3, 19.5, 19.2, \\ & 20.2, 19.8, 20.2, 21.7, 19.8, 19.7, 18.3, 19.3, 17.3, 18.5)\end{aligned}$$

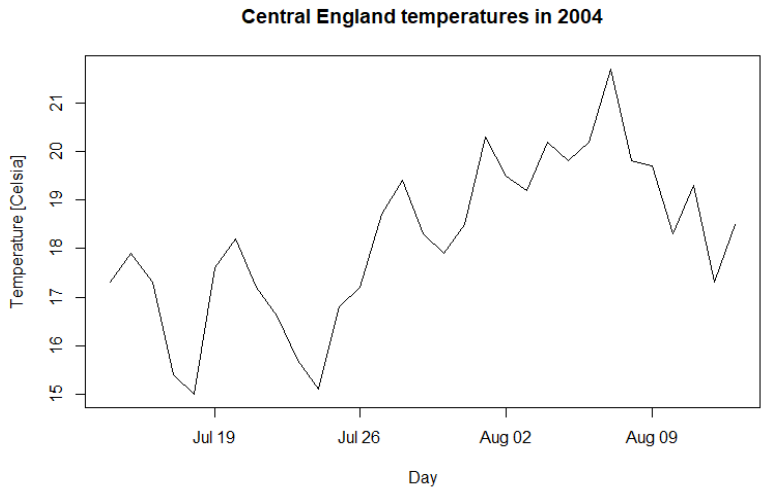


FIGURE 3.14: Central England daily temperatures time series data.

Solution. In the next, we, therefore, assume the Daily Temperature time series data follow no trend (i.e. we assume a constant trend), no seasonality and no cyclic pattern. Our goal is to estimate the values at times $T+1, ..., T+h$, i.e. the values $y_{t+1}, ..., y_{t+h}$. We will denote such estimates as $\hat{y}_{t+1|T}, ..., \hat{y}_{T+h|T}$, and we call them the forecasts.

Note, that the absolute simplest method for the forecast (at $h = 1, 2, \ldots$) is

$$\hat{y}_{T+h|T} = y_T \tag{3.20}$$

i.e. to naively assume that the most recent observation is the only important one for the forecast and that all previous observations have no importance. Such a forecast can be seen as a weighted average where all weight is put on the most recent observation. But what if the most recent value is a freak value and just due to a large amount of noise? Then such a simple model is not useful. We should try and look at past values y_1, ...,y_{t-1} as well, not just at the current value, y_t. The past values should have some information about the future of the time series as well.

Another simple method for the forecast (at $h = 1, 2, \ldots$) is

$$\hat{y}_{T+h|T} = \frac{1}{T}\sum_{t=1}^{T} y_t \tag{3.21}$$

This forecast put equal weight on all previous observations. However, this may not be a good idea as the most recent observations intuitively should be more relevant for the forecast than the old ones.

We really want something between these two extremes. We want to put larger weights on more recent observations than on observations from the distant past. This is the idea behind simple exponential smoothing. Forecasts are again calculated using weighted averages, however, the weights decrease exponentially as observations come from further in the past.

In the simple exponential smoothing method, we will assume that the overall trend is constant and that that there is no seasonality and no cyclic component in the time series y_1, ...,y_T. Our aim is to estimate the future values at times $T + 1, ..., T + h$. We will estimate the constant trend, and in doing so we will use all observations y_1,... ,y_T but we will give them various weights. For example, we predict y_{T+1} as a weighted sum:

$$\hat{y}_{T+1|T} = c_0 y_T + c_1 y_{Y-1} + c_2 y_{T-2} + \cdots \tag{3.22}$$

where $c_0, c_1, c_2, \cdots$ are the weights. We want to put higher weights on the more recent values and low weights on older values. In simple exponential smoothing, we use the following values, known as exponential weights

$$c_i = \alpha(1-\alpha)^i \tag{3.23}$$

which gives

$$\hat{y}_{T+1|T} = \alpha Y_T + \alpha(1-\alpha)y_T + \alpha(1-\alpha)^2 y_{T-2} + \cdots \tag{3.24}$$

Table 3.5 shows the weights attached to observations for four different values of α when forecasting using simple exponential smoothing.

The exponential behaviour of the weights. The weights $\alpha(1-\alpha)^i$ (in Equation 3.24) decrease geometrically, exponentially. For any α between 0 and 1, the weights attached to the observations decrease exponentially as we go back in time, hence the name "exponential smoothing".

i	Y	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$	$\alpha = 0.9$
0	Y_t	0.1	0.2	0.5	0.8	0.9
1	Y_{t-1}	0.09	0.16	0.25	0.16	0.09
2	Y_{t-2}	0.081	0.128	0.125	0.032	0.009
3	Y_{t-3}	0.0729	0.0729	0.0625	0.0064	0.0009
4	Y_{t-4}	0.06561	0.06561	0.03125	0.00128	0.00009
5	Y_{t-5}	0.059049	0.065536	0.015625	0.000256	0.000009

TABLE 3.5: Weights for four values of α in simple exponential smoothing.

The sum of the weights is 1 for large t. Why? Using the property of geometric series we have:

$$\alpha \sum_{i=0}^{t-1} (1-\alpha)^i = \alpha \frac{1-(1-\alpha)^t}{(1-\alpha)} = 1-(1-\alpha)^t \tag{3.25}$$

In other words, the sum of the weights even for a small value of α will be roughly equal to 1 for any reasonably large sample size (i.e. reasonably high value of t). Small α (i.e. close to 0) means we put a small weight on the current value of y (i.e. y_t), and a lot of weight on past values $(y_1, \ldots, y_{t-1})$ when predicting y_{t+1}.

Alternative form for simple exponential smoothing: Weighted average form. It can be shown easily that the forecast at time $T+1$ is equal to a weighted average of the most recent observation and the previous prediction $\hat{y}_{T|T-1}$

$$\hat{y}_{t|t-1} = \alpha y_{t-1} + \alpha(1-\alpha) y_{t-2} + \alpha(1-\alpha)^2 y_{t-3} + \ldots \tag{3.26}$$

$$\hat{y}_{t+1|t} = \alpha y_t + (1-\alpha)[y_{t-1} + \alpha(1-\alpha) y_{t-2} + \ldots] \tag{3.27}$$

$$= \alpha y_t + (1-\alpha)\hat{y}_{t|t-1} \tag{3.28}$$

hence we can calculate $\hat{y}_{t+1}$ iteratively, as

$$\hat{y}_{t+1|t} = \alpha y_t + (1-\alpha)\hat{y}_{t|t-1} \tag{3.29}$$

and we can do forecasts iteratively too, as

$$\hat{y}_{T+1|T} = \alpha y_T + (1-\alpha)\hat{y}_{T-1|T} \tag{3.30}$$

Hence we showed that the forecast at time $T+1$ is equal to the weighted average of the most recent observation y_T and the previous prediction $\hat{y}_{T|T-1}$.

Another alternative form for simple exponential smoothing: Component form. An alternative representation of time series is to write

them in a component form. For simple exponential smoothing, the only component included is the level, ℓ_t. The component form of simple exponential smoothing is given by *forecast equation*

$$\hat{y}_{t+h|t} = \ell \tag{3.31}$$

and the *smoothing equation*

$$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1} \tag{3.32}$$

where ℓ_t is the **level of the series** at time t. We can also interpret it as the smoothed value of the time series at time t. When we set $t = T$ then we get forecasts beyond the training data.

The forecast equation shows that the predicted value at time $t+1$ is the estimated level at time t. The smoothing equation for the level (also referred to as the level equation) gives the estimated level of the series at time t. If we replace ℓ_t with $\hat{y}_{t+1|t}$ and ℓ_{t-1} with $\hat{y}_{t|t-1}$ in the smoothing equation, we will get the weighted average form of simple exponential smoothing. The component form shown here is not helpful now, but will be helpful in the next two sections, as we add further components.

Forecasting from Simple Exponential Smoothing model. We want to predict the future, e.g. beyond the current data. A simple exponential smoothing model has a "flat" forecast function:

$$\hat{y}_{T+h|T} = \hat{y}_{T+1|T}, \qquad h = 2, 3, \ldots \tag{3.33}$$

This function gives so a called *point forecast*, i.e. just one value (point) for each time in the future. This function is flat, which means that all forecasts have the same value, equal to the last level component from time T. Importantly, these point forecasts will only be suitable if the time series has no trend and no seasonal component.

Example: Daily temperature data. (continues) We want to fit a simple exponential smoothing model to the daily temperature data, and to find the best smoothing parameter $\alpha = 0.6$. We can use an initial value y_0=y_1.

Solution. We need to calculate $\hat{y}_1, \hat{y}_2, \hat{y}_3, ..., \hat{y}_{31}$.Then we calculate the predictions by iterating equations (Equation 3.30) for $t = 1, ..., 32$:

$$\begin{aligned}
\hat{y}_1 &= \alpha y_1 + (1-\alpha)y_0 = 0.6 \times 17.3 + 0.4 \times 17.3 = 17.3 \\
\hat{y}_{2|1} &= \alpha y_1 + (1-\alpha)\hat{y}_1 = 0.6 \times 17.3 + 0.4 \times 17.3 = 17.3 \\
\hat{y}_{3|2} &= \alpha y_2 + (1-\alpha)\hat{y}_{2|1} = 0.6 \times 17.9 + 0.4 \times 17.3 = 17.66 \\
\hat{y}_{4|3} &= \alpha y_3 + (1-\alpha)\hat{y}_{3|2} = 0.6 \times 17.3 + 0.4 \times 17.66 = 17.44 \\
\hat{y}_{5|4} &= \alpha y_4 + (1-\alpha)\hat{y}_{4|3} = 0.6 \times 15.4 + 0.4 \times 17.44 = 16.22 \\
&\vdots
\end{aligned}$$

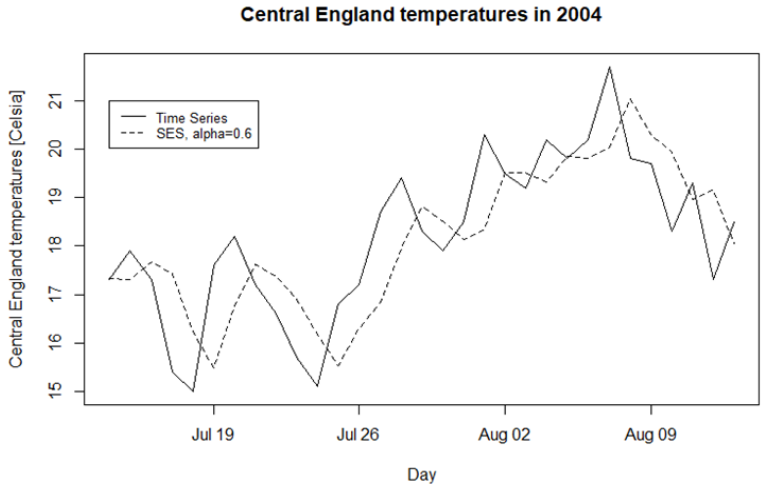


FIGURE 3.15: Daily temperatures time series (solid line) with simple exponential smoothing (dashed line), using parameter $\alpha = 0.6$.

Finding the optimal α, i.e. the optimisation. We have not discussed the choice of α yet. Obviously, $0 \leq \alpha \leq 1$. Large α means we put a lot of weight on the current observation, and small weight on the past. In other words, the speed at which the older observations are dampened (smoothed) is a function of the value of α. When α is close to 1, dampening is quick and when α is close to 0, dampening is slow.

We can compare the prediction estimate $\hat{y}_{t|t-1}$ against the actual data value y_t. In other words, we find the errors (residuals)

$$w_t = y_t - \hat{y}_{t|t-1} \tag{3.34}$$

The errors can be positive or negative, so it is better to work with squared forecast error values

$$w_t^2 = (y_t - \hat{y}_{t|t-1})^2 \tag{3.35}$$

Finally, we sum them all up, for all values of t, to get the sum of squared forecast errors

$$SSE = \sum_{t=1}^{T} w_t^2 = \sum_{t=1}^{T} (y_t - \hat{y}_{t|t-1})^2 \tag{3.36}$$

The sum SSE depends on α as well as y_0. We want to find such α and y_0 that minimise the sum of squared errors. Although, if time series are long enough the effect of y_0 gets small for large T.

Unlike the regression model where there are formulas (the closed-form solution of so-called normal equations) which return the values of the regression

coefficients that minimise the SSE (Equation 3.36), finding the optimal α for the simple exponential smoothing model is a non-linear minimisation problem, and we need to use an optimisation tool to solve it. One simple way (though may not be accurate) is to do a grid search i.e. to calculate SSE at prespecified values of α such as $0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$. Another way to do the optimisation is to use a more sophisticated algorithm (such as Newton Raphson algorithm) which is implemented in function `ses` in `R` (the name `ses` stands for simple exponential smoothing) which not only finds the optimal α but also the optimal initial value y_0.

In the next example, we will have a look and see what happens at various values of α. We will judge which α seems the most optimal: via a subjective visual inspection and then via maximising SSE (i.e. grid search).

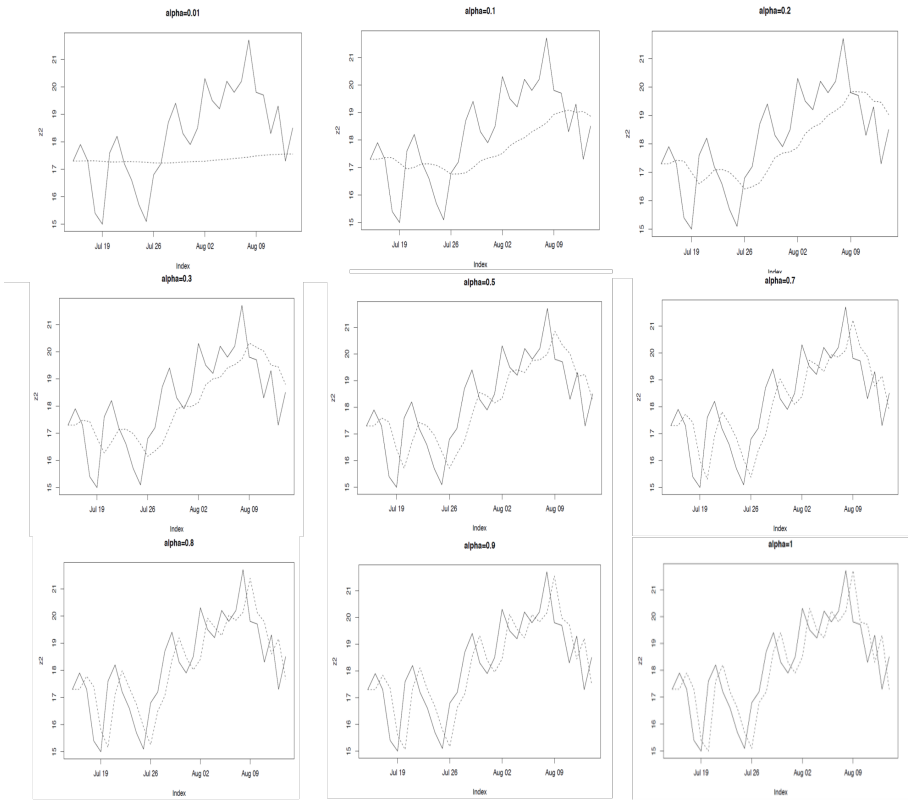


FIGURE 3.16: Daily temperatures time series and simple exponential smoothing at various α values.

Example. Daily temperatures. (continues) To get a visual insight into how simple exponential smoothing looks at various values of α we can

construct multiple plots in Figure 3.16. We note that $\alpha = 0.01$ is very small, so the predicted values are almost an average of past data. We see that α controls how much we use the current value, or average value for our next prediction. Checking the plots (such as Figure 3.16) should not be used to choose the α value. This is because choosing it by eye is not the best way of seeing the best α value: we get tired as we would have to construct many plots, and such an approach is subjective. We need a better system of finding the best α value. An objective way of finding the α value is to quantify numerically the errors and choose the α value that gives the smallest sum of squared errors.

α	SSE
0.01	99.50
0.05	79.69
0.1	65.14
0.2	51.88
0.3	46.10
0.4	43.16
0.5	41.59
0.6	40.74
0.7	40.31
0.8	40.22
0.9	40.50
1.0	41.22

TABLE 3.6: Central England daily temperatures data and SSE for several values of α in simple exponential smoothing.

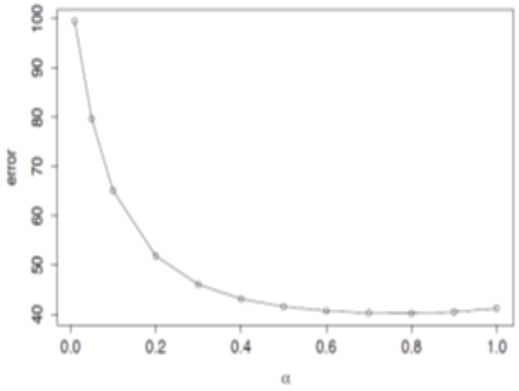


FIGURE 3.17: Daily temperatures time series and simple exponential smoothing SSE at various α values.

Hence, via doing our grid search, we found that the best choice of α is 0.8,

because SSE is the smallest (see Table 3.6 and Figure 3.17).

Caution! The optimal SES model is not the one that copies the data! In other words, the optimal model is not the one that is the closest to the data points. As we saw in Figure 3.17, the model closest to the data points is the model with $\alpha = 1$, i.e. the model that puts all weight on the most recent observation. However, that is not the optimal model. The optimal model is the one whose α minimises SSE, so it is $\alpha = 0.8$ in the Kings life span data example.

Holt's exponential smoothing model. Holt extended simple exponential smoothing into a model that is suitable for time series with a trend component (see [32]). Holt's method contains a forecast equation, and two smoothing equations (one for the level and one for the trend). For Holt's model, the *forecast equation* is

$$\hat{y}_{t+h|t} = \ell + hb_t \tag{3.37}$$

level equation

$$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \tag{3.38}$$

and *trend equation*

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1} \tag{3.39}$$

where ℓ_t is an estimate of the level of the series at time t, b_t is an estimate of the trend (slope) of the series at time t, α is the smoothing parameter for the level ($0 \leq \alpha \leq 1$), and β is the smoothing parameter for the trend ($0 \leq \beta \leq 1$).

Holt's smoothing is similar to simple exponential smoothing, but more complex and hence more flexible. It is more complex because it has an extra parameter β in addition to the parameters α and y_0. As we see from the level equation, α still controls how much we use the current value to estimate the level and how much we use the older values. The extra parameter β controls how much weight we put on the change in the last two levels (hence the slope) for our next prediction. The trend equation shows that b_t is a weighted average of the estimated trend at time t based on $\ell_t - \ell_{t-1}$ and b_{t-1}, the previous estimate of the trend. We use Holt's method when there is a trend in the data.

Finding the optimal values of parameters for Holts' exponential smoothing model. To find the optimal values of the parameters, we need to search over combinations of α, β and y_0 to see which gives the smallest SSE value. In other words, to find the optimal parameters we use the criterion of minimal SSE.

Forecasting from the Holts' exponential smoothing model. The forecast function (Equation 3.37) is no longer flat but it has a trend. The h-step-ahead forecast is equal to the last estimated level plus h times the last estimated trend value (see Equation 3.37). Hence, at each future time $T + h$,

the forecasts is a linear function of h.

Example. Daily temperatures. (continues) The data and the SES and Holt models' predicted values are plotted in Figure 3.18. Visually, it seems that both models appear to have very similar predicted values. So visually, it appears that there is not much difference in the way they fit the Daily temperatures time series data. We will use quantitative criteria to compare these two models, in Section 3.4.3.

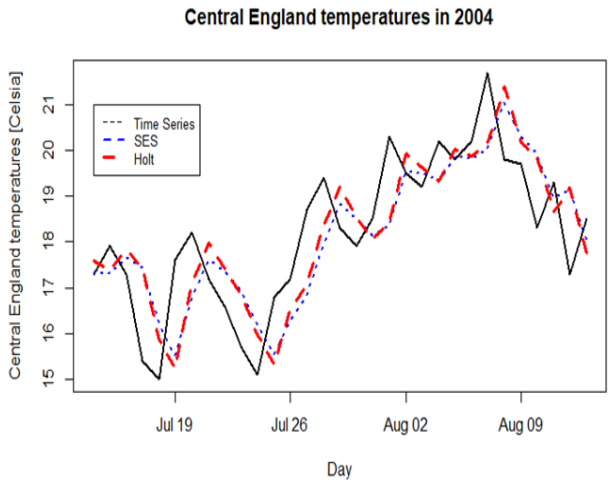


FIGURE 3.18: Daily temperatures time series with Simple Exponential and Holt smoothing.

Holt-Winter's seasonal exponential smoothing model. If in addition to the trend we also have seasonality in our time series data, then we should not use simple exponential smoothing or Holt's smoothing. We should use a so-called Holt-Winter's model. It is a model that was extended by Holt and Winters (see [63]). The component form representation of Holt-Winter's model contains a forecast equation and three smoothing equations (one for the level, one for the trend, and one for the seasonality). The *forecast equation* is

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(k+1)} \tag{3.40}$$

the *level equation* is

$$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \tag{3.41}$$

the *trend equation* is

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1} \tag{3.42}$$

and the *seasonal equation* is

$$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \tag{3.43}$$

where are now four parameters to optimise: α, β, γ and y_0. The parameters α and β play the same role as in Holt's exponential smoothing. The parameter γ controls how much we use the seasonal component in our next prediction ($0 \leq \gamma \leq 1-\alpha$). The seasonal equation shows a weighted average between the current seasonal index, $(y_t - \ell_{t-1} - b_{t-1})$ and the seasonal index of the same season last year (hence m time periods ago).

Finding the optimal values of parameters for Holt-Winter's seasonal exponential smoothing model. To find the optimal values of the parameters, we need to search over combinations of α, β, γ and y_0 to see which gives the smallest SSE value. This optimisation is already implemented in function `ses` `R` (see in Section R Lab).

Forecasting using Holt-Winter's model. The forecast function is not flat, it has a trend and seasonality. The $h-$step ahead forecast is calculated via the forecast Equation 3.40 and is equal to the last estimated level plus h times the last estimated trend value, plus the estimated seasonality.

Example. Overseas visits. (continues) We next fit two HW models to Overseas data: one where the seasonality is additive (i.e. added to the trend and noise), and one where the seasonality is multiplicative (i.e. multiplied with the trend and noise), see Figure 3.19. Visually both models appear to do well, although the additive model seems to not be able to fit well the peak and valley in the second year.

All the model estimates are: $\alpha = 10^{-4}$, $\beta = 10^{-4}$, $\gamma = 10^{-4}$ (see the R Output at the end of this chapter, in the Section R Lab 1).

3.4.3 Choosing the best-fitting model

We learned how to fit models to time series data: time series regression models and three types of exponential smoothing models. Next, we need to know how to select the best-fitting model among a set of candidate models (as we discussed in Section 3.1.2).

Criteria to find the best fitting model. To find the best-fitting model from a set of candidate models we need a model selection criterion. Actually, there are several model selection criteria and each criterion may recommend a different model. We now discuss such criteria.

SSE criterion. This criterion utilises the sum of squared residuals (SSE) statistic from Equation 3.36. From the model candidates, can we choose the model with the smallest SSE? We can use it to compare models with same

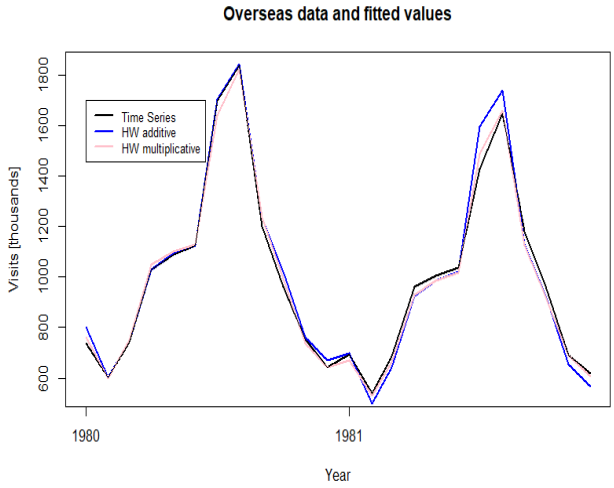


FIGURE 3.19: Overseas visits and two fitted HW smoothing models: additive and multiplicative models.

number of parameters. However, if the models have varying numbers of parameters then we should not use SSE, because SSE will always drop with the addition of another (even irrelevant) predictor. Note, that we already used SSE to compare several SES models to find the optimal *alpha* (see e.g. Figure 3.17), which was OK to do as all the models were having the same number of parameters (y_0 and α) thus the same complexity.

Then the *maximum likelihood estimator* for the variance σ_w^2 of the noise component of time series is

$$\hat{\sigma}_q^2 = \frac{SSE}{T} \tag{3.44}$$

where q is the number of parameters to be estimated (such as y_0 and α for the SES model, or β_0 and β_1 in regression coefficients). Such an estimator is biased as it underestimates the variance σ_w^2. From the model candidates, can we choose the model with the smallest $\hat{\sigma}_q^2$? No, because as we said this estimator is biased, so it is not used for a model selection.

RMSE criterion. An unbiased estimator for the standard deviation σ_w of random errors is

$$s_q = RMSE = \sqrt{\frac{SSE}{T-(q+1)}} \tag{3.45}$$

where $RMSE$ denotes the *mean squared error (RMSE)* statistic, where

- $q = 1$ for a linear regression model

- $q = 2$ for a quadratic regression model
- $q = 3$ for a cubic regression model
- $q = 2$ for a simple exponential smoothing (SES) model (the two parameters are α and y_0)
- $q = 3$ for Holt's model
- $q = 4 + s$ for Holt-Winter's (HW) model, where $s = 12$ for monthly data, $s = 4$ for quarterly data

In RMSE we penalise for the complexity of the model. How? As q gets larger, it makes RMSE larger. So if we add another parameter (thus increase q) which does not decrease SSE much, such a small SSE decrease may be overturned by the increase in q. Penalising for the complexity is desired, as it let us choose the most parsimonious model as well as the model that generalises for future data. When choosing the best-fitting model, we look for a model with the smallest RMSE.

AIC criterion. Akaike (see [5], [6], [7]) suggested the following statistic:

$$AIC = \log \hat{\sigma}_q^2 + \frac{T + 2q}{T} \tag{3.46}$$

where $\log \hat{\sigma}_q^2$ is given by Equation 3.44 and q is the number of parameters in the model. The acronym AIC denotes the *Akaike's Information Criterion.* AIC is the amount of discrepancy between data and the data-generating mechanism. So we want AIC to the be smallest possible. The value of q yielding the minimum AIC specifies the best model. The idea is roughly that minimizing $log\, \hat{\sigma}_q^2$ would be a reasonable objective, except that it decreases monotonically as q increases. Therefore, we ought to penalize the error variance by a term proportional to the number of parameters. So just like RMSE, AIC also helps to find the most parsimonious model as well as the model that generalises well for future data. The choice for the penalty term given by Eq 3.46 is not the only one, and considerable literature is available advocating different penalty terms. See e.g. [52], page 51 for more discussion.

Bayesian Information Criterion (BIC) Another option for a correction term is based on Bayesian arguments, as in Schwarz (1978), which leads to the following statistic:

$$BIC = \log \hat{\sigma}_q^2 + \frac{q\, log\, T}{T} \tag{3.47}$$

BIC is also called the Schwarz Information Criterion. The penalty term in BIC is larger than in AIC, consequently, BIC tends to choose smaller (i.e., simpler) models. Various simulation studies showed that BIC does well at getting the

correct model in large samples, whereas AICc tends to be superior to BIC in smaller samples where the relative number of parameters is large (see [40]) for detailed comparisons.

R-squared and Adjusted R-squared statistic. When comparing several candidates of regression models we can also use *R-squared* statistic

$$R^2 = 1 - \frac{SSE}{SST} \tag{3.48}$$

where $SST = \sum(y_t - \bar{y})^2$ is total sum of squares, and $\bar{y}$ is the grand average. When we compare models with the same number of parameters (hence same q), then the one with the smallest R^2 is the best fitting model. However, we cannot use R^2 to compare two models with different numbers of parameters. This is because R^2 always increases. If we add an irrelevant explanatory variable into the model, thus increasing its complexity, the R^2 will increase a little bit, thus making R^2 not a useful metric.

Instead of R^2 statistic, it is better to use its generalisation *adjusted R-squared*

$$R^2_{adj} = R^2 - (1 - R^2)\frac{q}{T - (1 + q)} \tag{3.49}$$

which is essentially having the same penalisation for complexity as s^2_w from 3.45.

Caution! When comparing a statistic across several model candidates, we must make sure they are calculated from the same dataset. So, as a consequence, they all have the same T.

Further model selection criteria are Mallows Cp (by Mallows [39]) and bias-corrected AIC criteria (AICc), which we do not consider in this book.

Example. Kings life span data. (continues) How to select the best model from the two models that we estimated earlier?

Solution. We can calculate several statistics and compare them: SSE, RMSE, R^2, R^2_{adj}, AIC and BIC. The `R` code is in Question 2, Section R Lab 2. All the statistics are summarised in Table 3.7. We note that the quadratic model does better in terms of RMSE, R-squared adjusted, AIC and BIC. We do not comment on the fact that the quadratic model has better SSE and R-squared, as this is expected: a model with more predictors will always have better SSE and R-squared. So if we have to choose between the two models, we choose the quadratic model. Does this mean that the quadratic model should be used for forecasting? We do not know the answer to that yet, as we need to check if the assumptions of the model are satisfied (such as if the errors are normal and identically distributed with zero mean and common variance) which is what we will do next.

Model	**q**	**SSE**	**RMSE**	R^2	R^2_{adj}	**AIC**	**BIC**
Linear 3.17	2	9423.694	15.34902	0.1628	0.1419	352.5500	357.7630
Quadratic 3.19	3	7986.085	14.30984	0.2905	0.2542	347.5979	354.5485

TABLE 3.7: Summary statistics for linear and quadratic models of Kings life span data.

Example. Overseas visits. (continues) How to select the best model from the two models that we estimated earlier for the Overseas data, using the first two years?

Solution. We can calculate several statistics and compare them: RMSE, AIC and BIC. The `R` code is in Question 1 Section 1. All the statistics are summarised in Table 3.8. We note that the multiplicative model does better in terms of RMSE, AIC and BIC. Does this mean that the multiplicative HW model should be used for forecasting? We do not know the answer to that yet, as we need to check if the assumptions of the model are satisfied (such as if the errors are normal and identically distributed with zero mean and common variance) which is what we will do next.

Model	**q**	**RMSE**	**AIC**	**BIC**
Additive	16	49.61922	297.6834	317.7104
Multiplicative	16	28.21509	264.1459	284.1728

TABLE 3.8: Summary statistics for additive and multiplicative HW model of Overseas visits data (January 1980 - December 1981).

We now know how to find the best-fitting model or models. This is however not enough. Next, such model(s) must be subjected to further checks, also called *goodness-of-fit checks* of the model, which we do in the next section. If the model fails the goodness-of-fit check we cannot trust the forecasts from such a model.

3.4.4 Think ZINC!

ZINC assumptions. When using a model for forecasting we also need to provide prediction intervals. If we construct a prediction interval using the formulae (Equation 3.52), then we need to be aware of the fact that this formula was built using four assumptions, which we will call ZINC assumptions:

$$\begin{aligned} &(Z)\, zero\, mean\, of\, errors\, for\, each\, time\, point\, t \\ &(I)\, independence\, of\, errors \\ &(N)\, normal\, distribution\, of\, errors \\ &(C)\, common\, variance\, of\, errors \end{aligned} \tag{3.50}$$

where by errors we mean the errors of the fitted model, also called residuals, w_t, defined in Eq. 3.18,3.34.

Hence, in order to use the formulae 3.52 we need to check all four assumptions. If at least one of the assumptions is not satisfied, then we cannot trust the prediction intervals from that formula (Equation 3.52) and from that best-fitting model. If all assumptions are satisfied, then we can trust the prediction intervals.

How do you decide if the best-fitting model is also a well-fitting model? We need to do a goodness-of-fit analysis which is we need to check if ZINC assumptions (Eq. 3.50) are satisfied. This is a goodness-of-fit analysis that we need to do in order to decide if a model is a well-fitting model:

- (Z) **Zero mean of errors for each t.** To check if residuals have zero mean, it is important to make a plot of residuals vs. time. If such a plot shows no patterns, then we say that we have no evidence against the assumption of zero means of errors. A model with patterns in errors should not be used for forecast and is hence inferior to a model with no pattern in error plot. Zero means of errors for each t is representing a model whose mean behaviour is correct, e.g. if data consists of a linear trend and a noise, and if such data are modelled with a linear regression model, then the mean of the residuals will be zero at each t.

- (I) **Independence of errors.** To check if residuals are not correlated, we can use the autocorrelation function (from Section 3.3.3). If errors are uncorrelated and normally distributed, then they are independent.

- (N) **Normality of errors.** To check the normality of errors (residuals), we can use a normal-probability plot of the residuals (e.g. Kolmogoro-Smirnov test or Shapiro-Wilk test and the p-value of the tests).

- (C) **Common (or constant) variance of errors.** To check if residuals have a common variance, it is important to make a residuals vs. time plot. If the plot shows constant variance for all times, then we say that we have no evidence against the assumption of the common variance of residuals.

Caution! Remember, making a forecast is easy, but making a good forecast can be difficult! The difficulty is in crafting a set of candidate models, then choosing the model that is the best fitting model that later shows to be a well-fitting model, too. This has also been discussed earlier in Section The forecasting steps 3.1.2.

Example. Kings' life span. (continues.) We found the quadratic model to fit data better than the linear model, according to several model selection criteria (Table 3.7). Next, we need to check if the quadratic model is a well-fitting model for the Kings' life span data, i.e. to do a goodness-of-fit analysis of the quadratic model.

Solution. We do the goodness-of-fit checks in R. The R code is in Section R Lab (see Question 2).

First, we check if the residuals of the quadratic model have zero mean (Assumption Z). We use the plot on the left in Figure 3.20. By visual inspection, we see no patterns in the residual plot, in other words, we see that for each value of time, the mean of residuals is roughly zero.

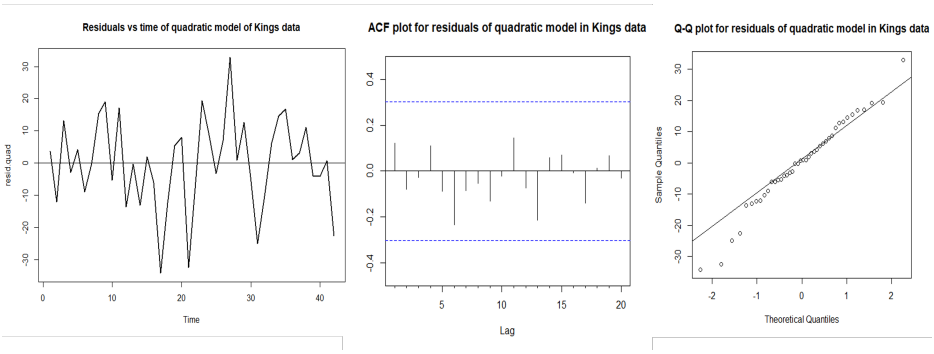


FIGURE 3.20: Goodness of fit analysis of the quadratic model for Kings life span data.

Next, we check if the residuals of the quadratic model are independent of each other (Assumption I). We use the plot in the middle in Figure 3.20. We see that all ACF values are within the 95% confidence bands. Since at least 95% of calculated ACF values are within the 95% confidence bands, we do not have evidence against independence. So we conclude independence of the residuals.

Next, we check visually if residuals of the quadratic model are normally distributed (Assumption N). We use the plot on the right in Figure 3.20. By visual inspection, we see some deviations from the 45-degree line, but it is unclear if they are bigger beyond chance and hence significant. So we did the Shapiro-Wilk normality test on residuals from the quadratic model and we got a test statistic W = 0.97771 with a p-value = 0.5744. In this test, the hypothesis H0 is that residuals do have a normal distribution. Since 0.5744 is bigger than 0.05, we conclude that we do not have evidence against H0, at 0.05 level of significance. So we conclude that the residuals are normally distributed.

So we conclude that all assumptions of residuals are satisfied.

Example. Overseas visit data. (continues.) Next, we will check the goodness-of-fit of the multiplicative HW model fitted to the first 2 years of Overseas visits data.

Solution. We do the goodness-of-fit checks in R. The R code is in Section R Lab (see Question 1).

First, we check if the residuals of the HW model have zero mean (Assumption Z). We use the plot on the left in Figure 3.21. By visual inspection, we see no pattern in the residual plot. In other words, we see that for each value of time, the mean of residuals is zero. So assumption Z is violated.

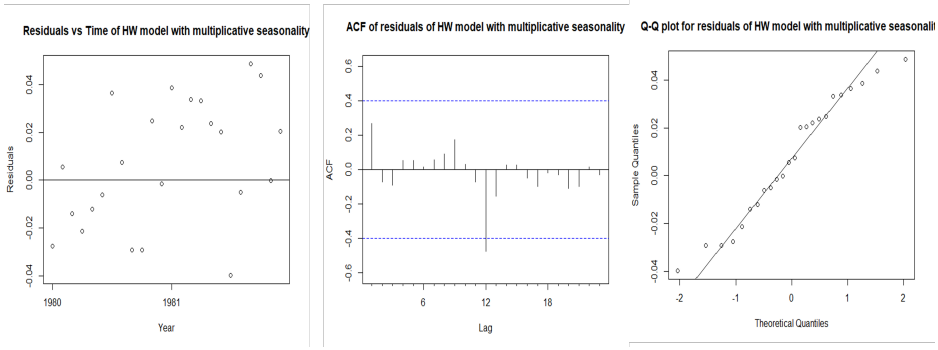


FIGURE 3.21: Goodness of fit analysis of the multiplicative HW model of Overseas monthly visits 2 years data January 1980 - December 1981.

Next, we check if the residuals of the multiplicative HW model are independent of each other (Assumption I). We use the plot in the middle in Figure 3.21. We see that all ACF values except one are within the 95% confidence bands. Since at least 95% of calculated ACF values are within the 95% confidence bands, we do not have evidence against independence. So we conclude independence of the residuals.

Next, we check if residuals of the multiplicative HW model are normally distributed (Assumption N). We use the plot on the right in Figure 3.21. By visual inspection, we see some small deviations from the 45-degree line, but it is unclear if they are too big and hence significant. So we did the Shapiro-Wilk normality test of the residuals giving p-value = 0.3397. In this test, the hypothesis H0 is that residuals do have a normal distribution. Since 0.3397 is more than 0.05, we conclude that there is not enough evidence against H0, at a 0.05 level of significance. So we conclude that the residuals are normally distributed.

Last we check if the residuals of the HW model have a common variance (Assumption C). We use the plot on the left in Figure 3.21. By visual inspection, we that the spread of the residuals is constant, for all values of time the variance is the same or similar.

In summary, the multiplicative HW model that we with to the first two years of Overseas visits data satisfies all four ZINC assumptions.

Overseas visits. (continues.) Next, we will check the goodness-of-fit of the additive HW model fitted to the first 2 years of Overseas visits data. The information criteria found the additive HW model worse than the multiplica-

tive model. However, as an education exercise, we also do a goodness-of-fit check of the additive HW model of Overseas visits data, of the first two years.

Solution. The graphs are in Figure 3.22.

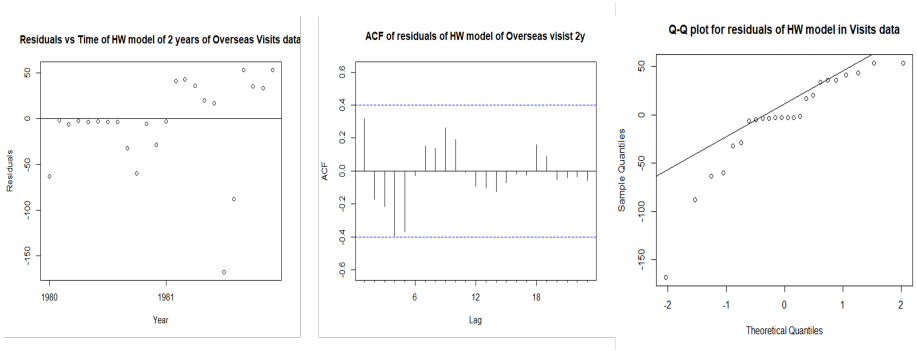


FIGURE 3.22: Goodness of fit analysis of the additive HW model of Overseas visits 2 years data.

Next, we check if the residuals of the HW model are independent of each other (Assumption I). We use the plot in the middle in Figure 3.22. We see that all ACF values are within the 95% confidence bands. Since at least 95% of calculated ACF values are within the 95% confidence bands, we do not have evidence against independence. So we conclude independence of the residuals.

Next, we check if residuals of the HW model are normally distributed (Assumption N). We use the plot on the right in Figure 3.22. By visual inspection, we see some deviations from the 45-degree line, but it is unclear if they are too big and hence significant. So we did the Shapiro-Wilk normality test of the residuals gives a test statistic W = 0.85463 with a p-value = 0.002667. In this test, the hypothesis H0 is that residuals do have a normal distribution. Since 0.002667 is less than 0.01, we conclude that we do have strong evidence against H0, at 0.01 level of significance. So we conclude that the residuals are not normally distributed, i.e. the normality assumption is violated.

Last we check if the residuals of the HW model have a common variance (Assumption C). We use the plot on the left in Figure 3.22. By visual inspection, we that the spread of the residuals is not constant, for low values of time the variance is small, and then the variance increases.

In summary, three assumptions are not satisfied. We should not be using the additive HW model for forecasting.

3.4.5 Prediction intervals

We had several candidate models that we fit to time series data $Y_1, \cdots, Y_T$, we discussed how to find out which model fits the data the best, and then we discussed how to check if such a best-fitting model is also a well-fitting

model. Next, we want to use the best-fitting and well-fitting model to make the forecasts, i.e. to use the time series to estimate the value y_{T+h}, where $h = 1, 2, \ldots$ is the forecast horizon. There are several types of forecasts that are useful for stakeholders: point forecasts, three-point forecasts, and prediction interval forecasts (see also Section 3.1.3).

Point prediction (or point forecast) is a point in a graph, e.g. the number of visitors to the UK in January 2023 being predicted as 1.5 million. Point forecasts can be obtained from the fitted model as we mentioned in Sections 3.4.2 and 3.4.1. We can calculate point forecasts for January till December 2023, i.e. 12 forecasts, and we can join them on a plot. Such a plot is then called *forecasted central path*.

Point forecast is not sufficient for decision making. If we go for a holiday to Portugal, knowing that the most likely temperature during the day is 25 degrees Celsius is not enough for us to make a decision about what clothes to pack. Knowing that 80% likely the temperatures will be in the range of 18 and 33 degrees Celsius is more helpful for us, we will know that we need to pack some light and some warmer clothes too. Then we can comfortably travel as long as we are ok to take the risk that there is a 10% chance that the temperature will be lower than 18, and a 10% chance that the temperature will be above 33 degrees Celsius.

In some application domains, the experts will insist that a *three-point estimate* (forecast) is provided. So in the case of the temperatures, a three-point estimate can be 18, 25, or 33 degrees of Celsia. Such estimates are commonly used by decision-makers in critical application scenarios such as in the military. A three-point estimate can be defined as the first point being a 10% percentile, the second point being the 50% percentile and the third point being the 90% percentile - as we illustrated in the previous paragraph. A three-point estimate is easy to communicate, it is better than a one-point estimate, yet it is still limited in terms of provided information for decision-making (see also [22]).

Sometimes truly rational decision-making requires knowing the full probability distribution of potential future values, i.e. providing all percentiles which is called a *probabilistic forecast*. To communicate such a probabilistic forecast requires some thought. One way to do it is to calculate and plot a set of relevant prediction intervals. For example in Figure 3.24 we have 50 and 95% prediction intervals, in Figure 3.4 we show 80 and 95% prediction intervals, and in Figure 3.5 we show the full range of intervals: 10,20,...90% prediction intervals.

Prediction interval what needs to be communicated to stakeholders for requested forecast horizons, h. For each h, a prediction interval is a set of two numbers: one lower bound and one upper bound. For example, the number of visitors to the UK in March 2023 is predicted as 1.1 to 1.7 million. In other words, we want to add a bound on our point forecast, i.e. a prediction interval. A wide interval will mean that we are not certain about

the forecasted value, and a narrow interval will mean that we are more certain. Intuitively we want:

- The width of the prediction interval to be larger with larger h as it will reflect that predicting a more distant future is more uncertain than a less distant future,
- The width of the prediction interval will be large if the past behaviour of the time series had a large noise variance, so naturally we always look into the errors (also called remainders or residuals) to see how to figure out the width of the prediction interval. We do like it when the residuals are independent and normally distributed with zero means, and common constant variance because then the construction of the prediction interval is easy to do (which we discuss in the next paragraphs).

What is a prediction interval (PI)? It is an interval that will contain the true value of the future observation (at horizon h) with a pre-specified required probability. Formally, a $(1-\alpha)\times 100\%$ prediction interval, for a future observation Y_{T+h}, is an interval PI= $(Lower\,Bound, Upper\,Bound)$ such that

$$P(Lower\,Bound < Y_{T+h} < Upper\,Bound \,|\, y_1, \ldots, y_T) = 1 - \alpha \qquad (3.51)$$

i.e. with 95% probability the Y_{T+h} will be contained in this interval, where $\alpha = 0.05$.

The idea of prediction interval is similar to confidence interval (CI), but its calculation and interpretation are different. The prediction interval tells us where one future value can be at time point $T + h$. The confidence interval tells us where the mean of future values can be at time point $T + h$. Intuitively, the PI is wider than CI, because predicting one realisation is harder than predicting the mean of several realisations. In other words, we are less sure where one observation will be than where the mean of several observations will be. Sometimes stakeholders need PI, sometimes CI, as we discussed in Section 3.1.3. And as we will see in the R-Lab section at the end of this chapter, the function `predict` provides both CI and PI.

How to construct a prediction interval? Assume we have time series data and we found the best fitting model among a set of candidate models. If such a model also passed the goodness-of-fit check criteria (see ZINC assumptions), then we can use a so-called parametric approach to PI construction:

$$\left(\hat{y}_{T+h} - z_{1-\alpha/2}\,\hat{\sigma}_h, \hat{y}_{T+h} + z_{1-\alpha/2}\,\hat{\sigma}_h\right) \qquad (3.52)$$

where $\hat{\sigma}_h$ is an estimate of the standard deviation of the h-step forecast distribution, $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of normal distribution. So $\alpha = 0.05$ gives $z = 1.96$ and hence a 95% prediction interval. This interval has several properties:

- PI gets wider with more noise in data, as seen by having $\hat{\sigma}_h$ on both sides of the PI.
- For each value h, PI is wider for larger $1 - \alpha/2$. In other words, a 99% PI will be wider than 95% PI, for each h.
- The more far away future we want to predict (i.e. h is large), the wider PI will be. This is achieved via$\hat{\sigma}_h$ increasing with h.
- This interval is valid for prediction from time series regression models as well as from exponential smoothing models, where there is just one difference: how $\hat{\sigma}_h$ is calculated. The calculation of $\hat{\sigma}_h$ can be complex, it depends on the model used. Generally, if we only do a 1-step ahead forecast (i.e. h=1), then $\hat{\sigma_h}$ is very close to our estimate of the standard deviation of errors: $\hat{\sigma_q} = RMSE$. To see how $\hat{\sigma}_h$ is calculated for exponential smoothing models please read [34]).
- The PI is called a parametric interval because it uses the fact that the residuals are having normal distribution, which can be described by two parameters: the mean and variance. It assumes that the mean of residuals is zero at any time point in data and that the variance is constant at any time point in data. It also uses the fact that the residuals are independent of each other.
- Most books use t-quantiles instead of z-quantiles, thus making it more accurate.

For more discussion on PI, do read Section 3.5.

Example. Kings' life span. (continues) We have found that the quadratic model fits the data better. We are asked the following questions:

- Question 1. Construct a forecast for the next three kings, provide point estimate as well as 80% and 95% prediction interval. Plot the data, the quadratic model and the prediction intervals – all on the same plot. Interpret the forecast intervals.
- Question 2. Is the next king likely to live at least 70 years? Or is there a chance of the next king living less than 70 years?
- Question 3. Are all three next kings likely to live at least 70 years?
- Question 4. Are we satisfied with the model? Do we trust the forecast? Is there another model we should try?

Solution. We can use `R` to do all the calculations, see Section R Lab 2.

The answer to Question 1: For kings 43, 44 and 45 the point estimate of their longevity is: 81.1, 83.6 and 86.2 years, respectively.

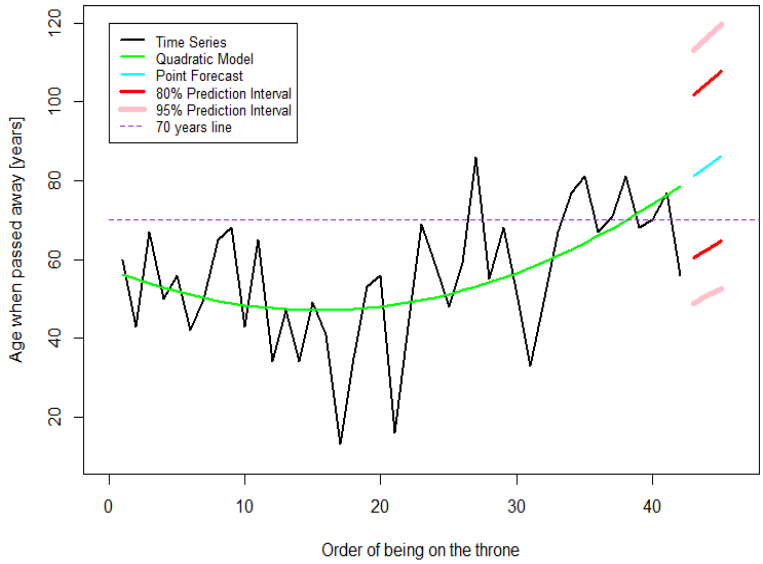


FIGURE 3.23: Kings life span data and the forecasts for the next three Kings.

We are 80% confident, that the kings 43, 44 and 45 will live till at ages: (60.3, 101.8), (62.4, 104.7) and (64.6, 107.8) years, respectively.

We are 95% confident, that the kings 43, 44 and 45 will live till at ages: (48.9, 113.2), (50.8, 116.4) and (52.6, 119.7) years, respectively.

We note that the 95% confidence gives a wider prediction interval, as expected. We also note that the prediction intervals are very wide, which means that there is a lot of uncertainty about the longevity of the kings.

Regarding Question 2: Is the next king likely to live at least 70 years? The next king is king number 43. We predict for him to most likely live till 81.1 years, and with 80% or 95% confidence we predict he will live (60.3, 101.8) or (48.9, 113.2) years, respectively. The number 70 is in both prediction intervals, so it is supported by the data.

Is there a chance that the next king will live less than 70 years? To put this into mathematical language, we are asked to find P(43rd King lives $\geq$ 70 years). We do know that

$$P(43rd\,King\,lives\,\geq 81.1\,years) = 0.5$$

i.e. the 43rd King has a 50% chance to live 81.1 or more years. So, since 70

is less than 81.1, the 43rd King has more than 50% chance to live at least 70 years, according to the quadratic model.

Regarding Question 3: Are all three next kings likely to live at least 70 years? We are asked to find

$$P(43rd\,King\,lives\,\geq 70\,years,\,and\,44th \geq 70\,years,\,and\,45th\,\geq 70\,years)$$

We know that the Kings are independent (since Acf was uncorrelated), so since each has at least 50% chance to leave at least 70, then there is at least $0.5^3 = 0.125 = 12.5\%$ chance that all three live at least 70 years, according to the quadratic model.

Regarding Question 4: Are we satisfied with the model? Do we trust the forecast? Are there other models we should try? A reasonable answer: In overall, it seems that the quadratic model is a good choice, as all ZINC assumptions are satisfied. Obviously, in checking the Z assumption we used a visual judgement, so our conclusion was subjective. There is a potential to try a 3rd-order polynomial for the predictions, as the quadratic model seems to be increasing too fast for the last kings, and hence it may be too optimistic for the future kings. A polynomial of higher order (e.g. third order) can be more suitable to capture the fact that the longevity stabilized for the last 10 kings. This can be checked quantitatively using the model selection criteria.

Example. Overseas visits. (continues) We have fit a Holt-Winters model to the Overseas Visits time series data from Jan 1980 – Dec 1981. Next, we use the time series data to predict the overseas visits in Jan 1982. All output is in Section R Lab 1.

Using the HW model, a point prediction for January 1982 is $\hat{y}_{T+1} = 693.0357$ thousand visitors. The 95% prediction interval is (581.5528, 696.5187). The 80% prediction interval is (601.4497, 676.6218) thousands of visitors. The interpretation is as follows: with 95% confidence, the number of visitors in January 1982 is predicted to be between 581.5528 thousand and 696.5187 thousand. For this we used data from January 1980 to December 1981, we assumed that the random errors are independent, normally distributed with zero mean and constant variance and we used the past data (i.e. we did not use any other information).

Figure 3.24 shows prediction intervals for the next three years (hence 36 prediction intervals). Again: for these forecasts, we used data from January 1980 to December 1981, we checked ZINC assumptions that they are all satisfied, and we used data only to do these forecasts.

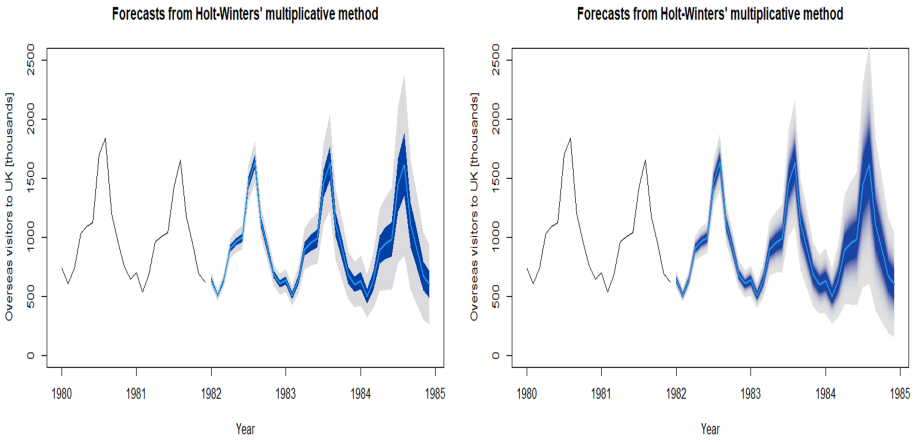


FIGURE 3.24: Overseas visits to UK analysed via the multiplicative HW model. LEFT: 50% and 95% prediction intervals (see the dark blue and light grey shaded areas, respectively). RIGHT: Fan chart, from 51 to 99% prediction intervals.

3.5 Tips to think and act like a risk expert

Here we will give tips and tricks on prediction intervals, on how to communicate the risk when time series have been used to estimate the risks. In doing so we will tie up what we said in Section 3.1.1 with the later sections.

3.5.1 Remember there is no such a thing as a free lunch!

The formula Eq. 3.52 for prediction interval looks rather notorious. We have seen it so many times earlier. It is beautifully simple in that it is symmetric around the point forecast $\hat{y}_{T+h}$ and it uses the quantiles of normal distribution. But there is no such thing as a free lunch! The simplicity of this PI comes with a catch. The catch is that the model must satisfy the four ZINC assumptions as well as the fifth assumption that we can use past data to forecast the future. The ZINC assumptions are: the residuals (Z) have zero mean, (I) are independent of each other, (N) are normally distributed, and (C) have constant variance, i.e. we assumed that the forecasting model is a well-fitting model. What we do in forecasting is predict the time series values beyond the range of time 1,... T (more discussion in Section 3.5.3). If these five assumptions are true, then it is safe to use the formula 3.52. To see the intuition behind ZINC, let us look at the assumption of normality of the residuals. It means that in the past the random variation of data around the

model was following a normal distribution. We assume in future the random variation around $\hat{y}_{T+h}$ will be the same, we assume it will again be normal, thus it is OK to use the quantile of normal distribution in our formula 3.52. In summary, this prediction interval 3.52 will only be valid if the residuals do satisfy these five assumptions.

What if the best fitting model did not pass ZINC assumptions? Assume we found a best-fitting model among the set of candidate models. However such a model did not pass the goodness-of-fit check (i.e. the ZINC assumptions are not satisfied), then there are some options. Usually, the following steps help:

1. It helps to start with Assumption Z, as it means that the fitted values from the model are not really going through the data well. If assumption Z is violated (irrespective of the other three assumptions), then we should try to see if there is a better model that we may not think of earlier. Violation of Z happens if the residuals (errors) vs time plot shows a pattern. If residuals show a quadratic pattern, then a remedy would be to add a quadratic predictor ($time^2$) into the model thus having a new model candidate. Violation of Z also happens if we have forgotten to add an important predictor into our model. Adding such a predictor may do the trick of making our model pass the Z Assumption. (Often this trick may make other assumptions to be passed too.)

2. If Assumption Z is satisfied but Assumption C is violated (irrespective of the other two assumptions), then the residuals have a variance that increases (or decreases) with the overall trend. Then we should try to fit a multiplicative model if using exponential smoothing, or we should try to log transformation of Y if using time series regression models, and then see if such a new model passes the C assumptions. (This remedy may make other assumptions to be passed too.)

3. If Assumptions Z and C are satisfied, but Assumption N of normality is not satisfied (irrespective of the assumption I) while the histogram of residuals is unimodal but skewed, then we should try a model where Y is in logarithmic scale. Sometimes this helps to get residuals that satisfy the normality. But sometimes it does not help.

4. Lastly, if the Assumptions Z and C are satisfied, but normality cannot be achieved or residuals are still correlated, then realise we did what we could and we conclude we have the best possible model, even if it does not satisfy all ZINC assumptions. We can use the model to get the point forecasts, but for the construction of PI, we must not use the Eq. 3.52. For the PI construction, we need to use a non-parametric approach, a kind of computer simulation method (such as Bootstrap) where we simulate future sample paths by using the past residuals, however, it is beyond the

goal of this book. For bootstrap methods of obtaining the prediction intervals, we recommend reading: [56] for forecasting from linear and nonlinear regression models, [44] for auto-regression time series forecasting, and [64] for Holt-Winters forecasting.

Why don't we just go with a non-parametric approach always? Why do we fuss in trying to find a model that satisfies all ZINC assumptions? Because the non-parametric approach will not work well if Z or C are violated, and because the non-parametric approach always gives wider PI, than the parametric approach. (Though there is no elegant theory suggesting how much wider, the research suggests that the answer depends on the data at hand.)

3.5.2 Be a pro at visualising the risk and uncertainty

As we said earlier, the best way to visualise the risks is via plots that include both the central forecast path, as well as the prediction intervals. We showed examples: in Figure 3.24 we have 50 and 95% prediction intervals, in Figure 3.4 we show 80 and 95% prediction intervals, and Figure 3.24 we show the full range of intervals: 51 to 99% prediction intervals. Note, that the last figure is called a "fan chart", so it can be seen as a special case of prediction interval plots. Note, that Hyndman and Athanasopoulos ([34]) call the plots with prediction intervals as "Visualization of probabilistic forecasts".

Example. Overseas visits. We used Overseas visits data and generated one-step-ahead prediction and prediction interval, as well as predictions with horizon up to $h = 36$ months, in Figure 3.24. What risks, opportunities and uncertainty we face after we did forecast from the time series? Several thoughts follow:

- Main uncertainty is that we do know if the forces that drove the time series in the past will remain the same in the future. What if government issues a regulation that will restrict the travel to UK and hence affect the number of overseas visits? We may not even know that such regulation can happen. We may not be able to estimate the probability of such a regulation. We may not be able to estimate the effect of such regulation on the total number of overseas visits. This is a situation of *uncertainty about the model*. There is always uncertainty about the model choice (e.g. linear regression or Holt-Winters or other) and its parameters being estimated. But if we have a sufficient number of data and if all the assumptions are satisfied (normality of errors etc) then we can say we reduced the uncertainty about the model.
- Let us assume that the future will be like the past (i.e. the forces that drove the time series will remain unchanged) and let us assume we have a good fitting model. Then we still face some degree of uncertainty, because there are random fluctuations that we cannot forecast. We can however

estimate the probability distribution of such random fluctuations, and then we can use it to construct the prediction intervals. This is *uncertainty about outcome due random fluctuations*.

- We found that for January 1982 the expected number of visitors is 693.0357 thousand, and the 95% prediction interval is (581.5528, 696.5187) thousand. This is expressing all possible values of yield that can happen in future.

- We are facing some risks. For example, there is a 50% chance that the number of visitors will be lower than 693.0357 thousand, and a 50% chance that it will be above 693.0357 thousand, in January 1982.

- We are facing a risk that in January 1982 the number of visitors will be 581.5528 thousand or lower, with a probability of 2.5%. We are facing a risk (or opportunity, if we are a travel agency) that the number of visitors can be 693.0357 thousand or higher (as suggested by the 95% prediction interval).

3.5.3 When relying on a model alone is a wrong idea

In previous sections, we stressed that each statistical method (or model) comes with assumptions. We listed several assumptions that we make when we do forecasting from time series regression models or exponential smoothing models, and we discussed how to check those assumptions (see the ZINC assumptions in Section 3.4.5).

There is one additional assumption we make when forecasting from time series: we assume that the "future will be like the past". And then, we use the model to predict the values in the future, i.e. beyond the range of observed time points $1, \ldots, T$. Is that reasonable to do? There are two scenarios when this is not reasonable:

- **What if the external forces that drove the values of our time series will change in future?** In the Example Daily temperatures, we aimed to predict temperatures using the past temperatures. But what if all governments will reinforce drastic measures to reduce the pollution on Earth? This will affect the time series in future, e.g. it may cause the temperatures to stop increasing, thus causing a so-called *change point* in time series. If we only use the past temperature data, we have no way to predict future temperatures. We need to be cautious. The "future will be like the past" will not always be true. In such a situation, we need to find ways to incorporate expert judgement into our forecasts.

- **What if the external forces that drove the time series do not change while there will be an internal cause for change in the time series?** What if we are predicting the number of visitors to the UK

using past visitors' data? What if the UK government is not planning any change in travel policies? And what if the UK travel market is almost saturated in the sense that it cannot handle any more visitors, i.e. is saturated? And what if the past data do not show any sign of saturation? Then surely, any forecast that is purely based on the past data will be an overestimation of future visits. In such a situation, we need to find ways to incorporate expert judgement into our forecasts. Bayesian time series forecasting has ways of incorporating expert judgment into the forecast but is not in the scope of this book, though we showed one forecast example, which was purely based on expert judgements in Figure 3.5.

Caution! When we communicate the data-based forecast to stakeholders, it is our duty to remind them that our forecasts are only true when the ZINC assumptions are satisfied as well as when the "future will be like the past" assumption is satisfied.

3.6 Summary

We learned in this chapter:

1. Time series analysis involves analysing data points that are collected at regular intervals over time to identify patterns, trends, and other relationships that can be used to forecast future values. Time series methods are often used in quantitative risk analysis to evaluate the probability and size of risks over time.

2. We learned the basic tools: moving averages (MA), auto-correlation (ACF), simple exponential smoothing (SES), Holt's exponential smoothing (also called double exponential smoothing), Holt-Winters (HW) exponential smoothing (also called triple exponential smoothing). We learned how to decide which model is the most suited for the data at hand. We learned how to use time series data to make predictions, how to quantify and communicate the prediction uncertainty and plot it all into a plot called a fan chart. We learned how one can make recommendations from time series about potential future events.

3. Overall, time series methods are a powerful tool for quantitative risk analysis because they allow analysts to identify patterns and trends that may not be immediately apparent in raw data, as well as they also allow forecasting - which can help organisations make better-informed decisions about risk management.

3.7 Further reading

In risk analysis, time series methods can be used to analyse historical data on various *risk factors*, such as market trends, economic indicators, or weather patterns. This information can be used to (A) develop predictive models that estimate the likelihood of future events or (B) identify risk factors, which can be used to inform risk management strategies. Time series have a vast number of resources.

In our chapter, we used several resources, which we list here:

1. For further understanding of applied time series and R, we recommend this online book on time series: Rob J Hyndman and George Athanasopoulos's book "Forecasting: Principles and Practice", available at https://otexts.com/fpp2/ The book presents the statistical theory on basic to intermediate level.

2. For further understanding of applied time series and more examples in R, we recommend the book "Time Series Analysis and Its Applications. With R Examples. Fourth Edition." by Shumway and Stoffer [52]. The book presents a statistical theory on an intermediate to advanced level. It has Chapter 2 on Time Series Regression and Exploratory Data Analysis, where they introduce multiple linear regression and then give attention to more topics: exploratory data analysis for preprocessing of nonstationary time series (for example, trend removal), the concept of differencing and the backshift operator, variance stabilisation, as well as nonparametric smoothing of time series.

3. For a quick way to start with R and then run time series analysis in R, we recommend this online book by Avril Coghlan [15].

4. We focused on the frequentist approach of using time series analysis for risk and uncertainty estimation. There are Bayesian extensions which are also popular and more flexible than the frequentists' methods. They are, however, more advanced and outside the scope of this chapter. For an interested reader, we recommend [52].

5. To learn about how to identify risk factors, we recommend the book by Hanck, Arnold, Gerber and Schmelzer [29]. It is accessible at https://www.econometrics-with-r.org/14-ittsraf.html, and it shows the R code as well. It has Chapter 14 devoted to time series, as well as further time series topics such as estimation of dynamic causal effects, vector autoregression, and cointegration.

6. To get a wider knowledge of the visualisation of time series, we recommend Shurkhovetskyy et al. (2017) "Data Abstraction for Visualizing Large Time Series" [53].

3.8 R Lab

Here, we practice using R to analyse time series. The first three questions are about the three datasets we discussed in the previous sections. Thus the first three questions are provided with a solution R code. Here we explain the R code and show the output from R. The interpretation of the R output was explained in the previous sections.

Then after the first 3 questions, you are asked to solve more questions, for which a solution is available upon request.

1. [**Purpose: Analysis of Overseas visits data.**] Here we analyse the Overseas visits data. The whole analysis is split into several sections. Each section is defined by its own goal.

   First, we import the dataset Overseas visits to UK data, that we used in this chapter. The data file is a CSV format: Risk-2021-DataVisitsToUK.csv. We recreate Figure 3.1 via using the `autoplot` function.

```
# R Code
# ------------------------------------------------------------
# Goal: To import Overseas visits data and plot its first 2 years' data
# ------------------------------------------------------------
#
# Import data into R and call it mydata
mydata<-read.csv("Risk-2021-DataVisitsToUK.csv")
# Check the structure of the imported object
str(mydata)
# Create a ts structure called mydata2
mydata2 <- ts(mydata$Visits,start=c(1980,1),end=c(1981,12), frequency=12)
# Check the structure of the new object mydata2. R should say that it is
# a Time-Series structure.
str(mydata2)
# Next we attach the library for plotting (there are other libraries)
library(ggfortify)
# Next we plot the data
autoplot(mydata2) +
   ggtitle("Monthly Visits ") +
   xlab("Month") +
   ylab("Visits[Thousands]")

R Output
> str(mydata)
'data.frame': 483 obs. of  4 variables:
 $ X     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Month : chr  "JAN" "FEB" "MAR" "APR" ...
 $ Year  : int  1980 1980 1980 1980 1980 1980 1980 1980 1980 1980 ...
```

```
 $ Visits: int  739 602 740 1028 1088 1124 1699 1839 1200 963
    ...
> mydata[1:3,]
  X Month Year Visits
1 1   JAN 1980    739
2 2   FEB 1980    602
3 3   MAR 1980    740

> str(mydata2)
 Time-Series [1:24] from 1980 to 1982: 739 602 740 1028 1088
    1124 1699 1839 1200 963 ...

```

Next, we plot all 40 years of monthly overseas visits to the UK data from Jan 1980 till Mar 2020, i.e. we recreate the Figure 3.2.

```
# R Code
# ------------------------------------------------------------
# Goal: To plot all 40 years of the Overseas visit time series
    .
# ------------------------------------------------------------
#
mydata<-read.csv("Risk-2021-DataVisitsToUK.csv")
autoplot(mydataALL) +
   ggtitle("Overseas Visits to UK") +
   xlab("Month") +
   ylab("Visits[Thousands]
```

Next, we practice moving averages to estimate the trend in Overseas visits data, in R. We thus recreate the Figure 3.7 via using the `ma` function.

```
# R Code
# ------------------------------------------------------------
# Goal: Detect the trend in Overseas visits data using the "ma
    " function.
# ------------------------------------------------------------
#
# Need to attach the library forecast, as it contains the
    function "ma"
library(forecast)
# Import the data
mydata<-read.csv("Risk-2021-DataVisitsToUK.csv")
# We try the moving average method of 12 and plotting the
    estimated trend
Visits_trend = ma(Visits_Series, order = 12, centre = T)
plot(Time2, Visits_Series,ylab="Visits [thousands]",
     xlab="Time [Year]",type="l",ylim=c(500,4500))
lines(Time2, Visits_trend,col="red",lwd=2)
title("Moving average of 12 observations")
legend(dt1,4500,legend=c("Time series", "Estimated trend"),
       col=c("black","red"), lty=1, lwd=c(1,2),cex=0.8)
```

Next, we are practising time series decomposition of Overseas visits in R. We recreate the Figure 3.8 via using the `decompose` function.

```
# R Code
```

```
# ------------------------------------------------------------
# Goal: Time series decomposition of Overseas visits data
# by using the function "decompose".
# ------------------------------------------------------------
#
# Need to attach the library forecast
library(forecast)
# Import the data
mydata<-read.csv("Risk-2021-DataVisitsToUK.csv")
str(mydata) # shows the structure of the data, for a quick
    check
head(mydata,10) # shows first 10 rows of data, for a quick
    check
# Next, we create Visits_Series_ts object that is in a "ts"
    format,
# because the decompose function only accepts data in a ts
    format.
# Make sure this structure knows that data are collected 12
    times
# per year, it starts in January 1980.
Visits_Series_ts = ts(Visits_Series, frequency = 12, start=c
    (1980,1))

# Now the decomposition of the time series in one single line:
decompose_visits = decompose(Visits_Series_ts, "additive")

# The plots
plot(as.ts(decompose_visits$x))
plot(as.ts(decompose_visits$trend))
plot(as.ts(decompose_visits$seasonal))
plot(as.ts(decompose_visits$random))

# Or simply:
plot(decompose_visits)
```

Next, we are practising forecasting from Overseas visits in R, via a computer simulation of potential future paths and by using the ETS model. Thus, here, we will recreate Figure 3.3 that shows four possible future paths for overseas visits to the UK. First, we fit an ETS model to the Jan 1980-Dec 1981 data (see row 3, below). Then we use the fitted ETS model to forecast four possible future scenarios for Jan 1981-Dec 1992. We use function `simulate` to simulate future scenarios. For each scenario, we set the random seed generator to a fixed value (see `set.seed` in rows 6, 9, 12 and 15, below) so each time we run this code we always get the same 4 simulated futures scenarios. Then we can plot the computer-generated potential future time series (see lines 4, 7, 10, 13, 16 and 17).

```
# R Code
# ------------------------------------------------------------
# Goal: To obtain obtain 4 potential future scenarios, in
    Overseas visits data.
# ------------------------------------------------------------
#
library(forecast)
```

```
# We fit ETS model to the time series data.
fit <- ets(mydata2)
plot(mydata2, xlim=c(1980,1992.1),ylim=c(400, 2400),ylab="
    Visits[Thousands]",xlab="Year")
# simulation 1
set.seed(10)
lines(simulate(fit, 120), col="red",lwd=2)
# simulation 2
set.seed(555)
lines(simulate(fit, 120), col="yellow",lwd=2)
# simulation 3
set.seed(323)
lines(simulate(fit, 120), col="green",lwd=2)
# simulation 4
set.seed(44)
lines(simulate(fit, 120), col="blue",lwd=2)
legend(1980, 2400,legend=c("Simulation 1", "Simulation 2", "
    Simulation 3", "Simulation 4"), col=c("red", "yellow", "
    green", "blue"), lty=1, lwd=2,cex=0.8)
```

Above, you are encouraged to change the values of the seeds, which will lead to different future potential time series paths. We used the ETS model to simulate future paths, but we could have used any other relevant model. We only used a simple functionality of the function `simulate`. To learn more about function `simulate` please visit: https://rdrr.io/cran/forecast/man/simulate.ets.html

Next, we will use the same data, the same ETS model, but instead of plotting several potential future time series paths, we will provide 80% and 95% prediction intervals for overseas visits to the UK. We accomplish this by using the function `forecast` and by specifying how far the prediction needs to go: 10 years, hence 120-time units (see $h = 120$ in row 2). The time units are months because the time series were collected at each month. Hence we will recreate Figure 3.4.

```
# R Code
# ------------------------------------------------------------
# Goal: To forecast Overseas visits using ets function
# ------------------------------------------------------------
#
# Remember, the structure mydata2 contains 2 years of data
    only.
fit <- ets(mydata2)
plot(forecast(fit,h=120),ylim=c(400, 2400),ylab="Visits[
    Thousands]", xlab="Year")
```

Next, we are practising creating an autocorrelation function for the monthly Overseas visits time series in R. We use the R function `Acf` (row 4). We generate the ACF figure 3.9 using R function `ggAcf`.

```
# R Code
# ------------------------------------------------------------
# Goal: To obtain ACF for 40 years of Overseas visits data.
# ------------------------------------------------------------
```

```
#
mydata<-read.csv("Risk-2021-DataVisitsToUK.csv")
mydataALL<-ts(mydata$Visits,start=c(1980,1),end=c(2020,3),
    frequency=12)
ggAcf(mydataALL)
my.acf<-Acf(mydataALL)

R Output
> my.acf
  Autocorrelations of series 'mydataALL', by lag
0     1     2     3     4     5     6     7     8     9    10
1.000 0.910 0.820 0.743 0.655 0.577 0.530 0.562 0.632 0.709
    0.772
11    12    13    14    15    16    17    18    19    20
0.851 0.915 0.845 0.756 0.684 0.598 0.521 0.478 0.507 0.570
21    22    23    24    25    26
0.647 0.705 0.781 0.839 0.770 0.686
```

Next, we calculate the ACF of monthly visits data again. But this time we do it step by step, without using the R built-in function `Acf`. We do it for lags 1 (rows 3-6) and 6 (rows 8-11). Note that the values are close but not the same as from the built-in function `Acf`, due to rounding.

```
# R Code
# ------------------------------------------------------------
# Goal: To obtain ACF at lag 1 and 6, for Overseas visits data
    .
# ------------------------------------------------------------
#
# First we obtain ACF for lag 1.
length(mydataALL) # T=483
Yt <-mydataALL[2:483]
Yt.lag1<-mydataALL[1:482]# lagged series, lag=1
cor(Yt,Yt.lag1) # correlation at lag 1.
length(mydataALL) # T=483
Yt <-mydataALL[7:483]
Yt.lag6<-mydataALL[1:477]# lagged series, lag=6
cor(Yt,Yt.lag6) # correlation at lag 6.

R Output
cor(Yt,Yt.lag1) # correlation at lag 1.
[1] 0.9130594
cor(Yt,Yt.lag6) # correlation at lag 6.
[1] 0.5426583
```

```
# ------------------------------------------------------------
# Goal: Fit an additive and a multiplicative HW model to
# 2 years of Overseas visits data.
# ------------------------------------------------------------
library(forecast)
# Data: Overseas Visits to UK data
mydata<-read.csv("Risk-2021-DataVisitsToUK.csv")
head(mydata,10) # shows the first 10 rows of data, for a quick
     check
# extract years 1980 and 1981 only, and put them into a ts
    object.
```

```
mydata.visits.ts<-ts(mydata$Visits,start=c(1980,1),end=c(1981,12),frequency=12)
# Create the Time for the first visit
dt1.visits <- as.Date("1980-01-01")
dt1.visits
# The length of our time series
T<-24 # 24 months only
# Then we create a whole column of n days:
Time.visits<-seq(dt1.visits, length = 24, by = "months")
str(Time.visits) # Checking that Time is indeed a date format.
# Fit the HW model with additive seasonality
hw.visits.additive <- hw(mydata.visits.ts,seasonal="additive",h=24)
summary(hw.visits.additive)
# Fit the HW model with multiplicative seasonality
hw.visits.multiplicative <- hw(mydata.visits.ts,seasonal="multiplicative",h=24)
summary(hw.visits.multiplicative)
# Next, we plot the data and the fitted values from the two models.
plot(Time.visits,mydata.visits.ts, ylab="Visits [thousands]",xlab="Year",
     lwd=2,type="l")
lines(Time.visits, hw.visits.additive$fitted,col="blue",lwd=2)
lines(Time.visits, hw.visits.multiplicative$fitted,col="pink",lwd=2)
# We redraw the original data, just to make the plot look nicer.
lines(Time.visits,mydata.visits.ts)
legend(Time.visits[1],1700, legend=c("Time Series", "HW additive", "HW multiplicative"),
       col=c("black","blue","pink"), lty=c(1,1,1),
       lwd=c(2,2,2),cex=0.8)
title("Overseas data and fitted values")

R OUTPUT:
> summary(hw.visits.additive)

Forecast method: Holt-Winters' additive method

Model Information:
Holt-Winters' additive method

Call:
 hw(y = mydata.visits.ts, h = 1, seasonal = "additive")

  Smoothing parameters:
    alpha = 1e-04
    beta  = 1e-04
    gamma = 1e-04

  Initial states:
    l = 1109.4158
    b = -8.7048
    s = -334.2999 -253.2824 0.8936 201.4947 803.099 653.9046
           69.9524 25.6193 -43.8918 -336.9516 -488.2117
    -298.3261
```

```
58
59    sigma:  85.943
60
61        AIC      AICc       BIC
62  297.6834  399.6834  317.7104
63
64  Error measures:
65                       ME     RMSE      MAE        MPE
66  Training set  -5.874759 49.61922 33.54391 0.01322887
67
68  MAPE      MASE      ACF1
69  3.559648 0.4158336 0.3167459
70
71  > summary(hw.visits.multiplicative)
72
73  Forecast method: Holt-Winters' multiplicative method
74
75  Model Information:
76  Holt-Winters' multiplicative method
77
78  Call:
79   hw(y = mydata.visits.ts, h = 1, seasonal = "multiplicative")
80
81    Smoothing parameters:
82      alpha = 0.054
83      beta  = 0.0477
84      gamma = 2e-04
85
86    Initial states:
87      l = 1108.865
88      b = -7.8516
89      s = 0.6556 0.7436 0.9872 1.2152 1.7783 1.5844
90             1.0807 1.0403 0.9809 0.6941 0.5493 0.6903
91
92    sigma:  0.0459
93
94        AIC      AICc       BIC
95  264.1459  366.1459  284.1728
96
97  Error measures:
98                      ME     RMSE      MAE       MPE
99  Training set  5.367679 28.21509 22.71365 0.6492115
100
101 MAPE      MASE      ACF1
102 2.24464 0.2815742 0.1951689
```

Next, we do a goodness-of-fit analysis of the additive HW model of the 2 years of Overseas visits data.

```
1  # R Code
2  # ------------------------------------------------------------
3  # Goal: To do a goodness-of-fit analysis of the additive HW
       model for first
4  # 2 years of Overseas visits data.
5  # Use the HW model with additive seasonality.
6  # ------------------------------------------------------------
7  # Plot of resid vs time
```

```
plot(Time2[1:24], fit.hw.visits$residuals,xlab="Year",ylab="
    Residuals",
     main="Residuals vs Time of HW model of 2 years of
    Overseas Visits data")
abline(0,0)
# ACF for residuals HW model of Overseas data data
Acf(hw.visits.additive$residuals,main="ACF of residuals of HW
    model of Overseas visits 2y",lag=30)
# ACF for original data. We plot it just to see how much ACF
    we removed with
# HW model.
Acf(mydata.visits.ts,main="ACF of 2 years of Overseas visits
    data",lag=30)
# Check assumption N: normality of residuals
# Create Q-Q plot
qqnorm(hw.visits.additive$residuals,main="Q-Q plot for
    residuals of HW model in Visits data")
# Add straight diagonal line to plot
qqline(hw.visits.additive$residuals)
# Get p-value
shapiro.test(hw.visits.additive$residuals)
# normality test for residuals
shapiro.test(hw.visits.additive$residuals)

R OUTPUT:
> # normality test for residuals
> shapiro.test(hw.visits.additive$residuals)

  Shapiro-Wilk normality test

data:  hw.visits.additive$residuals
W = 0.85463, p-value = 0.002667
```

Next, we do a goodness-of-fit analysis of the multiplicative HW model of the 2 years of Overseas visits data.

```
# R Code
# ------------------------------------------------------------
# Goal: To do a goodness-of-fit analysis of the multiplicative
     HW model for first
# 2 years of Overseas visits data.
# ------------------------------------------------------------
# Plot of resid vs time
plot(Time2[1:24], hw.visits.multiplicative$residuals,xlab="
    Year",ylab="Residuals",
     main="Residuals vs Time of HW model with multiplicative
    seasonality")
abline(0,0)
# ACF for residuals HW model of Overseas data data
Acf(hw.visits.multiplicative$residuals,main="ACF of residuals
    of HW model with multiplicative seasonality",lag=30)
# Check assumption N: normality of residuals
# Create Q-Q plot
qqnorm(hw.visits.multiplicative$residuals,main="Q-Q plot for
    residuals of HW model with multiplicative seasonality")
# Add straight diagonal line to plot
qqline(hw.visits.multiplicative$residuals)
```

```
# normality test for residuals, to get pvalue
shapiro.test(hw.visits.multiplicative$residuals)

R OUTPUT
> shapiro.test(hw.visits.multiplicative$residuals)

  Shapiro-Wilk normality test

data:  hw.visits.multiplicative$residuals
W = 0.9546, p-value = 0.3397
```

Next, we use the multiplicative Holt-Winters smoothing model that we fit into the first two years of data, and we will do the forecasts for the next three years, in R. We find the best parameter values, using R. And we create the Figure 3.24.

```
# R Code
# ------------------------------------------------------------
# Goal: Do use multiplicative HW model of first two years of Overseas visits data,
# and then construct the prediction intervals.
# ------------------------------------------------------------
#
# Fit HW multiplicative model again, let R choose the best parameters.
# calculate the forecast for the next 36 months.
# Ask for prediction intervals: 50% and 95%.
library(forecast)
fit.hw.visits <- hw(mydata.visits.ts,seasonal="multiplicative",h=36,level=c(50,95))
# Print the forecasts on the computer screen in Console
summary(fit.hw.visits.multiplicative)
# plot the model and forecasts.
plot(fit.hw.visits,xlab="Year",ylab="Overseas visitors to UK [thousands]",ylim=c(0,2500))
# Next, finally (!), we create a fan plot
fit.hw.visits <- hw(mydata.visits.ts,seasonal="multiplicative",h=36,fan=TRUE)
plot(fit.hw.visits,xlab="Year",ylab="Overseas visitors to UK [thousands]",
    ylim=c(0,2500))

R OUTPUT:
> # Print the forecasts on the computer screen in Console
> summary(fit.hw.visits.multiplicative)

Forecast method: Holt-Winters' multiplicative method

Model Information:
Holt-Winters' multiplicative method

Call:
 hw(y = mydata.visits.ts, h = 24, seasonal = "multiplicative")

  Smoothing parameters:
    alpha = 0.054
```

```
    beta  = 0.0477
    gamma = 2e-04

  Initial states:
    l = 1108.865
    b = -7.8516
    s = 0.6556 0.7436 0.9872 1.2152 1.7783 1.5844
           1.0807 1.0403 0.9809 0.6941 0.5493 0.6903

  sigma:  0.0459

     AIC     AICc      BIC
264.1459 366.1459 284.1728

Error measures:
                    ME     RMSE      MAE       MPE
Training set 5.367679 28.21509 22.71365 0.6492115

 MAPE      MASE      ACF1
2.24464 0.2815742 0.1951689

Forecasts:
         Point Forecast   Lo 80     Hi 80     Lo 95     Hi 95
Jan 1982      639.0357  601.4497  676.6218  581.5528  696.5187
Feb 1982      508.1726  478.1289  538.2162  462.2247  554.1204
Mar 1982      641.8192  603.4559  680.1825  583.1476  700.4908
Apr 1982      906.4547  851.2590  961.6503  822.0402  990.8691
May 1982      960.8800  900.7475 1021.0125  868.9153 1052.8447
Jun 1982      997.6307  932.8662 1062.3952  898.5819 1096.6794
Jul 1982     1461.7490 1362.4141 1561.0839 1309.8293 1613.6686
Aug 1982     1639.6424 1522.0507 1757.2340 1459.8015 1819.4833
Sep 1982     1119.8271 1034.4936 1205.1606  989.3207 1250.3335
Oct 1982      909.2438  835.2401  983.2474  796.0650 1022.4225
Nov 1982      684.4661  624.7467  744.1855  593.1332  775.7991
Dec 1982      603.1469  546.6067  659.6870  516.6762  689.6175
January 1983      634.7129  570.7146  698.7113  536.8359
    732.5900
Feb 1983      504.7331  449.9860  559.4801  421.0047  588.4614
Mar 1983      637.4727  563.1310  711.8144  523.7769  751.1685
Apr 1983      900.3125  787.5460 1013.0790  727.8510 1072.7740
May 1983      954.3653  826.1590 1082.5717  758.2907 1150.4400
Jun 1983      990.8630  848.3326 1133.3935  772.8815 1208.8446
Jul 1983     1451.8273 1228.6043 1675.0502 1110.4372 1793.2174
Aug 1983     1628.5069 1361.3648 1895.6491 1219.9483 2037.0656
Sep 1983     1112.2176  917.9235 1306.5117  815.0704 1409.3648
Oct 1983      903.0617  735.3758 1070.7476  646.6083 1159.5152
Nov 1983      679.8098  545.8784  813.7411  474.9795  884.6400
Dec 1983      599.0414  474.0458  724.0369  407.8772  790.2056
```

2. [**Purpose: Analysis of Kings life span data in R.**] Here, we will practice fitting linear regression and quadratic regression models to Kings life span data. Below the analysis is done in several sections, each having its own goal.

   First input data into R, then we create the Figure 3.11 of the Kings time series data, using the following R code:

```
# R Code
# ------------------------------------------------------------
# Goal: To input the Kings' life span data into R and plot it.
# ------------------------------------------------------------
# We will type data directly into R, by creating a vector y:
y<-c(60,43,67,50,56,42,50,65,68,43,
65,34,47,34,49,41,13,35,53,56,
16,43,69,59,48,59,86,55,68,51,
33,49,67,77,81,67,71,81,68,70,
77,56)
# Next, we create the time variable
Time<-c(1:length(y))
# Let us plot the response variable with respect to time
plot(Time, y, ylab="Age of death [years]",
   xlab="Order of being on the throne",type="l")
# Next, we generate a data frame with these variables
dataset.kings<-data.frame(Time, y)
```

Note that we did not put data into `ts` format, as this was not needed. Next, we estimate the linear regression model:

```
# Rcode
# ------------------------------------------------------------
# Goal: Fit linear model to Kings' life span data.
# ------------------------------------------------------------
#
# We will fit a simple linear regression to this time series,
# where there is only one predictor: the time
kings.lin<-lm(formula= y~Time, data=dataset.kings)
# Next, we obtain detailed information on our regression using
# the summary() command. Note that kings.lin is now an object
# created by R, it contains the data and the estimated
    parameters.
# So we need to extract those parameters and summaries:
summary(kings.lin)

R Output:
> summary(kings.lin)

Call:
lm(formula = y ~ Time, data = dataset.kings)

Residuals:
    Min      1Q  Median      3Q     Max
-39.833  -9.125   1.942  10.742  27.717

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  43.5679     4.8227   9.034 3.32e-11 ***
Time          0.5450     0.1954   2.789  0.00805 **
---
#Signif. codes:  0    ***    0.001    **    0.01    *    0.05
       .    0.1         1

Residual standard error: 15.35 on 40 degrees of freedom
Multiple R-squared:  0.1628,  Adjusted R-squared:  0.1419
F-statistic:  7.78 on 1 and 40 DF,  p-value: 0.008051
```

Next, we plot the fitted values from the linear regression on the same plot as the actual data:

```
# R Code
# ------------------------------------------------------------
# Goal: To plot the fitted linear model together with Kings
    time series data.
# ------------------------------------------------------------
#
plot(Time, y, ylab="Age when King died [years]",xlab="Order of
    being
   on the throne",type="l")
y.predict.lin <-
   kings.lin$coefficients[[1]]+kings.lin$coefficients[[2]]*
    Time
# Since Time is already ordered, we can superimpose the fitted
    (predicted) values:
lines(Time, y.predict.lin, col="blue",lwd=2)
# We calculate the residuals
resid.lin<-y-y.predict.lin
# Next, we plot the residuals vs time
plot(Time, resid.lin, type="l",ylim=c(-40,40),
  ylab="Residuals in model with linear term")
```

Next, we are practising estimation of a model with quadratic trend, to Kings time series data. We estimate the quadratic trend using the following R code:

```
# R Code
# ------------------------------------------------------------
# Goal: To fit quadratic regression to the Kings' life span
    data.
# ------------------------------------------------------------
#
# Data is already entered in R (see the previous question).
# Next we fit the quadratic model with two covariates: Time
    and Time^2
# We store the results into a structure called kings.quad
kings.quad<-lm(y~Time+I(Time^2),data=dataset.kings)
# Or alternatively
# dataset.kings$Time2<-dataset.kings$Time^2
# kings.quad<-lm(y~Time+Time2,data=dataset.kings)
#
summary(kings.quad)

Output:
> summary(kings.quad)

Call:
lm(formula = y ~ Time + I(Time^2), data = dataset.kings)

Residuals:
    Min      1Q  Median      3Q     Max
-34.189  -6.073   0.762   8.459  32.915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 57.61934    6.95262   8.287 3.93e-10 ***
Time        -1.37109    0.74575  -1.839   0.0736 .
I(Time^2)    0.04456    0.01682   2.650   0.0116 *
---

Signif. codes:  0    ***    0.001    **    0.01    *    0.05
       .    0.1         1

Residual standard error: 14.31 on 39 degrees of freedom
Multiple R-squared:  0.2905,  Adjusted R-squared:  0.2542
F-statistic: 7.986 on 2 and 39 DF,  p-value: 0.001239
```

The above output from R was discussed earlier in this Section 3.4.1.

Next, we show the R code to produce the Figure 3.13:

```
# R Code
# ------------------------------------------------------------
# Goal: to plot the King's life span data with the estimated
    linear and quadratic models.
# And to create the residual plot for the quadratic model.
# ------------------------------------------------------------
#
plot(Time, y, ylab="Age when King died [years]",xlab="Order of
     being
   on the throne",type="l")
y.predict.quad <-
  kings.quad$coefficients[[1]]+
  kings.quad$coefficients[[2]]*Time+
  kings.quad$coefficients[[3]]*Time^2
# alternatively: y.predict.quad<-predict(kings.quad)
# Since Time is already ordered, we can plot it:
lines(Time, y.predict.lin, col="blue",lwd=2)
lines(Time, y.predict.quad, col="green",lwd=2)
# We see the quadratic model has better residuals
resid.quad <- y - y.predict.quad
plot(Time, resid.quad, type="l",ylim=c(-40,40),
   ylab="Residuals in model with quad term")
```

Next, we are practising obtaining the values of the information criteria for the linear and quadratic models of Kings' life span data. This is so we can then compare the two models and tell which model is better. Below we show the R code only. The interpretation and discussion was Section 3.4.3. We use the following criteria: SSE, RMSE, R-squared, R-squared adjusted, AIC and BIC.

```
# R Code
# ------------------------------------------------------------
# Goal: Get the values of the information criteria for the
    linear model of Kings' life span data.
# ------------------------------------------------------------
#
# First we calculate the SSE, and RMSE by using their
    definitions:
# We assume we already fitted the linear model and we stored
    the fitted model
```

```
# in a structure called y.predict.lin
# Next, we get fitted values
y.predict.lin <-
  kings.lin$coefficients[[1]]+kings.lin$coefficients[[2]]*Time
# Alternatively we could do: y.predict.lin<-predict(kings.lin)
# We get the residuals
resid.lin<-y-y.predict.lin
# Get length of the data
n= dim(dataset.kings)[1]
n
# We get SSE
SSE.lin = sum(resid.lin^2)
SSE.lin
# We get degrees of freedom
kings.lin$df.residual
# We get RMSE
RMSE.lin = sqrt(SSE.lin / kings.lin$df.residual)
RMSE.lin
# Here we use summary function to get RMSE, R-squared and R-
    squared adjusted
summary(kings.lin)
# Next we get AIC and BIC
AIC(kings.lin)
BIC(kings.lin)

R output:
> n
[1] 42
> SSE.lin
[1] 9423.694
> kings.lin$df.residual
[1] 40
> RMSE.lin
[1] 15.34902
> summary(kings.lin)

Call:
lm(formula = y ~ Time, data = dataset.kings)

Residuals:
    Min      1Q  Median      3Q     Max
-39.833  -9.125   1.942  10.742  27.717

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  43.5679     4.8227   9.034 3.32e-11 ***
Time          0.5450     0.1954   2.789  0.00805 **
---
% Signif. codes:  0    ***    0.001    **    0.01    *    0.05
        .    0.1         1

Residual standard error: 15.35 on 40 degrees of freedom
Multiple R-squared:  0.1628,  Adjusted R-squared:  0.1419
F-statistic:  7.78 on 1 and 40 DF,  p-value: 0.008051

> # Next we get AIC and BIC
> AIC(kings.lin)
```

```
[1] 352.55
> BIC(kings.lin)
[1] 357.763
```

The output above shows that for the linear model, we have: SSE=9423.694, RMSE=15.34902 (the same as the Residual standard error 15.35 in the output of the function `summary`), R-squared=0.1628, R-squared adjusted = 0.1419, AIC=352.55, BIC=357.763.

```
# R Code
# ------------------------------------------------------------
# Goal: Get the summary statistics for the quadratic model of
    Kings life span data.
# ------------------------------------------------------------
#
# First we calculate the SSE, and RMSE by using their
    definitions:
# We assume we already fitted the quadratic model and we
    stored the fitted model
# in a structure called y.predict.quad
# Next, we get fitted values
y.predict.quad <-
  kings.quad$coefficients[[1]]+
  kings.quad$coefficients[[2]]*Time+
  kings.quad$coefficients[[3]]*Time^2
# Or alternatively we could do: y.predict.quad<-predict(kings.
    quad)
# We get the residuals
resid.quad<-y-y.predict.quad
# Get length of the data
n= dim(dataset.kings)[1]
n
# We get SSE
SSE.quad = sum(resid.quad^2)
SSE.quad
# We get degrees of freedom
kings.quad$df.residual
# We get RMSE
RMSE.quad = sqrt(SSE.quad / kings.quad$df.residual)
RMSE.quad
# Here we use summary function to get RMSE, R-squared and R-
    squared adjusted
summary(kings.quad)
# Next we get AIC and BIC
AIC(kings.quad)
BIC(kings.quad)

R Output:
> n
[1] 42
> SSE.quad
[1] 7986.085
> kings.quad$df.residual
[1] 39
> RMSE.quad
[1] 14.30984
```

```
> summary(kings.quad)

Call:
lm(formula = y ~ Time + I(Time^2), data = dataset.kings)

Residuals:
    Min      1Q  Median      3Q     Max
-34.189  -6.073   0.762   8.459  32.915

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 57.61934    6.95262   8.287 3.93e-10 ***
Time        -1.37109    0.74575  -1.839   0.0736 .
I(Time^2)    0.04456    0.01682   2.650   0.0116 *
---
% Signif. codes:  0    ***    0.001    **    0.01    *    0.05
        .    0.1         1

Residual standard error: 14.31 on 39 degrees of freedom
Multiple R-squared:  0.2905,  Adjusted R-squared:  0.2542
F-statistic: 7.986 on 2 and 39 DF,  p-value: 0.001239

> # Next we get AIC and BIC
> AIC(kings.quad)
[1] 347.5979
> BIC(kings.quad)
[1] 354.5485
>
```

The output above shows that for the quadratic model, we have: SSE=7986.085, RMSE=14.30984 (the same as the Residual standard error 14.31 in the output of the function `summary`), R-squared=0.2905, R-squared adjusted = 0.2542, AIC=347.5979, BIC=354.5485.

We found that the quadratic model is a better fit than the linear model. Hence in the next, we are going to check if the quadratic model is also a good enough fit. Next, we practice how to do a goodness-of-fit test of the quadratic model of Kings' life span data, in `R`.

```
# ------------------------------------------------------------
# Goal: Check if the quadratic model is a well-fitting model
    of the Kings lifespan data
# ------------------------------------------------------------
# R code
#
# Get errors from the quadratic model.
resid.quad<-y-y.predict.quad
# Add a zero line
abline(0,0)
# Check assumption Z: zero means of residuals
plot(Time, resid.quad,main="Residuals vs time of quadratic
    model in Kings life span data")
# Check assumption I: independence of residuals
library(forecast)
Acf(resid.quad,20,main="ACF plot for residuals of quadratic
    model in Kings life span data")
```

```
# Check assumption N: normality of residuals
# Create Q-Q plot
qqnorm(resid.quad,main="Q-Q plot for residuals of quadratic
    model in Kings life span data")
# Add straight diagonal line to the plot
qqline(resid.quad)
# Get p-value of normality of residuals test
shapiro.test(resid.quad)
# Check assumption C: constant variance of residuals
# We check it by visual inspection of the plot we did for
    assumption Z.

R OUTPUT

> # Get p-value of a normality test
> shapiro.test(resid.quad)

  Shapiro-Wilk normality test

data:  resid.quad
W = 0.97771, p-value = 0.5744
```

```
# R Code
#
    ------------------------------------------------------------------

# Goal: We are going to predict the Age at death for Time
    values 43, 44 and 45,
# i.e. the next three kings. So first we create a data frame
    containing
# the values of Time for the next three kings.
#
    ------------------------------------------------------------------

#
# New time points for which we want the forecast
mydatanew<-data.frame(Time=c(43,44,45))
# Next we can use the function predict for the prediction.
    Luckily, R is an
# object-oriented software. This means that R will first look
    into what structure
# the object kings.quad is, it will recognise that it is an
    object created by
# lm() function, so R will figure out how to do the prediction
     using its
# function called predict(). We just need to provide the new
    values of Time
# (i.e. the new values of the predictor variable) for which we
     want the predictions.
# Hence getting the point predictions for the three times is
    as simple as:
predict(kings.quad,newdata=mydatanew)
# Or getting the 95% prediction intervals is also simple:
predict(kings.quad,newdata=mydatanew,interval="prediction",
    level=0.95)
# Or 80% prediciton intervals
```

```
predict(kings.quad,newdata=mydatanew,interval="prediction",
    level=0.80)
#
# NOTE 1:
# In quadratic regression, we have two predictors: Time and
    Time^2.
#
# Note 2:
# Further detail of the predict function for the linear
    regression model can be
# found in the R documentation.
# > help(predict.lm)
#
# Finally, using the quadratic model we will plot the data
    with the prediction:
# Plot the y_hat for the quadratic model, with curve width lwd
    =2
mydatanew<-data.frame(Time=c(43,44,45))
plot(Time, y, ylab="Age when passed away [years]",
     xlab="Order of being on the throne",type="l",ylim=c
    (10,120),xlim=c(0,46),
     lwd=2)
lines(Time, y.predict.quad,col="green",lwd=2)
kings.pred.int80<-predict(kings.quad,newdata=mydatanew,
    interval="prediction",level=0.80)
kings.pred.int95<-predict(kings.quad,newdata=mydatanew,
    interval="prediction",level=0.95)
lines(mydatanew$Time, kings.pred.int80[,1],col="cyan",lwd=2) #
     plotting 1st column
lines(mydatanew$Time, kings.pred.int80[,3],col="red",lwd=3) #
    plotting 3rd column
lines(mydatanew$Time, kings.pred.int80[,2],col="red",lwd=3) #
    plotting 2nd column
lines(mydatanew$Time, kings.pred.int95[,3],col="pink",lwd=5) #
     plotting 3rd column
lines(mydatanew$Time, kings.pred.int95[,2],col="pink",lwd=5) #
     plotting 2nd column
lines(c(0,140),c(70,70),col="purple", lty=2)
legend(0,120, legend=c("Time Series", "Quadratic Model", "
    Point Forecast",
       "80% Prediction Interval", "95% Prediction Interval", "
    70 years line"),
       col=c("black","green","cyan","red","pink","purple"),
    lty=c(1,1,1,1,1,2),
       lwd=c(2,2,2,3,5,1),cex=0.8)
title("Forecasts the next 3 Kings using the quadratic model")

R OUTPUT:

> # Hence getting the point predictions for the three times is
     as simple as:
> predict(kings.quad,newdata=mydatanew)
       1        2        3
81.05488 83.56055 86.15534
> # Or getting the 95% prediction intervals is also simple:
> predict(kings.quad,newdata=mydatanew,interval="prediction",
    level=0.95)
```

```
       fit      lwr      upr
1 81.05488 48.87500 113.2348
2 83.56055 50.76011 116.3610
3 86.15534 52.64689 119.6638
> # Or 80% prediction intervals
> predict(kings.quad,newdata=mydatanew,interval="prediction",
    level=0.80)
       fit      lwr      upr
1 81.05488 60.31472 101.7950
2 83.56055 62.42043 104.7007
3 86.15534 64.55891 107.7518
```

3. [**Purpose: Analysis of central England temperatures data.**] Here we will do simple exponential smoothing and Holt's exponential smoothing for Temperature data.

```
# R Code
# ------------------------------------------------------------
# Goal: Input the England Temperature Data, plot it and then
    perform SES (Simple Exponential Smoothing). We will use
    alpha 0.6
# ------------------------------------------------------------
#
# Copy and paste data:
Y <-
c(17.3, 17.9, 17.3, 15.4, 15.0, 17.6, 18.2, 17.2, 16.6, 15.7,
  15.1, 16.8, 17.2, 18.7, 19.4, 18.3, 17.9, 18.5, 20.3, 19.5,
  19.2, 20.2, 19.8, 20.2, 21.7, 19.8, 19.7, 18.3, 19.3, 17.3,
  18.5)
# Check if Y was created correctly
Y
n<-length(Y)
# Create the Time variable
dt1 <- as.Date("2004-07-14")
dt1
# Then we create a whole column of n days:
Time<-seq(dt1, length = 31, by = "days")
str(Time) # Checking that Time is indeed a date format.
head(Time,13) # checking first 13 values of Time
# Initialise the vector Yhat of size 1xn, of values NA
# NA is how R codes missing values. So we created an empty
    vector.
Yhat=rep(NA,n)
Yhat
# Need to initialise the first value of Yhat. One way to do it
    is to simply copy the first value of Y.
Yhat[1]<-Y[1]
Yhat
## Create the constant alpha
alpha<-0.6
# Do calculate other values of Yhat. Use a 'for' function to
    create a loop:
for (i in seq(2,n)){
  Yhat[i]<- alpha *Y[i-1] + (1-alpha)*Yhat[i-1]
}
Yhat
```

```
# Now calculate the squared errors for each of the predictions
error<-Y-Yhat
# Squares of errors
error^2
# Finally, the sum of squared errors
SSE<-sum(error^2)
SSE
# Plot the model and fitted values
plot(Time,Y,ylab="Central England temperatures [Celsia]", xlab
    ="Day",type="l",main="Central England temperatures in 2004
    ")
# Add the plot of the fitted values
lines(Time, Yhat,lty=2)
# Next add a legend to the plot
legend(dt1,21, legend=c("Time Series", "SES alpha=0.6"), col=c
    ("black","black"), lty=c(1,2), lwd=c(1,1),cex=0.8)
```

Next, we present an alternative solution via the built-in function `ses` in R, from library `forecast`:

```
# R Code
# ------------------------------------------------------------
# Goal: An alternative solution to the above goal, now using
    the
# function ses from the library called forecast.
# ------------------------------------------------------------
#
# Attach the library forecast
library(forecast)
# Next use the function ses from the library forecast.
# The function ses is able to choose the most optimal alpha,
    but we will
# not use such functionality now, we will ask ses to use alpha
    =0.6
fit.ses.Y <- ses(Y,alpha=0.6)
# plot the model and fitted values (Yhat)
plot(Time,Y,ylab="Central England temperatures [Celsia]", xlab
    ="Day",type="l",main="Central England temperatures in 2004
    ")
lines(Time, fit.ses.Y$fitted,lty=2)
legend(dt1,21, legend=c("Time Series", "SES alpha=0.6"), col=c
    ("black","black"), lty=c(1,2), lwd=c(1,1),cex=0.8)

# Next, we check the values of the estimated parameters.
# Note that the output in R will show us that the ses model is
     using
# alpha=0.6
summary(fit.ses.Y)

R OUTPUT:

> summary(fit.ses.Y)

Forecast method: Simple exponential smoothing

Model Information:
Simple exponential smoothing
```

```
31
32 Call:
33  ses(y = Y, alpha = 0.6)
34
35   Smoothing parameters:
36     alpha = 0.6
37
38   Initial states:
39     l = 17.3391
40
41   sigma:  1.1852
42
43      AIC     AICc      BIC
44 118.9222 119.3508 121.7902
45
46 Error measures:
47 Training set
48   ME          RMSE      MAE
49   0.05262822 1.146358 0.9558096
50   MPE          MAPE      MASE       ACF1
51   0.002338004 5.352873 0.9558096 0.1829013
52
53 Forecasts:
54    Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
55 32       18.31794 16.79901 19.83687 15.99493 20.64094
56 33       18.31794 16.54658 20.08930 15.60887 21.02701
57 34       18.31794 16.32588 20.31000 15.27135 21.36453
58 35       18.31794 16.12731 20.50857 14.96765 21.66823
59 36       18.31794 15.94529 20.69059 14.68929 21.94659
60 37       18.31794 15.77628 20.85960 14.43081 22.20507
61 38       18.31794 15.61783 21.01805 14.18848 22.44740
62 39       18.31794 15.46817 21.16771 13.95960 22.67628
63 40       18.31794 15.32599 21.30989 13.74215 22.89373
64 41       18.31794 15.19027 21.44561 13.53458 23.10130
```

Next, we will do simple exponential smoothing again, for Central England temperature data again. However, this time we will find and use the optimal value of α. We find such optimal value, and for that, we will use the function `ses`.

```
1  # R Code
2  # ------------------------------------------------------------
3  # Goal: Input the Central England Temperature Data, plot it
       and then perform SES with optimal alpha.
4  # ------------------------------------------------------------
5  #
6  # Next use the function ses from the library forecast.
7  # The function ses is able to choose the most optimal alpha.
8  # What we do is we do not tell ses which alpha to use, and
       that way
9  # the function ses knows that the optimal alpha needs to be
       found first.
10 fit.ses.Y <- ses(Y)
11 # plot the model and fitted values (Yhat)
12 plot(Time,Y,ylab="Central England temperatures [Celsia]", xlab
```

```
    ="Day",type="l",main="Central England temperatures in 2004
    ")
lines(Time, fit.ses.Y$fitted,lty=2)
legend(dt1,21, legend=c("Time Series", "SES alpha=0.6"), col=c
    ("black","black"), lty=c(1,2), lwd=c(1,1),cex=0.8)

# Next, we check the values of the estimated parameters.
# Note that the output in R will show us that the ses model is
    using
# A different value of alpha (not 0.6)
summary(fit.ses.Y)

R Output:
> summary(fit.ses.Y)

Forecast method: Simple exponential smoothing

Model Information:
Simple exponential smoothing

Call:
 ses(y = Y)

  Smoothing parameters:
    alpha = 0.7743

  Initial states:
    l = 17.3835

  sigma:  1.1775

     AIC     AICc      BIC
120.5181 121.4070 124.8201

Error measures:
Training set
ME         RMSE      MAE         MPE        MAPE     MASE
0.03915644 1.138912 0.9796218 -0.03237187 5.45775 0.9796218
ACF1
0.05465299

Forecasts:
   Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
32       18.32345 16.81438 19.83251 16.01553 20.63136
33       18.32345 16.41487 20.23202 15.40453 21.24236
34       18.32345 16.08558 20.56131 14.90092 21.74597
35       18.32345 15.79888 20.84801 14.46245 22.18444
36       18.32345 15.54157 21.10532 14.06893 22.57796
37       18.32345 15.30613 21.34077 13.70885 22.93804
38       18.32345 15.08777 21.55912 13.37490 23.27199
39       18.32345 14.88324 21.76365 13.06211 23.58478
40       18.32345 14.69021 21.95668 12.76689 23.88000
41       18.32345 14.50693 22.13996 12.48659 24.16030
```

Next, we do Holts' exponential smoothing to Central England temperature data. We also recreate the Figure 3.18.

```
# R Code
# ------------------------------------------------------------
# Goal: Holts' exponential smoothing to Central England
# temperatures data and plot.
# ------------------------------------------------------------
# R code
#
# Plot of data and fitted values
plot(Time,Y,type="l",
     xlab="Day",
     ylab="Central England temperatures [Celsia]" ,main="
    Central England temperatures in 2004",lwd=2)
fit.holt.Y<-holt(Y)
lines(Time,fit.ses.Y$fitted,lty=3,col="blue",lwd=2)
lines(Time,fit.holt.Y$fitted,lty=2,col="red",lwd=3)
legend(Time[1],21, legend=c("Time Series", "SES", "Holt"),
       col=c("black","blue","red"), lty=2, lwd=c(1,2,3),cex
    =0.8)
```

In what follows, there are further R Lab questions for you to work on. Solutions are not provided here. They are provided upon request.

4. [**Purpose: to practice forecasting.**] You are asked to do forecasting for Central England Temperature data. Do all analyses in R.
   (a) Using R, and suitable criteria compare SES and Holt's models.
   (b) For the model that is better, do the goodness-of-fit analysis.
   (c) Do the forecast for 6-time points ahead. Do you trust your forecasts? Explain.

5. [**Purpose: to practice forecasting.**] You are asked to analyse securities yield time series data (Risk-2021-Yield-Data.csv). Do all analyses in R.
   (a) Use R to model the data in three ways: simple exponential smoothing, Holts, and Holt-Winters. Which is the best in terms of RMSE?
   (b) Generate the autocorrelation (correlogram) for the Yield time series.
   (c) Generate the autocorrelation (correlogram) for the errors of each of the three models.
   (d) Look at the plotted autocorrelation and judge if there is evidence against the model. Based on autocorrelation which model is the best? Based on autocorrelation can we trust the predictions of the model?
   (e) Fit a Holt-Winters model to the data.
   (f) Generate a one-step-ahead prediction and prediction intervals for that prediction.
   (g) State your assumptions upon which the interval is based.
   (h) If the assumptions are not satisfied, can we trust the prediction intervals?

6. [**Purpose: to practice forecasting.**] You are asked to analyse Plastic sales data. The plastics data set consists of the monthly sales (in thousands) of product A for a plastics manufacturer for five years. The dataset is stored as PlasticsMonthlySales.csv. Do all analyses in `R`.

   (a) Plot the time series of sales of product A. Visually, can you identify seasonal fluctuations, a trend or a cyclic component?

   (b) Calculate the logarithm of the time series and plot it again.

   (c) Use a classical additive decomposition to calculate the trend-cycle and seasonal indices.

   (d) Do the results support the graphical interpretation from part (a)?

7. **Purpose: to practice forecasting.**] You will work with Monthly data of totals of international airline passengers (Jan 1949- Dec 1960). The data are stored in the file AirPassengers.csv. Do all analyses in `R`.

   (a) Plot the time series data to get a feel for its structure. What type of time series decompositions should be done? Additive or multiplicative? Why?

   (b) Take the logarithm of the series and plot it. Does the log-transformed time series look suitable for additive or multiplicative decomposition?

   (c) Use the decompose function to do the time series decomposition (of original passenger data) automatically, twice: additive, and multiplicative decomposition. Create the plots and comment on the plots: do they differ? Why?

8. **Purpose: to practice forecasting.**] You are asked to analyse Yield data (Jan 1950-Dec 1970). They are stored in a file called Yield.csv. Do all analyses in `R`.

   (a) Plot the time series data. Does the seasonality appear additive or multiplicative? Why?

   (b) Do seasonal decomposition of the data assuming an additive model for the components, by using the `decompose` `R` function.

   (c) Next, do seasonal decomposition of the data assuming a multiplicative seasonality, by using the `decompose` `R` function and compare with the additive decomposition.

3.9 Exercises

Solve the following exercises by **using pen, paper and calculator**.

1. [**Purpose: Practicing simple exponential smoothing.** Given the following central England mean temperatures
$\mathbf{y} = (17.3, 17.9, 17.3, 15.4, 15.0, 17.6, 18.2, 17.2, 16.6, 15.7,$
$15.1, 16.8, 17.2, 18.7, 19.4, 18.3, 17.9, 18.5, 20.3, 19.5, 19.2,$
$20.2, 19.8, 20.2, 21.7, 19.8, 19.7, 18.3, 19.3, 17.3, 18.5)$
do the following:

   (a) Using simple exponential smoothing, and $\alpha = 0.4$, calculate the first 7 predictions (i.e. Y_t for $t = 1, \cdots, 7$).

   (b) Now calculate the squared errors for each of the predictions.

   (c) Finally, find the sum of squared errors.

   (d) Repeat the procedure for $\alpha = 0.3$. Which is the better value of α? Justify your answer.

2. [**Purpose: Practicing the decision process on what predictor variables should be used.**] You are asked to list the possible predictor variables that might be useful, assuming that the relevant data are available. (Adopted from Hyndman & Athanasopoulos, Forecasting: Principles and Practice, the online book).

   Case 1 A large car fleet company asked us to help them forecast vehicle resale values. They purchase new vehicles, lease them out for three years, and then sell them. Better forecasts of vehicle sales values would mean better control of profits; understanding what affects resale values may allow leasing and sales policies to be developed in order to maximize profits. At the time, the resale values were being forecast by a group of specialists. Unfortunately, they saw any statistical model as a threat to their jobs and were uncooperative in providing information. Nevertheless, the company provided a large amount of data on previous vehicles and their eventual resale values.

   Case 2 In this project, we needed to develop a model for forecasting weekly air passenger traffic on major domestic routes for one of Australia's leading airlines. The company required forecasts of passenger numbers for each major domestic route and for each class of passenger (economy class, business class and first class). The company provided weekly traffic data from the previous six years. Air passenger numbers are affected by school holidays, major sporting events, advertising campaigns, competition behaviour, etc. School holidays often do not coincide in different Australian cities, and sporting events sometimes move from one city to another. During the period of the historical data, there was a major pilot's strike during which there was no traffic for several months. A new cut-price airline also launched and folded. Towards the end of the historical data, the airline had trialled a redistribution of some economy class seats to business class, and some business class seats to first class. After several months, however, the seat classifications reverted to the original distribution.

3. [**Purpose: Practicing a visual inspection of plotted time series - recognising the trend, seasonal and cyclic changes.**] In the following Figure 3.25, there are four data series showing trend, seasonality, and cyclic behaviours. Which shows increasing seasonality with the increasing trend?

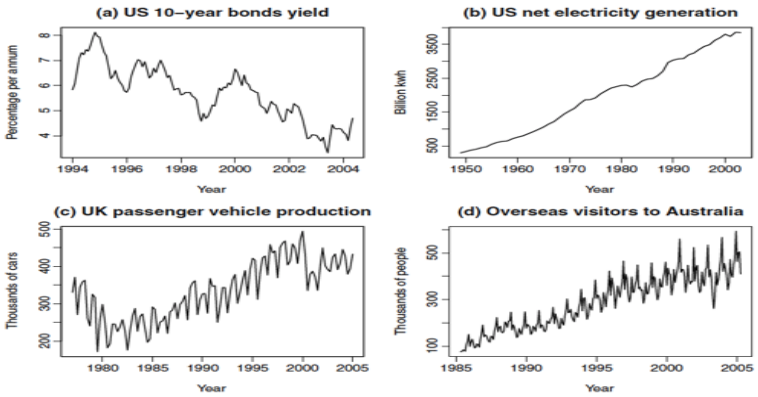


FIGURE 3.25: Four-time series showing patterns typical of business and economic data. Source: https://www.stat.berkeley.edu/ arturof/Teaching/STAT248/lab10$_part1.html$

4. [**Purpose: Practicing the estimation of covariance and correlation matrix.**] We have a short Olympic Women's 100 m data. Use these data to estimate the covariance matrix between time t and values Y. Use these data to estimate the correlation matrix.

	t_i	Y_t	u	v	$u \cdot u$	$u \cdot v$	$v \cdot u$	$v \cdot v$
	1928	12.2						
	1932	11.5						
	1948	11.9						
	1952	11.5						
E[]								

5. [**Purpose: Practicing the calculation of the lagged time series** Given a time series with the following values $\mathbf{y} = (5, 1, 4, -9, 3, -3, 7, 0, -1, 8)$ What is the time series lagged by 5? Use pen, paper and a calculator and clearly show your work.

6. [**Purpose: Practicing the calculation of autocorrelation.**] Use the following time series values $\mathbf{y} = (5, 1, 4, -9, 3, -3, 7, 0, -1, 8)$

   - Which pairs of values are used for calculating the autocorrelations r_3 and r_6?
   - Calculate the two autocorrelations, r_3 and r_6. Use a pen, paper and a calculator and clearly show your work.

- In any time series (not just this one), the autocorrelation at lag 0 is 1. Why is it?

7. [**Purpose: Practicing the calculation of moving averages for yearly time series data.**]

Year	**Average Temperature**
1659	8.83
1660	9.08
1661	9.75
1662	9.50
1663	8.58

   (a) Calculate the moving average of order 3 for the years 1660, 1661 and 1662.

   (b) Why can the moving average of order 3 not be calculated for the years 1659 and 1663

   (c) Is it possible to calculate a moving average of order 5 for these data?

8. [**Purpose: Practicing the calculation of moving averages for quarterly time series data.**] Here is a simple dataset of quarterly beer production data. (data from Hyndman's book, Chapter 6.2).

Year	**Quarter**	**Observation**
1992	Q1	443
1992	Q2	410
1992	Q3	420
1992	Q4	532
1993	Q1	433
1993	Q2	421
1993	Q3	410
1993	Q4	512
1994	Q1	449
1994	Q2	381
1994	Q3	423
1994	Q4	531

   (a) Calculate the moving average of order 4 using calculator.

   (b) Calculate the moving average of order 4 using R and compare.

4

Markov chains

Gabriela Czanner

Monika Kovacova

CONTENTS

4.1 Building a Markov Chain model 144
4.1.1 Terminology 145
4.1.2 The Markov property 147
4.1.3 One-step transition probabilities 148
4.1.4 Initial and one-step probability distributions 149
4.1.5 Multistep transition probabilities 153
4.1.6 Long-run prediction of the state 154
4.2 Further topics on Markov Chains 157
4.2.1 Commuter Cyril example 157
4.2.2 Regular Markov chains 159
4.2.3 Steady-state theorem 159
4.2.4 Interpretation of the long run distribution 160
4.2.5 Irreducible Markov Chains 163
4.2.6 Periodic and aperiodic Markov chains 165
4.3 Tips to think and act like a risk expert 166
4.3.1 Not all sequences are Markov Chains but some can be turned into Markov Chains 166
4.3.2 Sensitivity analysis 168
4.4 Summary 169
4.5 Further reading 169
4.6 R lab 170
4.7 Exercises 177

In this chapter, we are exploring a widely applicable probability model called a *Markov chain*, named after Russian mathematician A. A. Markov (1856-1922). He observed that many real-world phenomena can be modelled as a sequence of transitions from one *state* to another, while there is some *probability* about the transition. For example, a person can be healthy and each day the person

can transition to an unhealthy state with some probability, or it can stay in the healthy state. Or a homeowner's house has a certain value and each day it can transition to another value, it can go up or down, or stay the same. Or a bus driver is transitioning between several locations in a city.

In all these examples, we need to know the probabilities of transitioning from one state to another state or staying in the same state. Additionally, we need to know which state the system is in: where is the taxi driver now, how much money the gambler has now, and what the value of the house is today. Then we will be able to predict the next state. For the taxi driver, we do not need to know how he got into the current position, we need to know where he is now and the transition probabilities, and then we can guess where he will be next. For the house price, we do not need to know how the house price evolved in the past, we only need to know the current price and the transition probabilities, then we can make a guess what the next day price will be. For the gambler, we do not need to know how he got to the current state, we only need to know how much he owns now, and then we can estimate his next state.

Learning objectives

- Learn what Markov chains are and how they are mathematically defined.
- Learn matrix notation to facilitate Markov chain computations.
- Learn about regular chains and their exceptional property embodied in the Steady State Theorem.
- Learn about special types of Markov chains: irreducible, periodic and aperiodic.
- Learn to simulate Markov chains in `R`.

4.1 Building a Markov Chain model

We will start with an example, and then we introduce theory.

Example: Cereals buyers. We will consider the following scenario. Company K, the manufacturer of breakfast cereal, currently has 25% of the market. Data from the previous year indicates that 88% of K's customers remained loyal that year, but 12% switched to the competition. In addition, 85% of the competition's customers remained loyal to the competition, but 15% of the competitor's customers switched to K (Table 4.1).

Note, that the above implies that the decision to stay or to switch the cereals brand is happening at discrete time points, once a year. Without loss

		Next cereal by	
		K	Competition
Last cereal by	K	0.88	0.12
	Competition	0.15	0.85

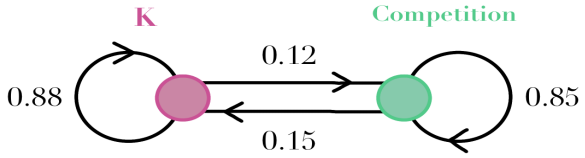


FIGURE 4.1: Cereal buyer's consumption state-transition diagram.

of generality, we can imagine that this decision is done by each customer on 1 January of each year. Assuming that these trends continue, we are asked to determine K's share of the market:

a) in 2 years,

b) in the long-run (i.e. the long-term prediction).

This cereals buyers' problem is an example of a brand-switching problem that often arises in the sale of consumer goods. We can construct a diagram in Figure 4.1, where the two circles represent the two states a customer can be in, and the arcs represent the probability that a customer makes a transition each year between states. Note the circular arcs indicating a transition from one state to the same state. The diagram is called the state-transition diagram.

In every such state transition diagram (Figure 4.1), the sum of probabilities on branches exiting a state must equal 1. For example, in Figure 1 the probabilities of exiting state K (i.e. state 1, the Brand K cereal) are 0.12 and 0.88, they add up to 1. We include in this calculation the probability 0.88 indicated by a loop in the state diagram, which simply means that the customer has a 0.88 probability of staying in state K (i.e. 1) in the next time step, once he/she has been buying K brand cereals in the present time step (here the steps are years).

4.1.1 Terminology

Define Y_0 to be the cereal brand which was bought by the customer in the initial time step, hence at time 0. Define Y_n be the cereal brand which is bought by the customer at time n (in this example n-th year). Since $Y_0, Y_1, Y_2, \ldots$ "occur" in sequence, they are often referred to as a chain. More precisely, this

particular sequence is a finite-state, discrete-time, time-homogeneous Markov chain. Each of these terms is explained below.

We will define Markov Chain mathematically.

- Let 1,..,s denotes the set of states (also called the state space) of the Markov Chain (in Cereal Example, 1 = buying company K cereals, and 2 = buying competitions cereals, so s=2)
- Let $Y_0, Y_1, Y_2, \ldots$ are random variables, indicating the states at times $n = 0, 1, 2, ...$
- Let y_n be the observed value of the state at time n.

For example, we can observe the following chain of states: $1 \to 2 \to 2 \to 1 \to 1$ see the Table 4.1:

Time	0	1	2	3	4	5
State that has occurred	$Y_0 = 1$	$Y_1 = 2$	$Y_2 = 2$	$Y_3 = 1$	$Y_4 = 1$	$Y_5 =?$

TABLE 4.1: Cereals buyer Bob's Markov chain realisation. Note, that in this example the time steps are years.

Table 4.1 shows the realised states for one customer, say, Peter, where 1 means buying K, 2 means buying from a competitor, i.e. his chain is $1 \to 2 \to 2 \to 1 \to 1$, in years 0 to 4, while we do not know what he is going to buy in year 5. Another customer, let us call her Anita, can follow the same model (the probabilities in Figure 1) and her realised sequence can be $2 \to 2 \to 2 \to 2 \to 1$ in years 0 to 4. These two chains of values are two realisations of the Markov Chain, one for Peter and Anita, the chains are different but they still are following the same model which is depicted by Figure 4.1 and can be worded as there is an 88% probability to stay with K in next year, 12% chance to switch from K to Competition, 85% probability to stay with Competition, and 15% probability to switch from Competition to K.

Markov chain is a stochastic process because the values appear randomly according to some probability rules. Time series are also examples of stochastic processes. But time series (like the number of car sales, or yields of shares) are a series of values measured on a continuous scale, hence they do not live in a finite space.

In the Cereals buyers example, in the Markov chain the time is discrete (see Figure 4.2). Hence the Markov chain is called to be a discrete-time Markov chain. The time is indexed by the discrete listing n=0,1,2.... In Cereals buyers' example, the time's steps are years (but they can be anything e.g. months, days, every 6 months).

In Cereals buyers' example the state variables are discrete i.e. they take on a finite or countably infinite number of states (i.e. the number of elements in the state space is finite or countably infinite). This is why this Cereals buyer's example Markov Chain is called a finite-state Markov chain.

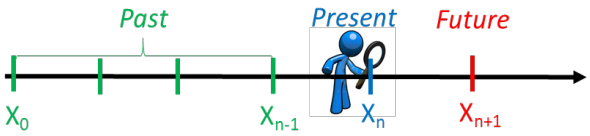


FIGURE 4.2: Illustration of the timeline in Markov Chain.

The Cereals buyers example is also time-homogeneous, in that the specified probabilities do not change over time (there is no n in Figure 1). One could imagine a different, more complicated model where the probabilities specified apply to young customers, but different probabilities apply once the customer gets older.

4.1.2 The Markov property

The Cereals buyers example has an important feature known as Markov property. It says that the value of the chain at time $n+1$ (i.e. the value y_{n+1}), only depends on the value of the chain at time n (i.e. on the value y_n). In other words, once the current state is specified, the path that brought the chain to that state is irrelevant. This has an important implication in how we estimate the value of the future state. Knowing the present state helps us to calculate the probabilities of the values of the future state, however, once we know the present state we won't get any additional information about the future by gathering the information about the past.

Consider, for example, the random walk of Cereals buyers example: if say for $n = 4$ we have $Y_4 = 1$ (brand K), then we know $Y_5 = 1$ or 2 with probability 0.88 or 0.12, respectively. It does not matter whether we speak about Peter or Anita, they have a different history of states, but they are both at state 1 at time n=4. Their probability distribution of the next state at time $n = 5$ is the same. This notion is formalised in the following definition.

Definition. Let sequence $Y_0, Y_1, Y_2, \ldots$ be a sequence of random variables (a chain) on some discrete state space (e.g. 1= customer buys from Company K, and 2= customer buys from Competitor, in the cereal example). This sequence is said to have the Markov property if, for any time index n and any set of states the following holds:

$$P(Y_{n+1} = s_{n+1}|Y_0 = s_0, Y_1 = s_1, \ldots, Y_n = s_n) = P(Y_{n+1} = s_{n+1}|Y_n = s_n) \tag{4.1}$$

The Markov property is also called the *memoryless property*. Because once we know the current state, that is all we need to find probabilities of the future state, we do not need to know the past states i.e. we can forget the past (Figure 4.3). This is not to say that the future states do not depend on the past. Indeed, the future depends on the past events. But, once we know the

current state, the future does not depend on the past. In other words, once we know the current state, knowing the past is not adding any new information about the future.

The probabilities $P(Y_{n+1} = s_{n+1}|Y_n = s_n)$ in Equation 4.1 are the one-step transition probabilities of the chain, or sometimes called just transition probabilities. These are the probabilities specified in the Cereals example (e.g. see Figure 1). It is critical to recognise that these are conditional probabilities: they specify the likelihood of the next member of the chain Y_{n+1} being in any particular state, given the current state of the chain Y_n.

Example: Cereals buyers. (continue) Let

- state 1 = customer buying K's cereal bran and
- state 2 = customer buying competition's cereal brand.

The sequence of successive cereal brands bought by customer Bob is characterised by four one-step transition probabilities. For example, it is stated that the customer "transitions" from Company K to the competitor with a probability 0.12, which means that for any time index, n,

$$P(Y_{n+1} = 2 \mid Y_{n=1}) = 0.12$$

This probability does not depend on the value of n, because the chain is time-homogeneous. Instead of writing $P(Y_{n+1}) = 2 \mid Y_n = 1) = 0.12$, we will sometime abbreviate with $P(1 \to 2) = 0.12$ to emphasize the idea of transitioning from one state to another. Thus, the complete set of one-step transition probabilities for a cereal customer is

$$P(1 \to 1) = 0.88, \quad P(1 \to 2) = 0.12$$
$$P(2 \to 1) = 0.15, \quad P(2 \to 2) = 0.85$$

The probabilities $P(Y_{n+1} = s_{n+1} \mid Y_0 = s_0, Y_1 = s_1, \ldots, Y_n = s_n)$ in Equation 4.1 are also conditional probabilities (we learned conditional probabilities in Chapter 2). The Equation 4.1 says, that the conditional distribution of the future states of the chain depends only upon the present state, not on the sequence of events that preceded it (i.e. not on past), as is also illustrated on Figure 4.3.

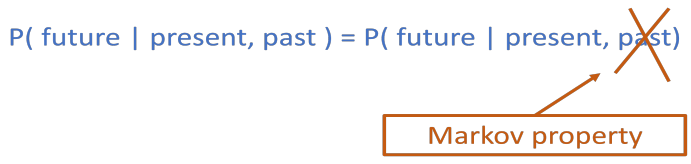


FIGURE 4.3: Markov property.

4.1.3 One-step transition probabilities

We introduced the notion of a Markov chain and its one-step transition probabilities. We can organise all transition probabilities into a matrix: the probability in i-th row and j-th column indicates the transition probability $P(i \to j)$:

Definition. Let $Y_0, Y_1, Y_2, Y_3 \ldots$ be a finite-state, time-homogeneous Markov chain, and index the states of the chain by the positive integers $1, 2, \ldots, s$. The (one-step) transition matrix of the Markov chain is the s×s matrix P whose (i, j)-th entry is given by

$$p_{i,j} = P(i \to j) = P(Y_{n+1} = j \mid Y_n = i)$$

for $i = 1, \ldots, s$ and $j = 1, \ldots, s$.

Example: Cereals buyers. (continue) We have two states, so $s = 2$. We called it State 1 when a customer buys from Company K, and State 2 is the customer busy from a Competition Manufacturer. The matrix P was 2×2 matrix:

$$\mathbf{P} = \begin{bmatrix} P(1 \to 1) & P(1 \to 2) \\ P(2 \to 1) & P(2 \to 2) \end{bmatrix} = \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix} =$$

So for example the entry of the matrix **P** in the second row and first column is

$$P(2 \to 1) = P(Y_{n+1} = 1 \mid Y_n = 2)$$

and it is equal to 0.15 in the Cereal Example. So if a customer is now buying a competitor cereal brand, then next time the customer will switch and buy from company K, with a probability of 0.15.

Note, that the matrix of one-step transition probabilities, **P**, must satisfy the following three conditions:

1. All the elements of P must lie between 0 and 1;
2. P must be a square matrix;
3. Each row of the matrix must sum to 1. This will always be the case: given that the chain is currently in some state i, it has to go somewhere in its next step (even if that entails remaining in state i). That is, for any state i and any time index n, we must have

$$\sum_{j=1}^{s} p_{i,j} = \sum_{j=1}^{s} P(i \to j) = \sum_{j=1}^{s} P(Y_{n+1} = j \mid Y_n = i) = 1$$

4.1.4 Initial and one-step probability distributions

Thus far, every probability we have considered has been conditional. For example, the entries of any one-step transition matrix indicate $P(Y_(n+1) = j|Y_n = i)$. In this section, we briefly explore unconditional probabilities which result from specifying a distribution for the random variable Y_0, the initial state of the chain. We will consider one case: modelling the initial state Y_0 as a random variable. Another case is to model the initial state as fixed, but we will not do it here.

Example: Cereals buyers. (continue) We will assume that we decided to call January 2000 as the initial year. On January 2000 we did a survey and found that 25% of customers are buying cereal K and 75% buy from competitors. That is, we have assigned the following initial distribution to the Markov chain:

State i	1 (K brand)	2 (competition brand)
$P(Y_0 = i)$	0.25	0.75

TABLE 4.2: Initial state distribution for Bob from Cereals buyers example.

This also means, that if we randomly chose a participant, then there is a chance of 0.25 that he/she is buying K and a chance of 0.75 that he/she is buying cereal from the competition. We will use the notation:

$$p^{(0)} = [0.250.75].$$

Unlike the conditional probabilities that comprise the transition matrix of the Markov chain, this initial distribution specifies the unconditional (aka marginal) distribution for the random variable Y_0. In what follows, we will sometimes refer to the bottom row of the table above or to p(0) as the "initial probability vector" "starting probability vector" "initial distribution" or "probabilities of the initial state".

What do we need to define a Markov chain? Two mathematical structures are sufficient to determine a Markov Chain fully:

1. The one-step transition matrix, **P**, of one-step transition probabilities,
2. The row vector, $p^{(0)}$, of probabilities of the initial state.

Now that we know the probabilities of the initial state, can we calculate the probability of future states? Now consider the random variable Y_1, the brand of cereal bought by the customer in year 1 (i.e. exactly 1 year after the initial date). What is the probability that he buys K after one time step, hence after 1 year? This can be determined by the law of total probability (see Chapter 2).

$$P(Y_1 = 1) = P(Y_1 = 1 \mid Y_0 = 1)P(Y_0 = 1) + P(Y_1 = 1 \mid Y_0 = 2)P(Y_2 = 2)$$

$$= (0.88)(0.25) + (0.15)(0.75) = 0.3325$$

The foregoing computation is the product of the initial probability vector with the second column of the matrix P. Both probabilities must add up to one (so we could have just calculated $P(Y_1 = 2) = 1 - 0.3325 = 0.6675$. So altogether, the unconditional probability mass function, $p^{(1)}$, for the random variable Y_1 (what the customer is buying one time step after the initial time) is:

State i	1 (K brand)	2 (competition brand)
$P(Y_1 = i)$	0.3325	0.6675

TABLE 4.3: Initial state distribution for Bob from Cereals buyers example.

An efficient way to determine the distribution of Y_1 is to compute all (here two) products simultaneously through matrix multiplication. If we multiply the transition matrix $\mathbf{P}$ on the left by a 1x2 row vector containing the initial probabilities for Y_0, we obtain

$$p^{(1)} = p^{(0)}\mathbf{P} \tag{4.2}$$

which is

$$p^{(1)} = p^{(0)}\mathbf{P} = [0.25, 0.75] \times \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix} = [0.3325, 0.6675]$$

Interpretation:

- So after one year (n=1), for a randomly chosen person there is a 33.25% chance that he is buying K and 66.75% competitor cereals. This is equivalent to the following:
- So after one year (n=1), 33.25% of the people are in state 1 - that is, buying K's cereal. So the expected market share of cereal K is 33.25%, after 1 year.
- This result makes intuitive sense, e.g. of the 25% currently buying K's cereal 88% continue to do so next year, while of the 75% buying the competitor's cereal, 15% change to buy K's cereal - giving a (fractional) total of $(0.25 \times 0.88) + (0.75 \times 0.15) = 0.3325$ buying K's cereal.

Note 1: We should always check that the total of $p^{(0)}$ adds up to 1, the total of $p^{(1)}$ adds up to 1, etc.

Note 2: When multiplying the vector $p^{(0)}$ with the matrix $\mathbf{P}$ the order of calculation is important. First, we write the vector and then the matrix (i.e., $p^{(0)}\mathbf{P}$) because:

$$p^{(0)}\mathbf{P} \neq \mathbf{P}p^{(0)},$$

$$p^{(1)}\mathbf{P} \neq \mathbf{P}p^{(1)}.$$

Note 3: We use small and thin letter p for a row vector of probabilities, and capital bold **P** for the transition matrix.

The method illustrated in the Cereal example can be generalised to find the unconditional distribution of the states Y_n in the chain after any number of transitions n, starting with a specified initial distribution for Y_0.

Theorem: Let $Y_0, Y_1, Y_2, Y_3 \ldots$ be a Markov chain with state space $1, \ldots, s$ and one-step transition matrix **P**. Let $p^{(0)}$ be a $1 \times s$ vector specifying the initial (starting) distribution of the chain, i.e. $p^{(0)} = (P(Y_0 = 1), \ldots, P(Y_s = 1))$. Then

$$p^{(1)} = p^{(0)}\mathbf{P}$$

More generally, if $p^{(n)}$ denotes the $1 \times s$ vector of marginal probabilities for Y_n

$$p^{(n)} = p^{(0)}\mathbf{P}^n$$

where $p^{(n)}$ is the probability vector representing the probability of each state at time n i.e. after n steps, and $\mathbf{P}^n$ is the one-step transition matrix **P** to the power of n.

Proof. The formula $p^{(n)} = p^{(0)}\mathbf{P}^n$ can be derived using the same mathematical approach displayed in Example Cereal for $p^{(1)}$. Now consider $p^{(2)}$, the vector of unconditional probabilities for Y_2. By the same reasoning, we have

$$p^{(2)} = p^{(1)}\mathbf{P}$$

The substitution $p^{(1)} = p^{(0)}\mathbf{P}$ then yields

$$p^{(2)} = p^{(0)}\mathbf{P}\mathbf{P} = p(0)\mathbf{P}^2$$

Analogically, we have for general n that

$$p^{(n)} = p^{(n-1)}\mathbf{P} = (p^{(0)}\mathbf{P}^{n-1})P = p^{(0)}\mathbf{P}^n,$$

as claimed. Hence the above is true for any n, by induction principle.

Example: Cereals buyers. (continue) Next, we want to know what is the state of the system in year 2. To answer such a question, we need to calculate the 2x1 vector of marginal probabilities: the probability that a customer will be in state 1 (buying K) after two years, and the probability of a customer being in state 2 (buying competitor brand) after two years. In two years from now (i.e. n=2), the state of the system is given by:

$$\begin{aligned}
p^{(2)} &= p^{(0)}\mathbf{P}^2 \\
&= [0.25, 0.75] \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix}^2 \\
&= [0.25, 0.75] \begin{bmatrix} 0.7924 & 0.2076 \\ 0.2595 & 0.7405 \end{bmatrix} \\
&= [0.3927, 0.6073]
\end{aligned}$$

Alternatively,

$$\begin{aligned} p^{(2)} &= p^{(1)}\mathbf{P} \\ &= [0.3325, 0.6675] \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix} \\ &= [0.3927, 0.6073] \end{aligned}$$

Interpretation: So two years from now, 39.27% of the people will be buying K's cereal, and 60.73% will be buying from the competitor. So if we randomly choose a customer, and if we do not know what was her/his initial state, then such a customer will be buying K's cereal two years from now with a probability of 39.27%, or competition cereals with a probability of 60.73%.

Note that $p^{(n)}$ is the probability vector of being in states $1, 2, \ldots s$ at time step n. These are marginal probabilities, as they give the probability of being in state i at n steps, i.e. they do not condition on anything. And $p_{(0)}$ is the vector of probabilities of initial states, i.e. at time 0.

4.1.5 Multistep transition probabilities

We now turn to the determination of multistep transition probabilities. Given that a Markov chain is currently in state i, what is the probability it will be in state j two steps later (i.e., after two transitions)? Three steps later? We begin by introducing the definition.

Definition. Let $Y_0, Y_1, Y_2, Y_3 \ldots$ be a time-homogeneous Markov chain. For any positive integer k, the k-th step transition probabilities are defined by

$$p_{i,j}^{(k)} = P^{(k)}(i \to j) = P(Y_{n+k} = j \mid Y_n = i), \tag{4.3}$$

where i and j range across all the states of the chain (typically $1, \ldots, s$. When $k > 1$ then we call these probabilities the **multistep transition probabilities**. For $k = 1$, i.e., one-step transition, we will typically revert to the previous notation: $P^{(1)}(i \to j) = P(i \to j)$. For $k = 1, 2, dots$ the $p_{i,j}^{(k)}$ are conditional probabilities because they give the probability of getting into state j given that we were in state i exactly k steps ago. The superscript (k) above in expression 4.3 does not indicate taking the k-th power; it indicates the state of the system (the customer or the market) in k steps. The matrix containing $P^{(k)}(i \to j)$ is called the **multistep transition matrix**, describing the transitions over k steps.

We have already seen how to calculate the probability of future states (see Theorem 4.1.4). We used matrix $\mathbf{P}^n$ to transition from an initial state into future states in n steps. Since the Markov chain is homogeneous, the matrix should allow us the transition from any time point to any future time point,

in other words, such a matrix is the matrix of transition probabilities that we are looking for.

Chapman-Kolmogorov equations. If a Markov chain has a one-step transition matrix $\mathbf{P}$, then the k-th step transition matrix of probabilities are the entries of the matrix $\mathbf{P}^k$, i.e. k-th power of matrix $\mathbf{P}$. In other words,

$$p^{(k)}(i \to j) = p_{i,j}^{(k)} \text{ is the (i,j)-th entry of matrix } \mathbf{P}^k$$

Example: Cereals buyers. (continue) Find the 2-step transition matrix, and use the theorem above to find the 2-step probability distribution. Suppose that the customer just bought cereals from the K brand. What is the probability that two years from now he will again be buying the K brand, i.e. $P^{(2)}(1 \to 1) = P(Y_{n+2} = 1 \mid Y_n = 1)$? To answer the question, we realise that the one-step transition matrix is

$$\mathbf{P} = \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix}$$

Hence the two-step-transition matrix is

$$\mathbf{P}^2 = \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix} \times \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix} = \begin{bmatrix} 0.7924 & 0.2076 \\ 0.2595 & 0.7405 \end{bmatrix}$$

Hence, two years after buying the K brand, a customer will be again buying the K brand with a probability of 79.24%. Similarly two years after buying a competitor brand, the customer will be again buying the competitor brand with a probability of 74.05%.

4.1.6 Long-run prediction of the state

When we first introduced the Cereal Example we also asked what happens in the long run, i.e. after a sufficiently long time, we want to know which states are likely to occur and with what probabilities. In other words, we are asked to find the long-term prediction of the states when n goes to infinity. This means we need to find the long-run probabilities of the states. This assumes that, eventually, the system will reach a long-run distribution in the sense that the state of the system at time n is equal to the state of the system at time $n-1$, i.e. the vectors of probabilities are not changing any more:

$$p^{(n)} = p^{(n-1)}$$

This does not mean that transitions between states no longer take place, they do, but they balance out so that the number in each state remains the same. Such long-run distribution is also called long-term distribution. It is also said

that, for large n, such a system is then in a *stationary state* as it is not changing any more. Some also call it the equilibrium state or steady state.

There are two basic approaches to calculating the long-term probability distribution:

1. Computationally: by calculating $p^{(n)}$ for $n = 1, 2, 3 \ldots$ and stop when $p^{(n-1)}$ and $p^{(n)}$ are approximately the same. This is easy for a computer to do but can be lengthy to do by hand.

2. Algebraically: to avoid the lengthy arithmetic calculations needed to write the vector $p^{(n)}$ algebraically for $n = 1, 2, 3 \ldots$, we use an algebraic shortcut. This is what we will do next.

   An algebraic solution to find the long-run prediction can be found by easy:

- In the long run the row-vectors of probabilities are the same:

$$p^{(n)} = p^{(n-1)}$$

- The left hand side (i.e. the $p^{(n)}$) can be written as $p^{(n)} = p^{(n-1)}\mathbf{P}$, while the right hand side is $p^{(n-1)}$, so due the equality we have

$$p^{(n-1)}\mathbf{P} = p^{(n-1)}.$$

- We will use this last equality to find the stationary vector $p^{(n-1)}$ of the probabilities.

Example: Cereals buyers. (continue) We next find the long-term behaviour of the Cereal Markov Chain example, via the algebraic approach. The vector $p^{(n)}$ has two elements because there are only two states in our example Cereal Manufacturer. Let $p^{(n-1)} = [Y_1, Y_2]$ then we need to find the two values Y_1 and Y_2 such that:

$$[Y_1, Y_2] = [Y_1, Y_2] \times \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix}$$

and such that $Y_1 + Y_2 = 1$. This means that we have to solve three equations of two unknowns:

$$\begin{aligned} Y_1 &= 0.88Y_1 + 0.15Y_2 \\ Y_2 &= 0.12Y_1 + 0.85Y_2 \\ 1 &= Y_1 + Y_2 \end{aligned}$$

Since $2 < 3$, it means that there is some redundancy in the three equations. In fact, it can be shown that the system of the above three equations is equivalent to (i.e. has the same set of solutions, (Y_1, Y_2): $0.12Y_1 - 0.15Y_2 = 0$

and $Y_1 + Y_2 = 1$. The equation $Y_1 + Y_2 = 1$ is essential, as without it we could not obtain a unique solution for Y_1 and Y_2. Solving the system of equations, we obtain $Y_1 = 0.5556$ and $Y_2 = 0.4444$. This means that in the long run, K's market share will be 55.56% and competitor's 44.44%. There will be customers who will be switching from one brand to another, but the split of the customers will not be changing. This can only happen, if the number of customers who switch from K to competition, is the same as the number of customers who switch from competition to K.

Comment 1: A useful numerical check (particularly for larger systems of equations) is to substitute the final calculated values back into the original equations to check that they are consistent with those equations.

Comment 2: Using a computational approach, $n = 12$ iterations were needed before the vectors $p(n)$ and $p(n-1)$ were approximately the same (see Section 1).

Comment 3: We were making an assumption that such a long-run state of the system exists. For example, the system with the following transition matrix and the initial state does not have a long-run probability distribution

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$
$$p^{(0)} \neq [0.5, 0.5]$$

Do you see why the system above does not have a long-run distribution? What if the initial vector was $p^{(0)} = [0.5, 0.5]$?

Comment 4: How do we come up with reasonable estimates of the transition probability matrix, P? The data needed to deduce transition probabilities can be easily gathered. As an example, such data used to be gathered for consumer brand switching by surveying customers individually. However, many supermarkets in the UK now have their own "loyalty cards" which are swiped through the checkout at the same time as a customer makes their purchases. These provide a mass of detailed information as well as other information such as the effect of promotional campaigns. Consider the supermarkets that are gathering a mass of data from which they can deduce brand-switching transition matrices. Do you think those matrices might be of interest (value) to other companies or not?

Comment 5: Should the transition probabilities be constant? You should note that transition probabilities ought not to be considered as fixed numbers, they are, in fact, numbers that we can influence or change.

Comment 6: Our Cereal example was very simple. The competition was represented by one state (so there were two states in total: K and competitor).

With more detailed data that state could be changed into a number of different states - maybe one for each competitor brand of cereal. We could also have different models for different segments of the market - maybe brand switching is different in rural areas to that in urban areas for example. Families with children would constitute another important segment of the market.

4.2 Further topics on Markov Chains

In this section, we continue with Markov Chains by bringing more complex examples.

4.2.1 Commuter Cyril example

In this section, we bring more examples, introduce regular Markov chains, discuss the interpretation of the long-run distribution, and introduce irregular and aperiodic Markov chains. We will also discuss one example of a Markov chain that does not live in time, but rather in space (e.g., a land).

Example: Cyril going to work. A man either drives his car or catches a train to work each day. Suppose he never goes by train two days in a row; but if he drives to work, then the next day he is just as likely to drive again as he is to travel by train. Before his first day of work he needed to go to his new workplace to bring all his paperwork, for that the man tossed a fair die and drove to the workplace if a 6 appeared. The state space of the system (i.e. the list of all the states) is $train, drive$ or simply t, d. Questions:

1. Now, the probability that the system changes from state t to state d in exactly 2 steps?
2. What is the probability that the system changes from state t to state d in exactly 4 steps?
3. What is the probability that on the fourth working day, the man is driving to work?

This is another example of a Markov chain since the outcome on any day depends only on what happened the previous day. This can be summarised in Table 4.4 and visualised on a one-step transition diagram in Figure 4.4.

The one-step transition matrix of this Markov chain is:

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

The first row of the matrix corresponds to the fact that he never goes by train

		Next flight by BA	Next flight by Competition
Last flight by	BA	0.85	0.15
	Competition	0.10	0.90

TABLE 4.4: Commuter Cyril's one-step transition probabilities for choosing the type of transportation from home to his work.

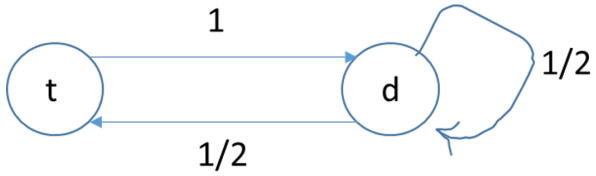


FIGURE 4.4: Commuter Cyril's one-step transition diagram for choosing the type of transportation from home to his work.

two days in a row, and so he will certainly drive the day after he travels by train. The second row of the matrix corresponds to the fact that the day after he drove to work, there is an equal probability that he will drive or go by train. Now, the probability that the system changes from, state t to state d in exactly 2 steps is calculated by using a second power of matrix $\mathbf{P}$, i.e. using matrix $\mathbf{P}^2$ (where the superscript is the power of 2): Hence the two-step-transition matrix is

$$\mathbf{P}^2 = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \times \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}$$

So the answer to the first question is that the probability that the system (the man who needs to travel to his work) changes from state t to state d in exactly 2 steps is 0.5.

Next, to answer the second question on the probability that the system changes from state t to state d in exactly 4 steps we need to calculate the fourth power of matrix P:

$$\mathbf{P}^4 = \mathbf{P}^2 \times \mathbf{P}^2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} = \begin{bmatrix} \frac{3}{8} & \frac{5}{8} \\ \frac{5}{16} & \frac{11}{16} \end{bmatrix}$$

So we have the four 4-steps transitions probabilities:

$$P^{(4)}(t \to t) = \frac{3}{8}, \quad P^{(4)}(t \to d) = \frac{5}{8}$$
$$P^{(4)}(d \to t) = \frac{5}{16}, \quad P^{(4)}(d \to d) = \frac{11}{16}$$

So the answer to the second question is that the probability that the system changes from state t to state d in exactly 4 steps is $\frac{5}{8}$.

To answer the third question (the probability that on the fourth day, the man is driving a car to work) we need the initial distribution. We were told, there was a day zero, when the man had to go to work to bring his paperwork. We were also told that the man tossed a fair die and drove to work if a 6 appeared, so he can bring his paperwork. This means that his initial probability distribution is $P^{(0)} = [\frac{5}{6}, \frac{1}{6}]$ (remember we always write the train probability first, then the driving a car). Then, after 4 time steps (here work days) the multi-step probabilities of each state are (from the Chapman-Kolmogorov equations):

$$P^{(4)} = P^{(2)}\mathbf{P}^4 = \begin{bmatrix} 5 \\ \overline{6}, \frac{1}{6} \end{bmatrix} \begin{bmatrix} \frac{3}{8} & \frac{5}{8} \\ \frac{5}{16} & \frac{11}{16} \end{bmatrix} = \begin{bmatrix} 35 \\ \overline{96}, \frac{61}{96} \end{bmatrix}$$

gives the probability of states on day four, i.e.

$$P_t^{(4)} = \frac{35}{96}, P_d^{(4)} = \frac{61}{96}$$

so, with probability $\frac{35}{96}$ he is taking a train, and with probability $\frac{61}{96}$ he is driving, on day four of his work.

4.2.2 Regular Markov chains

A finite-state Markov Chain with one-step transition matrix $\mathbf{P}$ is said to be a regular chain if there exists a positive integer n such that all of the entries of the matrix $\mathbf{P}^n$ are positive (i.e. non-zeros). In other words, for a regular Markov chain, there is some positive integer n such that every state can be reached from every state (including itself) in exactly n steps.

It is straightforward to show that if all entries $\mathbf{P}^n$ are positive, then so are all the entries of $\mathbf{P}^{(n+1)}$ and $\mathbf{P}^{(n+2)}$, and so on. Our Cereal Example is a regular chain, and since $\mathbf{P}$ has all entries positive (non-zero), and our Example Commuter Cyril is also a regular chain, as $\mathbf{P}^2$ has all entries positive (see previous section).

4.2.3 Steady-state theorem

What is so special about regular Markov chains? The transition matrices of regular Markov chains exhibit a rather interesting property.

The Steady-state theorem Let P be the one-step transition matrix of a finite-state regular Markov Chain. Then the following limit

$$\Pi = [\pi_1, \pi_2, \ldots, \pi_s] = \lim_{n \to \infty} \mathbf{P}^n$$

exists. Moreover, the rows of the limiting matrix Π are all identical, with all

positive entries. In other words, if the regular Markov Chain runs for a long time the $\mathbf{P}^n$ will stop changing, and it will be equal to a matrix with identical rows, and all elements are positive (non-zero). So in other words, in the long-run, every regular Markov chain reaches a steady state. We demonstrate the idea in the next example.

Example: Cereals buyers. (continue) In previous sections, we introduced Cereal Example. Its one-step transition matrix $\mathbf{P}$ is regular, as it has all elements positive (non-zero). Hence by Steady-state theorem the matrix $\mathbf{P}^n$ is converging to a limiting matrix, moreover, all rows of the limiting matrix are identical, with all positive entries. We can see this by calculating the powers in R (see Section 1):

$$P^1 = \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix}$$

$$P^2 = \begin{bmatrix} 0.7924 & 0.2076 \\ 0.2595 & 0.7405 \end{bmatrix}$$

$$P^{20} = \begin{bmatrix} 0.5563764 & 0.4436236 \\ 0.5545295 & 0.4454705 \end{bmatrix}$$

$$P^{40} = \begin{bmatrix} 0.5555571 & 0.4444429 \\ 0.5555537 & 0.4444463 \end{bmatrix}$$

$$P^{60} = \begin{bmatrix} 0.5555556 & 0.4444444 \\ 0.5555556 & 0.4444444 \end{bmatrix}$$

$$P^{200} = \begin{bmatrix} 0.5555556 & 0.4444444 \\ 0.5555556 & 0.4444444 \end{bmatrix}$$

We see that every row of the matrix $\mathbf{P}^{60}$ is identical to seven decimal places. The same holds for matrix $\mathbf{P}^{200}$. If we try even higher power, we will get the same matrix again and again. This example illustrates the central theorem of regular Markov chains: the Steady-state theorem, which says that for regular Markov chains the powers of transition matrix reach a limiting matrix which we will call as Π i.e., the capital Greek letter Pi. In Cereal example:

$$\Pi = \begin{bmatrix} 0.5555556 & 0.4444444 \\ 0.5555556 & 0.4444444 \end{bmatrix}$$

4.2.4 Interpretation of the long run distribution

If we let the vector $\Pi = [\pi_1, \pi_2, \ldots, \pi_s]$ denote each of the identical rows of the limiting matrix Π of a regular Markov chain, then the vector Π is called the long-run probability distribution of the Markov chain (also called stationary distribution, steady-state distribution).

Thus for the Cereals buyers example, the long-run distribution is $\Pi = [0.5556, 0.4444]$ (with 4 decimal places). The steady-state distribution can be

interpreted in several ways. We present three here (for further discussion see the recommended book by Carlton and Devore [13], p450):

1. If the "current" state of the Markov chain is observed after a large number of transitions, there is an approximate probability π_j of the chain being in the state j; that is for large n, $P(Y_n = j) \approx \pi_j$. Moreover, this holds regardless of the initial distribution of the chain (i.e. the unconditional distribution of the initial state Y_0).

   (a) The first sentence above is essentially the definition of the Π and it follows from the Steady-state theorem.

   (b) The second sentence above says that the effect of the initial state or the initial probability distribution of the process wears off as the number of steps of the process increase.

   (c) In the Cereal example, if we let the chain run for some amount of time, and then we randomly choose a customer, there is a 55.56% chance that she is buying K at that current time.

2. The long-term (i.e. long-run) proportion of time the Markov chain visits the j-th state is π_j. So, in the Cereal example, if we let the chain run for some amount of time, then after that this is what we can say about each customer: a customer spends about 55.56% of time buying brand K and 44.44% of time buying from competitor.

3. If we assign Π to be the initial distribution of Y_0 then the distribution of Y_n is also Π for any number of subsequent transitions n. For this reason, Π is customarily referred to as the stationary distribution of the Markov chain. Also, Π is sometimes called the fixed probability vector transition matrix $\mathbf{P}$.

Example: Cereals buyers. (continues) If the initial distribution is $p^{(0)} = [0.25, 0.75]$, then after 20 steps the market share is: $p^{(0)}\mathbf{P}^2 0 = [0.5556, 0.4444]$. We have seen, that from n=20, the $\mathbf{P}^n$ is not changing much. Also, we found that in the long term the proportion of the market is (0.5555556 0.4444444).

What if the initial market share is $[0.99, 0.01]$? What will be the market share after 20 steps (years)? What if the initial distribution is $[0.01, 0.99]$? What will be the market share after 20 years? The effect of the initial state wears off. We also saw this previous section with the Cereal Example. Now we will do it again with the Example of Commuter Cyril.

Caution! It is not possible to make long-range predictions with all transition matrices. However, for a regular transition matrix, it is always possible to make long-range predictions. In other words, a Markov process that has a regular transition matrix will have a steady state. Also, a Markov chain does not have to be regular for the limit of $\mathbf{P}^n$ to exist. (one example is in the book by Carlton and Devore, p449).

Example: Commuter Cyril. (continue) What is the long-term prediction for this Markov chain? We showed above that this is a regular Markov chain, so it will reach a steady state or equilibrium, let us call it: $\Pi = [\pi_1, \pi_2]$. We will use that $\pi_1 + \pi_2 = 1$, so we can simply reparameterise $\Pi = [x, 1-x]$ i.e. $x = \pi_1$ and $1 - x = \pi_2$. Furthermore, thanks to Theorem 4.1.4 the vector Π must satisfy:

$$[x, 1-x] \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = [x, 1-x]$$

which is a system of two equations of one unknown, in a vector and matrix notation. Multiplying the vector on the left-hand-side with the matrix, and then equating it to the vector on the right, we get:

$$\frac{1}{2}(1-x) = x$$
$$1 + \frac{1}{2}(1-x) = 1 - x$$

which is true if and only if

$$x = \frac{1}{3}$$

hence

$$\Pi = \left[\frac{1}{3}, \frac{2}{3}\right]$$

Thus, in the long run, the man will take the train to work one-third of the time and drive to work two-thirds of the time. Since there is a unique solution, we can say that the system will reach a steady state or equilibrium, if the transition probabilities do not change. The stationary distribution of this long-term prediction is this: $[\frac{1}{3}, \frac{2}{3}]$ i.e., with probability $\frac{1}{3}$ he will take train, and with probability $\frac{2}{3}$ he will drive his car.

Next, we calculate the powers of the matrix P in Example Commuter Cyril and estimate the limiting matrix. To answer this question we need to check if there exists a positive integer n such that all of the entries of the (power) matrix $\mathbf{P}^n$? Yes, this Markov chain is regular, because for n=2, the second power of P has all entries positive (non-zero):

$$P^2 = \begin{bmatrix} 0.5 & 0.5 \\ 0.25 & 0.75 \end{bmatrix}$$

And since the matrix P is regular then the steady-state exists according to State-State Theorem. Next, we can calculate the 10th and 20th power

$$P^{10} = \begin{bmatrix} 0.3339844 & 0.6660156 \\ 0.3330078 & 0.6669922 \end{bmatrix}$$

$$P^{20} = \begin{bmatrix} 0.333334 & 0.6666666 \\ 0.333333 & 0.6666667 \end{bmatrix}$$

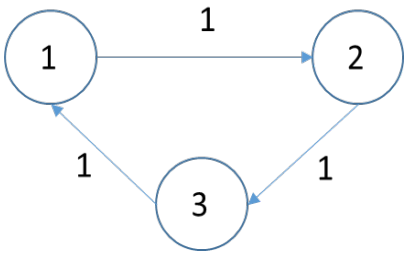


FIGURE 4.5: Bus driver Lubos's one-step transition diagram.

We, therefore, estimate that the limiting matrix must be:

$$\Pi = \begin{bmatrix} 0.333333 & 0.6666667 \\ 0.333333 & 0.6666667 \end{bmatrix}$$

4.2.5 Irreducible Markov Chains

The existence of a stationary distribution is not unique to regular Markov chains.

Definition. Let i and j be two (not necessarily distinct) states of a Markov chain. State j is accessible from state i (or, equivalently, i can access j) if $P^{(n)}(i \to j) > 0$ for some integer $n \geq 0$.

For $n = 0$, the symbol $P^{(0)}(i \to j) > 0$ is interpreted as the probability of going from i to j in zero steps, an so necessarily $P^{(0)}(i \to i) = 1$ for all i and $P^{(0)}(i \to j) = 0$ for $i \neq j$. In particular, this means that every state i is, by definition, accessible from itself.

Definition. A Markov chain is irreducible if every state is accessible from every other state in a finite number of steps.

It should be clear that every regular chain is irreducible (do you see why?). However, the reverse is not true: an irreducible Markov chain need not be a regular chain. See the next Example Bus driver.

Example: Bus driver Lubos. Consider the following Markov Chain. A bus driver follows his bus route from campus (state 1), to the nearby student housing complex (state 2), do downtown (state 3), and then back to campus. The associated Markov chain cycles endlessly. Assuming it starts at state 1, the chain is: $1 \to 2 \to 3 \to 1 \to 2 \to 3 \to 1 \ldots$. The one-step transition diagram is in Figure 4.5.

The one-step transition matrix is:

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

The powers of the matrix P are

$$\mathbf{P}^2 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \mathbf{P}^3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I, \mathbf{P}^4 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, ...$$

Where I is an identity 3×3 matrix. So we have that

$$\mathbf{P}^4 = \mathbf{P}^3\mathbf{P} = I\mathbf{P} = \mathbf{P}$$
$$\mathbf{P}^5 = \mathbf{P}^4\mathbf{P} = \mathbf{P}\mathbf{P} = \mathbf{P}^2$$
$$\mathbf{P}^6 = \mathbf{P}^5\mathbf{P} = \mathbf{P}^2\mathbf{P} = I$$
$$\mathbf{P}^7 = \mathbf{P}^6\mathbf{P} = I\mathbf{P} = \mathbf{P}$$
$$\mathbf{P}^8 = \mathbf{P}^7\mathbf{P} = \mathbf{P}\mathbf{P} = \mathbf{P}^2$$

and so on, the $\mathbf{P}^n$ is equal to one of $\mathbf{P}$, $\mathbf{P}^2$ and I for every positive integer n, and all three of these matrices contain some zero entries. Therefore, this is not a regular Markov Chain. Lubos can access any of the three locations in visits from any other location, so the chain is irreducible.

Question: What is its long-term prediction for the bus driver Lubos? Does it achieve a steady state? And if yes, what is the stationary distribution of this steady state?

Answer: Since it is not a regular Markov Chain we cannot apply the Steady-State theorem i.e. the theorem does not apply here and hence is not telling us if this has steady-state. So, we need another approach. We will do a direct proof. We will assume that such stationary distribution exists, we will call it $\pi = (x, y, z)$ and we will find it. (If we cannot find it then it does not exist). In order to find it, we will use the fact that a stationary distribution has to satisfy

$$[x, y, z] \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = [x, y, z] \tag{4.4}$$

and $x + y + z = 1$.

From 4.6 it follows that $x = z$, $y = x$, $z = y$. By applying the condition $x + y + z = 1$, we get that

$$[x, y, z] = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right] \tag{4.5}$$

is the stationary distribution of the steady-state. So in the long term, the chain achieves a steady state.

So, the interpretation of the above is that if the bus driver Lubos is equally likely to be at any of its three locations right now, it is also equally likely to be at any of those three places after the next transition. This is the stationary distribution of states for Lubos the bus driver even though the chain is not

regular. This means there are chains that are not regular, and yet achieve a steady–state in the long run (in the long term). In the next subsection we will look into them closely.

4.2.6 Periodic and aperiodic Markov chains

Definition. The period of a state i is defined as the greatest common divisor (gcd) of all positive integers n such that $P^{(n)}(i \to i) > 0$; if that gcd equals 1, then state i is called aperiodic.

Example: Bus driver Lubos. (continue) All three states have period 3, because for every state the period is gcd of $3, 6, 9, \ldots$ hence it is 3. Hence, all states are periodic with period 3, i.e. they are not aperiodic.

It can be shown that every state in an irreducible chain has the same period. A Markov chain is said to be aperiodic if that common period is 1 and is called periodic otherwise. In other words, irreducible Markov is called aperiodic if all states are aperiodic.

Why is periodicity important? It plays a role when we discuss long-term (i.e. limiting) distributions.

Theorem: A finite-state Markov chain is regular if, and only if, it is both irreducible and aperiodic.

How, do we check the periodicity of an irreducible Markov Chain? Consider a finite irreducible Markov chain. If there is a self-transformation in the chain ($P(i \to i) > 0$ for some i), then the chain is aperiodic.

Note: The bus driver Lubos's example is a case of an irreducible Markov chain that is not regular, it is irreducible and periodic. Hence, we cannot use the Steady-State theorem to tell if there is a steady state; nevertheless, it has a steady state, as we calculated previously. On the other hand, Going to Work and Cereal examples are regular, and hence irreducible and aperiodic.

Example: Fatima investigating forest health. Markov chain models have been used to study the pattern of diseased and healthy trees in forests. In one such model, it was assumed that the forest could be divided into areas of two types:

- gaps, which contained only healthy trees,
- and patches, which contain both diseased and healthy trees.

Each tree could be classified as

- a diseased tree (state 0),
- as a healthy tree (state 1)
- or as a healthy gap tree (state 2).

Fatima is tasked to investigate the forest's health. She decided to walk in the forest, choosing her path randomly. Each tree along a randomly chosen path through the forest is one of these three types. Suppose that the types of

trees Fatima encounters on her path can be modelled by a Markov chain. The transition matrix can be given by:

$$\boldsymbol{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0.3 & 0.1 & 0.6 \\ 0.1 & 0.3 & 0.6 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \end{array}$$

Approximately what proportion of the trees Fatima encounters on a long path are healthy?

Solution: The question is the same as asking for the stationary distribution of the chain but here instead of time, our Markov chain lives in space i.e. across the forest space. Hence we need to solve:

$$[x, y, 1-x-y] \begin{bmatrix} 0.3 & 0.1 & 0.6 \\ 0.1 & 0.3 & 0.6 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} = [x, y, 1-x-y] \tag{4.6}$$

which leads to the following equations

$$\begin{aligned} 0.3x + 0.1y + 0.1(1-x-y) &= x \\ 0.1x + 0.3y + 0.1(1-x-y) &= y \\ 0.6x + 0.6y + 0.8(1-x-y) &= 1-x-y \end{aligned}$$

which leads to

$$\begin{aligned} 0.1 &= 0.8x \\ 0.1 &= 0.8y \\ 0.2 &= 0.8x + 0.8y \end{aligned}$$

Solving these equations we obtain

$$x = \frac{1}{8}, y = \frac{1}{8}, z = \frac{3}{4}$$

The healthy trees are trees in states 1 and 2, so the proportion of trees that are healthy is given by $\frac{1}{8} + \frac{3}{4} = \frac{7}{8}$. Hence, Fatima will encounter approximately $\frac{7}{8}$ proportion of healthy trees encountered on a long path.

4.3 Tips to think and act like a risk expert

Here we will give two give tips and tricks.

4.3.1 Not all sequences are Markov Chains but some can be turned into Markov Chains

Not all sequences of random variables possess the Markov property. In econometrics (statistical methodology applied to economics, for example, most mod-

els for the closing price X_{n+1} of stock on the $n+1$st day of trading incorporate not only the previous day's closing price X_n but also information from many previous days (the data $X_{n-1}, X_{n-2}, \dots$ and so on). The likelihood that X_{n+1} will be £5 higher than X_n may depend on the stock's behaviour over all of last week, not just where it closed on day n.

That said, in some instances a model that includes more than a one-time-step dependence can be modified by reconfiguring the state space in such a way that it satisfies the Markov property.

For example, in the Snowy days example (see below Exercises), we assume that one can model tomorrow's weather based on today's conditions without incorporating any previous information. A more realistic model might assume that tomorrow's snow depth depends on today's and yesterday's weather. Suppose, for example, that tomorrow will be a snow day with a probability of 0.8 if both yesterday and today were snow days; with a probability of 0.6 if today was snow and yesterday was a green day; with a probability 0.3 if it was green day today and snow yesterday; with probability 0.1 if both previous days were green. Once again, let X_n = "the state of the weather on day n: G for green, S for snowy". Then the sequence $\{X_0, X_1, X_2, X_3 \dots\}$ does not satisfy the Markov property, because the conditional distribution of X_{n+1} depends on both X_n and X_{n-1} (the weather on previous two day's weather condition).

We can turn the sequence into a new sequence which has Markov property, which is what we show next. So for such sequence $\{X_0, X_1, X_2, X_3 \dots\}$ we can try to make the following modification: Let us define a new chain of variables Y_n

$$Y_n = (day\, n\, weather,\, day\, n+1\, weather) = (X_n, X_{n+1})$$

i.e. each Y_n is defined as a vector of two random variables X_n and X_{n+1}. So, for example, consider days 4 and 5:

- If snow depth was $\geq 50mm$ on day 4 but $< 50mm$ on day 5, then we denote it as $Y_4 = (S, G)$.
- The weather on day 6 depends on these previous two days, but they are now both contained in a single "variable" Y_4.
- In other words, Y_5, can be modelled entirely by knowing Y_4: Y_5' s first entry, X_5, matches the second entry of Y_4, and the probability distribution of the second entry of Y_5 (i.e., X_6) is determined by the rules given at the beginning of this example.

Thus with this modification, the sequence $\{Y_0, Y_1, Y_2, Y_3, \dots\}$ forms a Markov Chain. The state space is not $\{S, G\}$ but rather $\{(S,S), (S,G), (G,S), (G,G)\}$. The earlier weather rules can be expressed as one-step transition probabilities for this chain:

$$P((S,S)) \to (S,S)) = 0.8$$
$$P((S,G)) \to (G,S)) = 0.3$$
$$P((G,S)) \to (S,S)) = 0.6$$
$$P((G,G)) \to (G,S)) = 0.1$$

Four other transition probabilities can be found by considering the complements of the given transition events:

$$P((S,S)) \to (S,G)) = 1 - 0.8 = 0.2$$
$$P((S,G)) \to (G,G)) = 1 - 0.3 = 0.7$$
$$P((G,S)) \to (S,G)) = 1 - 0.6 = 0.4$$
$$P((G,G)) \to (G,G)) = 1 - 0.1 = 0.9$$

The final eight transition probabilities (with four states, there are 4^2=16 total one-step transition probabilities) are all 0, so we have:

$$P((S,G)) \to (S,S)) = 0$$
$$P((G,S)) \to (G,S)) = 0$$
$$P((G,G)) \to (S,S)) = 0$$
$$P((S,S)) \to (G,S)) = 0$$
$$P((S,G)) \to (S,G)) = 0$$
$$P((G,S)) \to (G,G)) = 0$$
$$P((G,G)) \to (S,G)) = 0$$
$$P((S,S)) \to (G,G)) = 0$$

Finally, the one-step transition matrix of $\{Y_0, Y_1, Y_2, Y_3 \dots\}$ is in the following table

		Next weather			
		(S,S)	(S,G)	(G,S)	(G,G)
Last weather	(S,S)	0.8	0.2	0	0
	(S,G)	0	0	0.3	0.7
	(G,S)	0.6	0.4	0	0
	(G,G)	0	0	0.1	0.9

4.3.2 Sensitivity analysis

In the examples above, we assumed that we know the probabilities. For example, we assumed precise probabilities of transition from K to Competition in the Cereal buyer example. In real life, we do not know them precisely, but hopefully, we have some imprecise answers. For example, in the Cereals example, we may estimate that the transition probability from K to Competition is between 0.1 to 0.15. In other words, we are facing uncertainty, as we do

not have the precise probabilities to put into the transition matrix. How can we resolve this problem? A way to resolve it is by conducting as sensitivity analysis.

Sensitivity analysis is an important tool of risk analysis (not just Markov Chains). It aids in reducing uncertainty by identifying high-impact parameters (such as probabilities). This can help in finding out which data (information) to acquire to reduce uncertainty on said parameters.

In Markov Chains, it is crucial to do sensitivity analysis. It can help in finding out how the result (the multistep probabilities and stationary distribution) depends on the specification of the probabilities or on outcome values. For example, we can calculate the stationary distribution with 0.1 and again with 0.15 and compare the stationary distributions to how much they differ. If they differ a lot, then the stationary distribution is sensitive to how we specify the transition probability from K to Competition.

4.4 Summary

We learned in this chapter:

1. Markov Chain is a finite-state stochastic process; it has a Markov "memoryless" property. Markov chain is defined by two components: initial probability distribution and a one-step-transition matrix.

2. We learned how to calculate k-step transition $P^{(k)}(i \to j) = P(Y_{n+k} = j|Y_n = i)$ using iterations or k-step transition matrix P^k (Chapman Kolmogorov equation).

3. We learned how to find long-term prediction, i.e. how the chain behaves after many steps (when k is very large).

4. We saw that some Markov chains will reach stationary distribution after a sufficient number of steps. We also call it a steady-state, equilibrium or long-run distribution. We learned how to calculate the stationary distribution $\pi = (Y_1, \ldots, Y_s)$ by finding a unique solution to equations: $\pi P = \pi$ and $Y_1 + Y_2 + \ldots + Y_s = 1$. We learned how to find out which Markov chains will reach a stationary distribution.

5. We learned about several types of Markov chains: regular/irregular, reducible/irreducible and periodic/aperiodic. We learned how to recognise them (see definitions), and we learned some useful properties. For example, the steady-state theorem says that a regular Markov Chain will always reach a stationary distribution. Then we found a chain that was not regular and still reached a stationary distribution (see the Example Bus driver Lubos).

4.5 Further reading

The chapter mainly followed the notation and structure in the book "Probability with Applications in Engineering, Science, and Technology" by Matthew A. Carlton and Jay L. Devore [13] (see their Chapter 6). However, we added several examples and discussions related to risk and uncertainty. Further recommended resources are:

1. For a dedicated R package we recommend [54].

2. Markov Chains are extremely popular and powerful with many applications. One of the most famous applications of Markov Chains is the Page rank algorithm to rank web pages. It was developed in the late 1920s by Sergey Brin and Larry Page, then graduates at Stanford University, when they worked on their project to organise the World Wide Web's information. In 1998 they published their work and founded Google. We invite the reader to google the history of the algorithm as well as how it was applied for page ranking.

4.6 R lab

We first show how to use R to solve problems using Markov Chains. First, several questions are provided with a solution. Then we give more questions to you to work on but without a solution.

1. **[Purpose: Getting experience of using R for simulation of Markov Chains, specifically to find the effect of initial distribution on limiting distribution from Section 4.1.]** Assume initial state: 0.25 and 0.75 for K and Competition brand of cereals, respectively. Next, we will use R to find the market share after n steps ($n = 1, \ldots, 100$). This is to find out the long-term prediction for the market share for each of the two brands of cereal. We will not simulate the Markov Chain; rather, we will evaluate the transition n times to get the multi-step probabilities distributions.

```
> states <- c(1,2) # 1 for K company, 2 for competition
> p_0<-c(0.25,0.75) # initial probability distribution
> P = matrix(c(0.88,0.12,0.15,0.85),nrow=2,ncol=2,byrow=TRUE)
# Next we initialise the multi-step probability distributions
    for states 1 and 2 at steps 1 to 100
> P_m1 <- rep(NA,100)
> P_m2 <- rep(NA,100)
# Next, the step 1 is to be done manually; this is an
    initialisation
```

```
> next_step <- NA
> next_step <- p_0 %*% P
> P_m1[1] <- next_step[1] # the marginal probability at time n
    -1
> P_m2[1] <- next_step[2] # at time n
# Next, all the other probability distributions calculated in
    a loop
> for (n in seq(2,100)){
>   next_step<-NA
>   next_step<-c(P_m1[n-1],P_m2[n-1])%*%P # <--- p(n) = p(n-1)
     P
>   P_m1[n]<-next_step[1]
>   P_m2[n]<-next_step[2]
> }
# plot the n-step probability distributions
> plot(seq(0,100),c(p_0[1],P_m1),type="l",xlab="n time steps",
     ylab="Prob after n steps",col="red",ylim=c(0,1))
> lines(seq(0,100),c(p_0[2],P_m2), col="blue")
> legend("topright",col=c("red","blue"),
       legend=c("State 1 (company K cereals)","State 2 (
    competition cereals"))
# Next, we plot the probabilities, we use the library(ggplot2)
> my.data<-data.frame(c(seq(0,100),seq(0,100)),
> matrix(c(p_0[1],P_m1,p_0[2],P_m2),nrow=202,ncol=1,byrow=
    FALSE),
                    as.factor(c(rep(1,101),rep(2,101))) )
> colnames(my.data)<-c("Time.Steps","Prob.After.n.steps","
    State")
> ggplot(my.data, aes(x=Time.Steps, y=Prob.After.n.steps, fill
    =State)) +
  geom_area()
# plot the differences for company K: proportion of market now
     minus one step ago
 we have to add the proportion (probability) at time 0.
> plot(diff(c(p_0[1],P_m1)),type="l",xlab="n",ylab="X at n - X
     at n-1")
# Next, we print the first 38 differences
> diff(P_m1[1:38])
# Long-term prediction:
> c(P_m1[100],P_m2[100])

OUTPUT:
> Changes in the probability distribution of K cereal
> diff(P_m1[1:5])
 [1] 6.022500e-02 4.396425e-02 3.209390e-02 2.342855e-02
    1.710284e-02
> Long-term prediction:
> c(P_m1[100],P_m2[100])
[1] 0.5555556 0.4444444
```

The R-code above estimated the changes in the share of the market and produced Figure 4.6. The share of K is smaller at year 0, but then, after about 12 years, it gets a higher and higher share of the market till it stops changing. So the long-term prediction for the market is a state that is not changing.

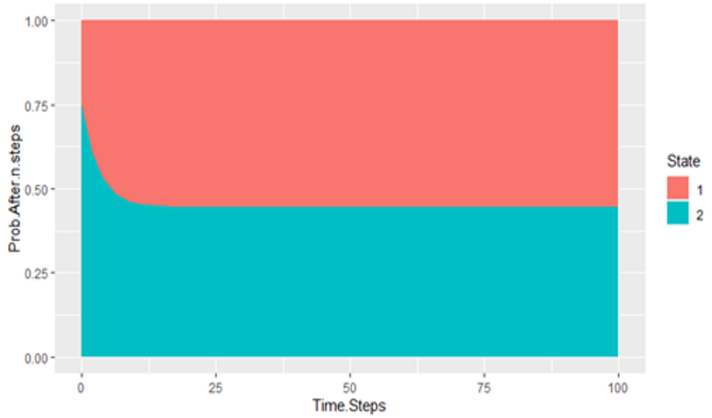


FIGURE 4.6: Cereals market shares over time, with the initial shares 0.25 and 0.75. The states are two: 1 = K brand cereal and 2 = Competition brand. The time steps are years, i.e. it is assumed that each customer is switching or staying with the same brand once a year.

Do we expect the actual market share to approach the long-term prediction for the market or not? In other words: is the calculated long-term prediction likely to happen? To answer this question we need to think about what can intervene with our long-term prediction. Based on the figure, it takes about 12-time steps (hence years) to reach the long-term prediction of the steady state. In our calculations, we assumed that the transition matrix $\mathbf{P}$ is not changing over the 12 years. Any changing circumstances can change the matrix $\mathbf{P}$. For example, if a new competitor enters the market this would likely render the transition matrix invalid and hence make it impossible for the actual market share to reach the calculated long-term prediction.

Next, we will change the initial state: 0.0001 and 0.9999 (K and competition cereals), so we can see if and how the market share is affected by the initial distribution. Use R to find the market share after n steps (n=1..100). We use the same R-code as above, with one change:

```
> Initial probability distribution
> p_0 <- c(0.0001, 0.9999)
```

Then the output is:

```
> OUTPUT:
> Changes in the probability distribution of K cereal
> diff(P_m1[1:5])
 [1] 6.022500e-02 4.396425e-02 3.209390e-02 2.342855e-02
     1.710284e-02
>  c(P_m1[100],P_m2[100])
[1] 0.5555556 0.4444444
```

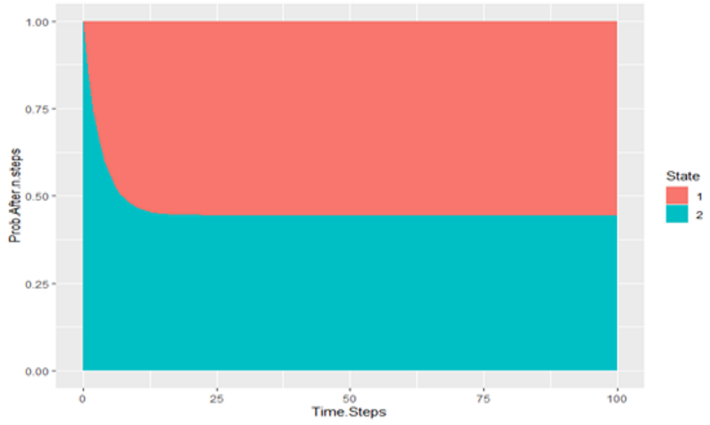


FIGURE 4.7: Cereals market shares, with the initial shares 0.0001 and 0.9999. The states are two: 1 = K brand cereal and 2 = Competition brand. The time steps are years, i.e. it is assumed that each customer is switching or staying with the same brand once a year.

When we compare the two calculations, the initial state 0.25 and 0.75 vs. 0.0001 and 0.9999 (Figures 4.6 and 4.7), we see that the effect of the initial state wears off. So the long-term prediction is the same as before.

In what follows, there are further R-lab questions for you to work on. A solution is not provided here but can be provided upon request.

2. [**Purpose: Getting experience of using R for simulation of Markov Chains, specifically simulation of individual customers from Section 4.1 - while using library called** `base`**.**]

```
> # Goal: This is a simulation of a Markov Chain.
> # We simulate a person buying either K (1) or a competitor
    brand (2).
> states<-c(1,2) # 1 for K, 2 for competitor
> # Initial probability distribution
> p_0 <- c(0.25,0.75)
> # P, one-step transition matrix
> P = matrix(c(0.88,0.12,0.15,0.85),nrow=2,ncol=2,byrow=TRUE)
> # Simulate the initial state of a customer
> X <- sample(states,1,TRUE,p_0)
> current<-X
> for (i in 1:100){
  nextstate<-sample(states,1,TRUE,P[current,])
  X<-c(X,nextstate)
  current<-nextstate
}
> X
OUTPUT: Three outputs (i.e. running the above code three times
    ) for three independent customers:
```

```
> X
  [1] 2 2 2 2 2 2 2 2 2 1 2 2 2 1 1 2 2 2 2 2 1 1 1 1 1 1 1 1
    1 1 1 1 1 1 1 1 1 1 1 1 1
 [42] 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1
    1 1 1 1 1 1 1 1 1 1 1 1 1
 [83] 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 1 1 1 1 1 2 1 1 1 1 1
    1 2 2 2 1 1 1 1 1 1 1 1 2
[124] 2 2 2 2 2 2 2 2 2 2 1 1 1 2 2 2 2 2 1 1 1 2 2 2 2 2 2 2 2
    2 1 1 1 1 1 1 1 1 1 1 1 1
[165] 1 1 1 1 1 1 1 1 1 2 2 2 2 1 2 1 1 1 1 1 1 2 2 1 1 1 1 1 1
    1 1 1 1 1 1 1 1 1 1
> X
  [1] 2 1 1 1 1 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2 2 2
    2 2 2 2 2 2 2 2 2 2 2 1 2
 [42] 1 1 1 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1
    1 1 1 2 2 1 1 1 2 2 2 2 2
 [83] 2 2 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1
    1 1 1 1 1 1 1 1 1 1 1 1 1
[124] 1 2 2 2 2 2 1 1 1 1 2 2 2 2 2 2 2 1 2 1 1 1 2 2 2 2 2 2 2
    2 2 2 2 1 2 1 2 2 2 2 2 2
[165] 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 2 2 2 2 2 2 2 2 1 1 1 1 2 2
    1 1 2 2 2 2 1 1 1
> X
  [1] 2 2 2 2 2 2 2 2 2 1 1 2 1 1 1 1 1 1 1 1 2 2 2 2 2 2 1 1 2
    2 1 2 2 2 2 2 1 1 1 1 1 1
 [42] 1 1 1 1 2 2 2 2 2 2 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2
    2 1 1 1 1 1 1 1 1 1 1 1 1
 [83] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2 2
    1 1 1 1 2 2 2 1 1 1 1 1 1
[124] 1 1 2 2 2 2 2 2 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2
    2 2 2 2 2 2 2 2 2 1 1 1 1
[165] 1 2 2 2 2 2 2 2 1 1 2 1 1 1 1 1 1 1 1 1 2 2 1 1 2 2 1 1 1
    1 1 1 1 1 1 1 1 1 1
```

3. **[Purpose: Defining matrix P in R, from the Example Commuter Cyril from Section 4.2.]** In R, create a 2-by-2 matrix P. It should contain the 1-step transition probabilities from the Example Commuter Cyril. Then calculate the matrix of 2-step transition probabilities, i.e. P^2.

```
> P<-matrix(c(0,0.5,1,0.5),2,2)
> P
     [,1] [,2]
[1,]  0.0  1.0
[2,]  0.5  0.5
> P%*%P
     [,1] [,2]
[1,] 0.50 0.50
[2,] 0.25 0.75
```

4. **[Purpose: Defining matrix P in R, from the Example Commuter Cyril from Section 4.2.]** Use R and Markov Theorem to find the n-step predictions and long-term predictions. Assume initial state: 5/6 and 1/6 for (train and drive). What is the long-term prediction for the commute behaviour of this person?

```
# R Code. This is not a simulation of a Markov Chain.
# Goal: This is a calculation of the multi-step probability distributions
# so we can numerically find the long-run behaviour of this Markov Chain.

# To initialise the two states.
states<-c(1,2) # 1 for Train,  2 for drive the car

# initial probability distribution
p_0<-c(5/6, 1/6)

step transition matrix
P=matrix(c(0,1,.5,.5),nrow=2,ncol=2,byrow=TRUE)                                    # matrix P
# P, one # multi-step probability distributions for states 1 and 2 at steps 1 to 100
P_m1<-rep(NA,100)
P_m2<-rep(NA,100)

# Step 1 must be done manually
next_step<-NA
next_step<-p_0%*%P
P_m1[1]<-next_step[1]
P_m2[1]<-next_step[2]
# Other steps in a loop
for (n in seq(2,100)){
  next_step<-NA
  next_step<-c(P_m1[n-1],P_m2[n-1])%*%P                                                # p^((n))=p^((n-1)) P
  P_m1[n]<-next_step[1]
  P_m2[n]<-next_step[2]
}

# plot the n-step probability distributions
plot(seq(0,100),c(p_0[1],P_m1),type="l",xlab="n time steps",
     ylab="Prob after n steps",col="red",ylim=c(0,1))
lines(seq(0,100),c(p_0[2],P_m2), col="blue")
legend("topright",col=c("red","blue"),
       legend=c("State 1 car","State 2 train"))
# library(ggplot2)
my.data<-data.frame(c(seq(0,100),seq(0,100)),
                    matrix(c(p_0[1],P_m1,p_0[2],P_m2),nrow=202,ncol=1,byrow=FALSE),
                    as.factor(c(rep(1,101),rep(2,101))) )
colnames(my.data)<-c("Time.Steps","Prob.After.n.steps","State")
ggplot(my.data, aes(x=Time.Steps, y=Prob.After.n.steps, fill=State)) +
  geom_area()
# plot the differences for Company K: proportion of the market now minus one step ago
# We have to add the proportion (probability) at time 0.
plot(diff(c(p_0[1],P_m1)),type="l",xlab="n",ylab="X at n - X at n-1")
# print the first 38 differences
```

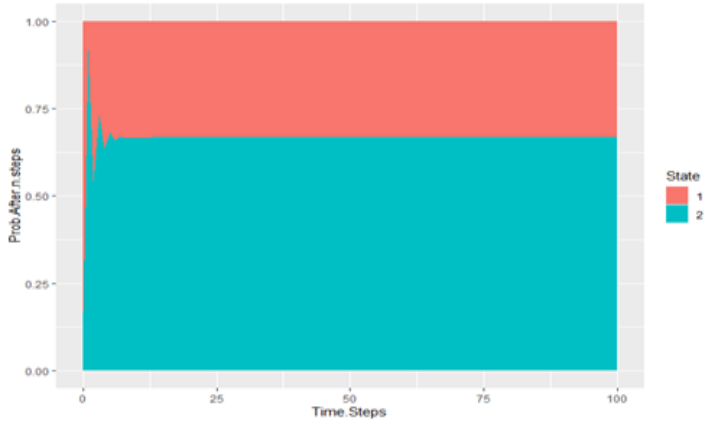


FIGURE 4.8: Changes in the market share over time, for the Cereals market shares example, with the initial shares 5/6 and 1/6. The states are two: 1 = K brand cereal and 2 = Competition brand. The time steps are years, i.e. it is assumed that each customer is switching or staying with the same brand once a year.

```
diff(P_m1[1:38])
# stationary distribution (probabilities in steady-state)
c(P_m1[100],P_m2[100])

OUTPUT:

Changes in the probability distribution of K cereal:
> diff(P_m1[1:5])
 [1]  3.750000e-01 -1.875000e-01  9.375000e-02 -4.687500e-02
    2.343750e-02
>  c(P_m1[100],P_m2[100])
[1] 0.3333333 0.6666667
```

There were some "oscillations" during the first 10 time steps (see Figure 4.8). But then the long-term prediction is that this person will achieve steady-state, in 1 out of 3 days he will use the train and in 2/3 days he will use the car.

Would you expect the actual commute of this person to approach the long-term prediction or not? In other words: is the calculated long-term prediction likely to happen? Think how many time steps it takes to achieve long-term prediction and what can intervene with your prediction during those time steps.

4.7 Exercises

Solve the following exercises by using pen, paper and calculator.

1. **[Purpose: Practicing Markov Chains topics from Section 4.1.]** In analysing switching by Business Class customers between airlines the following data has been obtained by British Airways (BA):

		Next flight by BA	Next flight by Competition
Last flight by	BA	0.85	0.15
	Competition	0.10	0.90

   Currently, BA has 30% of the Business Class market. Business Class customers make two flights per year on average. (Hint: Square the matrix first to get the yearly switching data).

   a) Draw the state-transition diagram.

   b) What would you forecast BA's share of the Business Class market to be after two years?

2. **[Purpose: Practicing Markov Chains topics from Section 4.1.]** A company is considering using Markov theory to analyse brand switching between three different brands of floppy disks. Survey data has been gathered and has been used to estimate the following transition matrix for the probability of moving between brands each month:

		To brand 1	To brand 2	To brand 3
From brand	1	0.80	0.10	0.10
	2	0.03	0.95	0.02
	3	0.20	0.05	0.75

   The current (month 1) market shares are 45%, 25% and 30% for brands 1, 2 and 3 respectively.

   a) What will be the expected market shares after two months have elapsed?

   b) What is the long-run prediction for the expected market share for each of the three brands?

   c) Would you expect the actual market share to approach the long-run prediction for the market or not (and why)?

3. **[Purpose: Practicing Markov Chains topics from Section 4.1.]** Vauxhall is currently investigating the behaviour of car fleet buyers in switching between companies. Preliminary investigations have revealed that fleet buyers usually have a number of companies from whom they buy, but that they tend to keep in mind a target percentage for each company. Market research indicates that, for the purpose of preliminary analysis, a fleet buyer's target percentage for Vauxhall cars can be regarded as being one of 100%, 70%, 50% or 20%. An in-depth study of previous buying behaviour has produced the transition matrix shown below for the probability of switching each year between target percentages:

		To target %			
		100	70	50	20
From target %	100	0.60	0.30	0.10	0.00
	70	0.00	0.70	0.30	0.00
	50	0.40	0.40	0.20	0.00
	20	0.00	0.20	0.50	0.30

The current situation is that, for every 100 fleet buyers, 5 have a target percentage for Vauxhall cars of 100%, 30 a target percentage of 70%, 45 a target percentage of 50% and 20 a target percentage of 20%.

a) Draw the state-transition diagram.

b) What will be the percentage of fleet buyers having a target percentage for Vauxhall cars of 50% in (i) 2 years' time and (ii) the long run?

c) Would you expect the actual percentage of fleet buyers having a target percentage for Vauxhall cars of 50% to approach the long-run figure calculated above or not (and why)?

d) What advantages and disadvantages can you think of in using Markov theory to forecast fleet buyers' behaviour in this way?

1. [**Purpose: Practicing Markov Chains topics from Section 4.2.**] The article "Markov Chain Model for Performance Analysis of Transmitter Power Control in Wireless MAC Protocol" (Twenty-first International Conference on Advanced Networking and Applications, 2007) describes a Markov chain model for the state of a communication channel using a particular "slotted non-persistent" (SNP) protocol, for hourly time scale. The channel's possible states are (1) idle, (2) successful transmission, and (3) collision. For particular values of the authors' proposed four-parameter model, the following one-step transition matrix of this Markov chain is:

$$\mathbf{P} = \begin{bmatrix} 0.50 & 0.40 & 0.10 \\ 0.02 & 0.98 & 0.00 \\ 0.12 & 0.00 & 0.88 \end{bmatrix}$$

   a) Draw the state transition diagram for this chain.

   b) Is this Markov chain irreducible?

   c) Is this chain aperiodic?

   d) We are told that $P(Y_0 = 1) = 0.1$, $P(Y_0 = 2) = 0.8$. Determine the probability distribution of Y_3, i.e. of the state of the communication channel three hours after the initial valuation. Interpret.

   e) At time 8 hours, what is the probability that the system is in a collision if it was idle at time 7?

   f) At time 8 hours, what is the probability that the system is in a collision if it was idle at times $0, 1 \ldots, 7$?

   g) At time 8 hours, what is the probability that the system is in a collision if it was successfully transmitting at times $0, 1 \ldots, 6$, and then at time 7 it was idle?

   h) Find $P(Y_0 = 1, Y_1 = 2, Y_2 = 1, Y_3 = 3)$ and interpret it.

   i) At time 5 hours, what is the probability that the system is in a collision if it was successfully transmitting at times $0, 1 \ldots, 4$?

   j) What is the long-term prediction for this channel? Does it achieve a stationary distribution?

   k) What proportion of time is this channel idle, in the long run?

   l) Would you expect the actual channel to approach the long-run prediction or not? Justify your answer.

2. **[Purpose: Practicing Markov Chains topics from Section 4.2.]** The article "Markov Chain Models of Negotiators' Communications" (Encyclopaedia of Peace Psychology 2012: 608-612) describes the following set-up for the back-and-forth dialogue between two negotiators. If at any stage a negotiator engages in a cooperative strategy, the other negotiator will respond with a cooperative strategy with a probability of 0.6. Otherwise, the response is described as a competitive strategy. Similarly, there is a probability of 0.7 that a competitive strategy offered at any stage of the negotiations will be met by another competitive strategy. Let Y_n= strategy employed at the n-th stage of a negotiation.

   a) Identify the state space for the chain, specify its one-step transition probabilities, and draw the corresponding state diagram.

   b) Construct the one-step transition matrix for the Markov chain Y_n= strategy employed at the n-th stage of a negotiation, assuming the states are (1) cooperative and (2) competitive.

   c) If negotiator A employs a cooperative strategy at some stage, what is the probability she uses a competitive strategy the next time? [don't forget that A's turns are two-time steps apart since B counter-negotiates in between.]

   d) Now introduce a third state (3) end of the negotiation. Assume that a Markov chain model with the following one-step transition matrix applies:

$$\mathbf{P} = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.3 & 0.4 & 0.3 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$$

   Given that the initial strategy presented is cooperative, what is the probability the negotiations end within three-time steps?

   e) Refer back to c). Given that the initial strategy presented is competitive, what is the probability the negotiations end within three-time steps?

3. **[Purpose: Practicing Markov Chains topics from Section 4.2.]** Markov chains are often used to model changing weather conditions; research literature in both meteorology and climate science is filled with Markov chain applications. The paper by Rodondi [49] provides data for several US cities on the daily transitions between "snow days", defined by a snow depth of at least 50 mm, and "green days" (snow depth ¡ 50mm).

   So, for each city, we can view the weather as a system, and the system can be in one of two states: green day or snowy. In probability language, this means that the state space of the system (i.e. the list of all the states) is G, S or simply green day, snow. Let X_n represent the snow status, either S for snow or G for green, on the nth recorded day. For New York City, the following one-step transition probabilities are provided in the article:

$$P(G \to G) = 0.964, \quad P(G \to S) = 0.036$$
$$P(S \to G) = 0.224, \quad P(S \to S) = 0.776$$

So, if today is a "green day" in New York, then there is a 96.4% chance that tomorrow's snow depth will also be below 50 mm, based on the available weather data (which, incidentally, stretches back to the year 1912 for New York). On the other hand, as the author notes, "the presence of a significant snow depth (accumulation) on the current day in Central Park (New York) has an approximately 1 in 5 chance of melting before the next day". Hence the one-step transition matrix is

		Next weather	
		G	S
Last weather	G	0.964	0.036
	S	0.224	0.776

And we have a sequence of random variables $X_0, X_1, X_2, X_3 \ldots$ which is a Markov Chain, i.e. it has Markov property because the probability of tomorrow's state is determined by the state today and by the transition matrix P i.e. we do not need to know the states of snow or green in days before today.

a) Identify the state space for the chain, specify its one-step transition probabilities, and draw the corresponding state diagram.

b) What is the long-run prediction for this weather?

c) Will it achieve a steady state?

d) What proportion of days will it be green and what proportion snowy, in the long run?

Part IV

DECISIONS UNDER RISK

5

Decisions under precise risk and under imprecise risk

Gabriela Czanner

Silvester Czanner

CONTENTS

5.1 Making decisions under precise risk 187
5.1.1 Measuring the risk with variance 187
5.1.2 What are the preferences of people toward risk? 190
5.1.3 Maximising expected utility of a decision maker 194
5.1.4 Probability premium 200
5.1.5 Monetary premium 202
5.2 Making decisions under imprecise risk 203
5.2.1 General strategy when probabilities are not known 203
5.2.2 Alternative decision criteria 204
5.3 Tips to think and act like a risk expert 208
5.3.1 Be aware of the criticism of the alternative criteria 208
5.4 Summary 209
5.5 Further reading 210
5.6 Exercises 212

When Cameron enters a casino and decides to bet money, assuming he is sufficiently skilled and rational, he can calculate all the possible outcomes that can happen and their probabilities. When he decides on his betting strategy in the casino, he knows he faces the risk of losing money or the chance of winning money. He does not know whether he will win or lose, but he knows the chances (probabilities). We call such a situation: precise risk (see also Chapter 1).

A different situation occurred in early 2020 when the Covid-19 pandemic started. Clinicians could articulate all future outcomes for a Covid-19 positive patient, but they could not give probabilities of each outcome. The Covid-19 situation was different from Cameron's casino situation: the Covid-19 situation

involved uncertainty about surviving or dying, as well it involved uncertainty about the probabilities of surviving and dying. We call such a situation: imprecise risk (see Chapter 1). Even though risk always involves less-than-complete information, there is less information in situations of imprecise risk.

Cameron deciding on betting in a casino is a clear-cut example of making decisions under precise risk. Such clear-cut decision-making situations under precise risk are unusual in real life. In real life, there are often uncertainties about outcomes or probabilities or consequences (see Figure 1.3). Life is more like an expedition into an unknown jungle rather than a visit to a casino. Yet, sometimes, decision-makers (sometimes wrongly encouraged or misinformed by data modellers) proceed as if they had reliable estimates of all outcomes, consequences and chances: they assume that the lack of knowledge is the same as random chance. Such a mistake is called the **tuxedo fallacy** [30], , and it leads to suboptimal, even fatal, decisions.

Various indications suggest that we have uncertainty in knowing the possible outcomes o consequences, including:

1. Variability within a sampled population or repeated measures leading to, for example, statistical margins-of-error and prediction intervals, e.g. high variability in weekly Covid-19 new cases will lead to wide prediction intervals,

2. Computational or systematic inadequacies of measurement, also called imprecision of measurement, e.g. underreporting the new weekly Covid-19 cases will lead to too low forecasts,

3. Limited knowledge and ignorance about underlying processes, e.g. if we do not know how Covid-19 spreads from one person to another, then we cannot employ the right data analytical methods for forecasting,

4. Expert disagreement, the credibility of a witness on a criminal legal case.

Note that another situation of imprecise risk is when the probabilities are only available with some imprecision. For example, we may estimate a probability of a person having cancer as 5 to 8% probability (which would be a case of a small probability uncertainty since 8-5 is 3%). Or we can estimate for another person the probability of cancer to be 5 to 20%, which would be a case of medium probability uncertainty. And for a third person, we may estimate the probability of cancer to be 5 to 80%, thus large probability uncertainty. If we estimate the probability to be between 0 to 100%, then this is a *complete probability uncertainty* (as a special case of imprecise probability).

Learning objectives

1. Learn that each decision maker has a different attitude to risk, and this attitude can be expressed via utility functions.

2. We will look into how to make decisions when facing precise risk. We will learn to consider the decision maker's utility in our recommendations.

3. Then we will learn how to decide in a situation of imprecise risk, where we do not know the probabilities of outcomes, i.e. complete probability uncertainty. We will explore several alternative decision criteria.

5.1 Making decisions under precise risk

Next, we look into several strategies that we use to advise a decision-maker who is facing a risk.

5.1.1 Measuring the risk with variance

Example. Makovnik Bakery. To motivate the decision under precise risk, we consider an example of the owner of a Slovak bakery called Makovnik. The owner is deciding where to open a new bakery. Figure 5.1 shows the probability distributions of possible weekly profits if the owner decides to locate the new bakery in either Kuchyňa, Lozorno, or Malacky. The means, standard deviations, and coefficients of variation for each distribution are displayed in Figure 5.1. Which city should the owner choose to open one new bakery?

We now outline three rules for making decisions under precise risk:

1. The rule of **maximum expected value**: choose the decision with the highest expected value. This rule employs the mean in order to make a decision.

2. The **mean-variance rules**: employ both mean and variance in choosing a decision. We discuss it below in more detail.

3. The rule of **minimum coefficient of variation**: choose the decision with the smallest coefficient of variation. *coefficient of variation* (CV) is a statistical measure of the relative dispersion of data points in a dataset around the mean. It represents the ratio of the standard deviation to the mean:

$$CV = E[X]/SD[X]$$

where X is the random variable representing all possible outcomes for opening the restaurant in, e.g. Kuchyňa, and each outcome has a known probability of happening, E[X] is the mean, and SD is the standard deviation. The CV is useful for comparing the degree of variation from one data series to another, even if the means are different from one another.

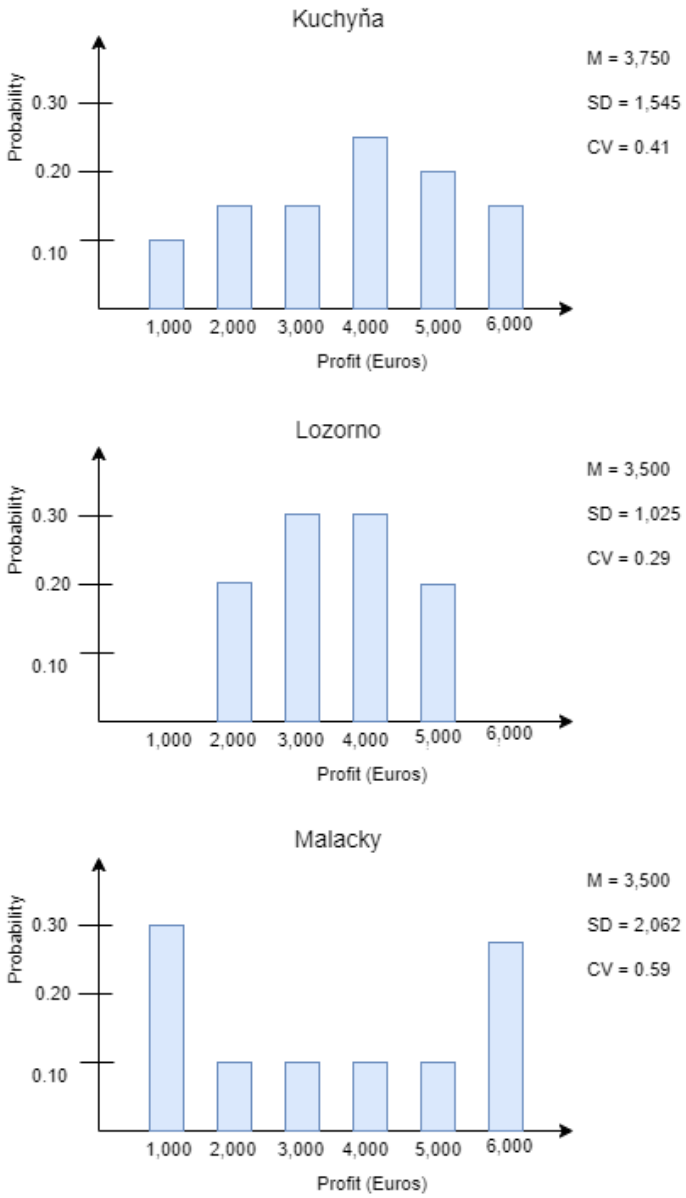


FIGURE 5.1: Weekly profits in Example Makovnik Bakery at three locations. M = Mean, SD = Standard Deviation, CV = Coefficient of variation.

The mean-variance rules are as follows: employ both mean and variance in choosing a decision. Given two risky decisions (designated A and B), the mean-variance rules for decisions under precise risk are

1. If decision A has a higher expected outcome and a lower variance than decision B, decision A should be made.

2. If both decisions A and B have identical variances (or standard deviations), the decision with the higher expected value should be made.

3. If decisions A and B have identical expected values, the decision with the lower variance (standard deviation) should be made.

The mean–variance rules are based on the assumption that a decision maker prefers a higher expected return to a lower, other things equal, and a lower risk to a higher, other things equal. It, therefore, follows that the higher the expected outcome and the lower the variance (risk), the more desirable a decision will be. Under rule 1, a decision maker would always choose a particular decision if it has a greater expected value and a lower variance than other decisions being considered. With the same level of risk, the second rule indicates decision maker should choose the decision with the higher expected value. Under rule 3, if the decisions have identical expected values, the decision maker chooses the less risky (lower standard deviation) decision.

Example. Makovnik Bakery. (Continues.) The expected weekly profit in Kuchyňa is \$3,750, in Lozorno is \$3,500, and in Malacky is \$3,500. Using the rule of maximum expected value, the owner will choose Kuchyňa. In this rule, the owner is not concerned (or is oblivious) with risk. Note that if the owner had been choosing between only the Lozorno and Malacky locations, the expected value rule could not have been applied because each has an expected value of \$3,500. In such cases, some other rule may need to be used.

How do we make decisions using the mean-variance rules?

- There is no location that dominates. Even though Kuchyňa has the highest expected value, it does not have the smallest spread among all three locations.

- However, Kuchyňa dominates Malacky because it has a higher expected value and a lower risk (rule 1).

- Lozorno also dominates Malacky regarding rule 3 because both locations have the same expected value (\$3,500), but Lozorno has a lower standard deviation, thus less risk. So, if a decision is made between Lozorno and Malacky, the location in Lozorno should be chosen.

- Next, how do we choose between Kuchyňa and Lozorno?

- If the owner compares the Kuchyňa and Lozorno locations, the mean–variance rules cannot be applied. Kuchyňa has a higher weekly expected profit ($3,750 >$3,500), but Lozorno is less risky (SD of Lozorno ¡ SD or Kuchyňa). Therefore, when making this choice, the owner must make a trade-off between risk and expected return, which would depend on the owner's valuation of higher expected return versus lower risk. We will use the coefficient of variation rule because it uses information on the expected value and dispersion and can be used to make decisions involving trade-offs between expected return and risk.

Using the rule of minimum coefficient of variation, the owner will calculate the CV first. They are

- CV (Kuchyňa) = 1,545 / 3,750 = 0.41
- CV (Lozorno) = 1,025 / 3,500 = 0.29
- CV (Malacky) = 2,062 / 3,500 = 0.59

Since Lozorno has the smallest CV, the owner will choose Lozorno, according to the Rule of minimal coefficient of variation.

Which rule is the best? At this point, you may wonder which of the three rules for making decisions under precise risk is the "correct one." After all, the owner of Makovnik Bakery either reached a different decision or did not decide, depending on which rule was used. Using the expected value rule, Kuchyňa was the choice. Using the coefficient of variation rule, Lozorno was chosen. According to mean–variance analysis, Malacky was out, but the decision between Kuchyňa and Lozorno could not be reached. If the decision rules do not all lead to the same conclusion, the owner must decide which rule to follow.

The art of decision-making under precise risk is closely associated with a decision-maker's (or stakeholder's) preferences concerning risk-taking. Decision makers (e.g. business owners, managers, prime minister, head of the school, people, you) can differ greatly in their willingness to take risks in decision-making. Some are quite cautious, while others may seek out high-risk situations.

In the next section, we present a theory of decision-making under precise risk that formally accounts for a people's attitude toward risk: if they like risk a lot, if they are okay with some level of risk, or if they avoid risk as much as possible.

5.1.2 What are the preferences of people toward risk?

Utility is a term economists use to describe the measurement of *usefulness* that a consumer obtains from any good. The utility may measure how much

one enjoys a movie or the sense of security from buying a deadbolt. Different people may have different utilities.

Example. Two lotteries. You are asked to choose between two lotteries: A and B.

- Lottery A: Get £3,125 for sure
- Lottery B: Win £4,000 with probability 0.75, and win £500 with probability 0.25

Which lottery do you prefer? The rule of maximum expected value advises choosing the one with a higher expected value (i.e. choosing the one that gives a higher expected profit). Is the rule of maximum expected value a good criterion for deciding between two lotteries?

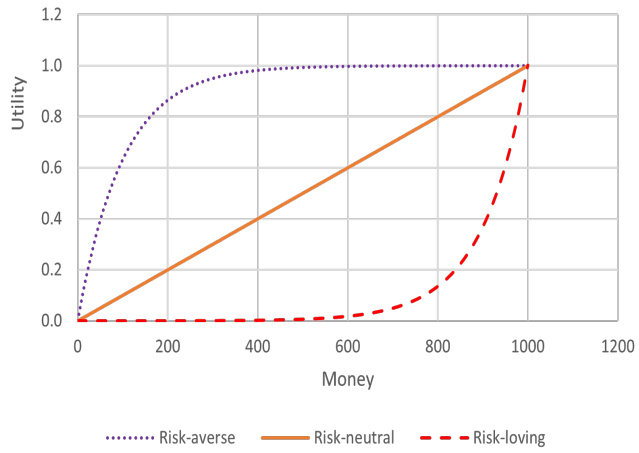


FIGURE 5.2: Example utility functions for risk-averse, loving and neutral people.

To answer the question, we take a close look at the two lotteries:

- The two lotteries have the same expected value: £3,125.
- The lottery A has no risk, and the variance is zero. So by the rule of maximum expected value, we cannot decide. However, by the mean-variance rules and the rule of minimal coefficient of variation, we decide on lottery A.
- Probably most people will choose Lottery A because they dislike risk. Such a person is called risk averse.

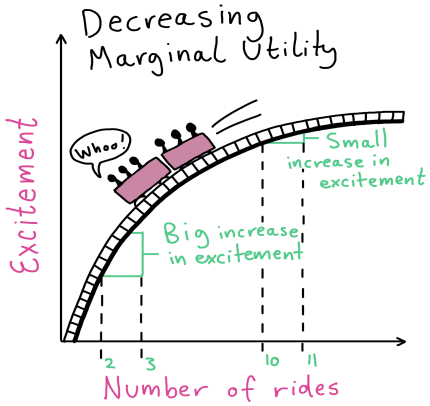


FIGURE 5.3: Diminishing (decreasing) marginal utility. The gain in happiness from one ride depends on the person's starting point. Giving an additional ride to a person who has done ten rides will increase his/her happiness less than giving an additional ride to a person who has done two rides.

- However, according to the rule of maximal expected value, both lotteries are equivalent. Expected value is not a good criterion for people who dislike risk. We call such people risk averse.
- If someone is indifferent between A and B, the risk is unimportant to him. Such a person is called risk-neutral.
- A risk-loving person is an investor who is willing to take on additional risk for an investment that has a relatively low additional expected return in exchange for that risk. For example, if there is a Lottery C: win £4,000 with a probability of 0.1 and 500 with a probability of 0.9, the expected return is £400+450=850. The risk-loving person would choose Lottery C even if it gives lower expected returns than Lottery A and Lottery B.

The Two lotteries example illustrates that the optimal decision depends on the utility of the decision maker, i.e. how he/she values the outcomes as well as the risk. Not only does a decision maker care about the utility that the money provides, but an individual also cares about the risk. A rational person values Lottery A more than Lottery B despite having the same expected monetary value. This is because Lottery A has a lower risk (in fact, no risk).

There are three types of people's attitudes toward risk (Figure 5.2):

- *Risk averse person.* His or her utility function is concave by definition. It

is a decision maker who makes the less risky of two decisions that have the same expected value.

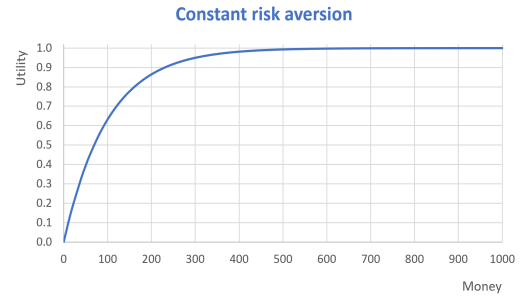
- *Risk loving person.* A decision maker who makes the riskier of two decisions that have the same expected value.
- *Risk neutral person.* A decision maker who ignores risk in decision-making and considers only the expected values of decisions.

We define utility formally as a function $U(x)$, where x is the money amount. In general, x can also be any goods or services, but in this book, we will consider monetary utilities only. A utility function must satisfy the following three properties:

- *Monotonically increasing utility.* This means that individuals prefer more money than less money, so the utility function is a strictly increasing function:
$$U'(x) > 0$$
- *Decreasing marginal utility.* This means that the same increase in money will cause a smaller increase in happiness if the person already owns more money. In other words, the gain in utility depends on the starting point, e.g. if I am rich or not. This property is also called diminishing marginal utility. Mathematically, such property is satisfied by concave functions (Figure 5.3)
$$U''(x) < 0$$

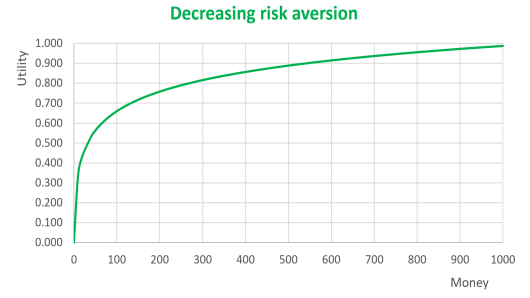
Utility functions for profit are often defined on continuous space of outcomes (like assets or wealth), with continuous probabilities, and hence the utilities are continuous functions. They are also called *monetary utilities*. We could use gains or losses in making a monetary decision, e.g. an investment decision. However, it makes more sense to use our current position (our assets). The decisions we make could depend on the starting point. For example, my decision may depend on whether I am rich or poor. I could make a riskier decision if I am rich than poor. Monetary utility functions are defined over a continuous variable (assets) or wealth.

Constant risk aversion is one of the types of risk aversion. Let us assume that such a person forms a portfolio with one risky and one risk-free asset. If the person experiences an increase in wealth, he/she will choose to keep unchanged the number of pounds of the risky asset held in the portfolio since her/his risk aversion is constant. An example utility curve for a person with a constant risk aversion is
$$u(x) = 1 - e^{-cx} \tag{5.1}$$
where $c = 0.01$ (Figure 5.4). In this case, the probability premium is constant for a given value of c.



Money	0	10	20	50	100
Utility	0.000	0.095	0.181	0.393	0.632

FIGURE 5.4: Utility of a person with constant risk aversion.



Money	0	10	20	50	100
Utility	0.000	0.343	0.435	0.562	0.659

FIGURE 5.5: Utility of a person with decreasing risk aversion.

Decreasing risk aversion is another type of risk aversion. Let us assume that such a person is forming a portfolio with one risky asset and one risk-free asset. If the person experiences an increase in wealth, he/she will choose to increase the number of pounds of the risky asset held in the portfolio since her/his risk aversion is decreasing. An example utility curve for a person with a constant risk aversion is

$$u(x) = \frac{ln(x+a) - ln(a)}{b} \tag{5.2}$$

where $a = 1$ and $b = 7$ (Figure 5.5). Upon visual inspection of the plot, we see that the utility curve is initially steeper and then flattens off.

5.1.3 Maximising expected utility of a decision maker

Imagine a situation with m possible decision options we can take, $d_1, ..., d_m$. There are also n possible outcomes when we make a decision, $\theta_1, ..., \theta_n$. And each outcome is associated with a probability, $P(\theta_1), ..., P(\theta_n)$. We also assume that each combination of decision (d_i) and event (θ_j) will have a consequence $C_{i,j}$ on us. It is useful to organise all values as shown in Figure 5.1.

		Outcome			
		θ_1	x	z	θ_2
Decision	d_1	$u(C_{1,1})$	$u(C_{1,2})$	$\cdots$	$u(C_{1,n})$
	d_2	$u(C_{2,1})$	$u(C_{2,2})$	$\cdots$	$u(C_{2,n})$
	$\vdots$	$\vdots$			$\vdots$
	d_m	$u(C_{m,1})$	$u(C_{m,2})$	$\cdots$	$u(C_{m,n})$
Probabilities	$P(\theta)$	$P(\theta_1)$	$P(\theta_2)$	$\cdots$	$P(\theta_n)$

TABLE 5.1: Decisions, states and utilities. This table illustrates how we organise all of the decision options, measures of uncertainty (such as probabilities) and utilities into a table format.

Each consequence $C_{i,j}$ in Table 5.1 has a utility assigned to it. A utility is a number that says how much the relevant stakeholder likes a consequence. The number $u(C_{i,j}$ is the utility of consequence $C_{i,j}$, i.e. the utility of making decision i and facing the outcome j. If $u(C_{i,j}) > u(C_{k,l})$ then this means that $C_{i,j}$ is better than $C_{k,l}$. An illustration of these concepts is shown in Piston's example.

Example. Piston. Manufacturer Peng is producing a piston for an engine (Figure 5.6), and she needs to decide whether to sell her item or whether to perform extensive (and hence expensive) testing and then sell it. We will assume that the piston is either free from flaws or has flaws and that extensive testing will catch all the flaws. The consequences of this situation are summarised in Table 5.2.

		Outcome	
		θ_1 Good part	θ_2 Poor part
Decision	d_1 Inspect	$C_{1,1}$	$C_{1,2}$
	d_2 Do not inspect	$C_{2,1}$	$C_{2,2}$

TABLE 5.2: Manufacturer Peng's table for utilities and consequences, in Piston example.

The value $C_{i,j}$ is the consequence of the decision i and outcome j. The

consequences are defined in a textual way. The value $C_{1,1}$ = the manufacturer's piston is free from flaws and the manufacturer decides to inspect it. The value $C_{2,1}$ = the manufacturer's piston is free from flaws, and the manufacturer decides not to inspect it. The value $C_{1,2}$ = the manufacturer's piston is flawed, and the manufacturer decides to inspect it. The value $C_{2,2}$=The manufacturer's piston is freely flawed, and the manufacturer decides not to inspect it. Of course, the manufacturer has no idea about the condition of the piston when she has to decide on whether to test it.

FIGURE 5.6: A piston.

The engineer prefers some scenarios more than others. Her preferences are expressed in utilities u in Table 5.3. The bigger u, the better utility or perceived usefulness. The $C_{1,1}$) consequence has the largest utility. For this manufacturer (the one whose utilities are in Table 5.3), it is the best consequence to hold in hands a piston that was just confirmed to be good by the test. The manufacturer decides on these utilities, or a data analyst elicits the utilities from the manufacturer. In this example, the piston manufacturer has a finite number of outcomes; hence we have a finite number of utilities.

		Outcome	
		θ_1 Good part	θ_2 Poor part
Decision	d_1 Inspect	$u(C_{1,1}) = 0.9$	$u(C_{1,2}) = 0.5$
	d_2 Do not inspect	$u(C_{2,1}) = 0.1$	$u(C_{2,2}) = 0.0$
Probability		$P(\theta_1) = 0.8$	$P(\theta_2) = 0.2$

TABLE 5.3: Mafucturer Peng's utilities in Example Piston.

How should the manufacturer decide so her utilities and probabilities are considered? She should calculate her expected utility for each decision (also

referred to as mean or average utility), and see which decision gives the maximal expected utility. This is what we will do next.

We will first find the average utility for each possible manufacturer decision. For example for decision d_1, we calculate $u(C_{1,1}P(\theta_1) + u(C_{1,2})P(\theta_2)$ (see the d_1 row in Table 5.4). What decision should the manufacturer take? The manufacturer should take decision d_1 according to the maximisation of the expected utilities since d_1 gives the maximal expected utility ($0.82 > 0.80$).

	Outcome		Expected utility
	θ_1 Good part	θ_2 Poor part	$E[u\|d_i]$
d_1	$u(C_{1,1}) = 0.9$	$u(C_{1,2}) = 0.5$	$0.9 \times 0.8 + 0.5 \times 0.2 = 0.82$
d_2	$u(C_{2,1}) = 0.1$	$u(C_{2,2}) = 0.0$	$0.1 \times 0.8 + 0.0 \times 0.2 = 0.08$
$P(\theta)$	$P(\theta_1) = 0.8$	$P(\theta_2) = 0.2$	

TABLE 5.4: Manufacturer Peng's expected utilities in Piston example.

So, this brings us to the **general formulae of expected utility** (also called average utility) for the case of discrete probabilities and discrete utility, i.e. for a finite number of outcomes:

$$E[u|d_i] = \sum_{j=1}^{n} u(C_{i,j})P(\theta_i|d_i) \tag{5.3}$$

which is a weighted sum of utilities, where the weights are equal to conditional probabilities, conditioned on decisions d_i. In our Piston example, we assumed that the probability of the event does not depend on the decision:

$$P(\theta_j) = P(\theta_j|d_i)$$

Next, we choose a **decision that maximises the expected utility**. The formulae Eq 5.3 is an expected conditional utility, conditioned on d_i, hence it is a function of d_i. Our task is to find the decision d_i that gives the highest conditional expected utility. Mathematically we are solving the following problem:

$$\arg\max_{d_i} E[u|d_i] = \arg\max_{d_i} \sum_{j=1}^{n} u(C_{i,j})P(\theta_i|d_i) \tag{5.4}$$

In the Piston example, the d_i that maximises the conditional expected utility is d_1, i.e. the decision to inspect.

Caution. A comment on Piston and Makovnik Bakery examples: In the Piston example, we were not provided with the profits or monetary values of the outcomes. We were provided with the utilities, and hence we calculated

the expected utilities. In principle, it may be possible to obtain estimates of the monetary values. In the Makovnik Bakery example, we were provided with the profits in monetary values, and hence it made sense to calculate expected profits. In principle, it is possible also to obtain data on the utilities of each outcome, which would take into consideration the decision maker's preferences for risk. Maximising expected utility can lead to a different decision than the one reached using the maximisation of expected profit rule (for example, read Chapter 15 of the book by Thomas and Maurice [59], pages 642-645).

In the expected utility of the piston manufacturer, we assumed a finite number of outcomes, hence we used a sum to obtain the expected utility. What if we have utility and probability defined on an interval, i.e. continuous probability and utility? In such case, we still need to calculate the conditional expected utility, but via an integral operation:

$$E[u|d_i] = \int u(x)p(x|d_i)\,dx \tag{5.5}$$

In Eq 5.5, an example of a risk-averse utility function is $10 \times 1 - e^{-cx}$, where $c = 0.01$ is a parameter and x is wealth. A choice of a probability density function depends on the problem at hand. One frequent choice is a Normal distribution. So then we would get the following integral to evaluate:

$$E[u] = \int_{-\infty}^{\infty} (1 - e^{-cx}) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{-(\mu-x)^2}{2\sigma^2}}\,dx \tag{5.6}$$

which is equal to

$$E[u] = 1 - e^{\frac{c^2\sigma^2}{2} - c\mu} \tag{5.7}$$

Example. Stock investment. Now we introduce a generic decision table for a stock investing scenario (Table 5.5). An investor has a capital equal to C dollars. He needs to decide if to invest or not invest in a stock. There is a chance that the stock will appreciate and give a net profit of 10, so the investor's capital will be $C + 10$ dollars. There is a chance that the stock will depreciate and cause a net loss of 10, so the investor's capital will be $C - 10$ dollars.

		Outcome	
		θ_1 Stock appreciates	θ_2 Stock depreciates
Decision	d_1 Invest	$C + 10$	$C - 10$
	d_2 Leave in bank	C	C

TABLE 5.5: Stock investment example.

What decision should be taken: d_1 or d_2? This depends on the value of C, probabilities of outcomes (probability of stock appreciation, θ_1, and probability of stock depreciation, θ_2; obviously the two probabilities add up to 1) and

on the utility of the person who invests his/her money.

In next, we will assume that $C = 10$ and $P(\theta_1) = p$. We are going to read the utilities off graphs. We will assume two different investors. One investor has a constant risk aversion and we will call him or his utility function I Figures (5.4). The second investor has a decreasing risk aversion and we will call him or his utility function II (Figure 5.5). For this stock example, the values of utilities obtained from the Figures are summarised in Table 5.6.

	Risk aversion	
Pounds	Constant (I)	Decreasing (II)
20	0.181	0.435
10	0.095	0.343
0	0.000	0.000

TABLE 5.6: Stock investment example's utilities. The values for the constant risk-averse investor were obtained from Figures 6.13. The values for the decreasing risk-averse investor are from Figure 6.14

.

There are three consequences that each of the two investors can face: 0, 10 or 20, i.e. having no profit, losing 10 or gaining 20. We now know the utilities of each investor for each of the consequences (Table 5.6). To advise on a decision, we can draw a decision table for each investor (Table 5.7).

Investor I	States	
	θ_1: Stock appreciates	θ_2: Stock depreciates
d_1: Invest	0.181	0.000
d_2	0.095	0.095
Investor II	**States**	
	θ_1: Stock appreciates	θ_2: Stock depreciates
d_1: Invest	0.435	0.000
d_2: Leave in bank	0.343	0.343

TABLE 5.7: Stock investment utilities for investors with constant risk aversion (Investor I), and for investors with decreasing risk aversion (Investor II). Here we assume a value of starting capital being C=10.

Next, we want to advise on a decision. Using the decision tables (Table 5.7) we will find the expected utility for each person, for each decision.

For the constant risk-averse investor (Investor I), the expected utility is

$$E[d_1] = 0.181p + 0(1-p) = 0.181p$$
$$E[d_2] = 0.095p + 0.095(1-p) = 0.095$$

hence, the two decisions d_1 and d_2 are giving equal utility, when $0.181p =$

0.095, hence when $p = 0.52$. Therefore, he should invest when $p > 0.52$, else not invest.

For decreasing risk-averse investor (Investor II), the expected utility is

$$E[d_1] = 0.435p + 0(1-p) = 0.435p$$
$$E[d_2] = 0.343p + 0.343(1-p) = 0.343$$

hence, the two decisions are giving equal utility, when $0.435p = 0.343$, hence when $p = 0.79$. Therefore, he should invest when $p > 0.79$, else not invest.

What happens when we change C into $C = 100$? Then the decision tables for the two persons are different, they are in Table 5.8.

Investor I	States	
	θ_1: Stock appreciates	θ_2: Stock depreciates
d_1: Invest	0.667	0.593
d_2: Leave in bank	0.632	0.632
Investor II	**States**	
	θ_1: Stock appreciates	θ_2: Stock depreciates
d_1: Invest	0.673	0.644
d_2: Leave in bank	0.659	0.659

TABLE 5.8: Stock investment utilities for investors with constant risk aversion (Investor I), and for investors with decreasing risk aversion (Investor II). Here we assume a value of starting capital being C=100.

Next, we need to update the table of expected utilities, see Table 5.9, where for the person with constant risk aversion we have

$$E[d_1] = 0.667p + 0.593(1-p) = 0.074p + 0.593$$

$$E[d_2] = 0.632p + 0.632(1-p) = 0.632$$

So, for investor I, the expected utilities of the two decisions are equal, i.e. $E[d_1] = E[d_2]$ if and only if p=0.52, which is the same as when the person started with the capital C=10, but this is expected as this is a person with constant risk aversion. For investor II, the expected utilities of the two decisions are equal if and only if p=0.52, which is a decrease from 0.79 when the person started with capital C=10. So investor II with initial capital C=100 is willing to invest at a smaller probability than when the person had only C=10 initial capital, which is expected as this person has a decreasing risk aversion.

5.1.4 Probability premium

A risk-averse person will never take a *fair bet*. A fair bet is one for which the expected payoff value is zero after accounting for the cost of the bet. For

	Risk aversion	
Pounds	Constant (I)	Decreasing (II)
$E[d_1]$	0.074p+0.593	0.033p+0.644
$E[d_2]$	0.632	0.659

TABLE 5.9: Expected stock investment example's utilities. Here we assume a value of starting capital being C=100.

example, suppose I offer to pay you 2 if a fair coin lands heads, but you must ante up 1 to play. The equal loss has a larger utility loss than the equal gain has a utility gain.

A risk-averse person needs a greater chance of winning to compensate for his/her risk aversion. But how much bigger should the chance be? The difference in chances is called the *Probability Premium*, and it depends on the person's utility function and depends on how much the person owns already. The steps to calculate the chances are:

1. Let the probability that makes a bet monetarily fair be p.
2. Let the probability that makes a bet acceptable, utility-wise, be P.
3. Then we calculate the difference $P - p$. This difference is the Probability Premium. It is the increase in the probability of changing a fair bet into an acceptable bet.

Example. A £50 bet. A decision maker with decreasing risk aversion. Assume that he owns an asset of £50 and decides if to bet it all on a fair bet. What is the probability premium?

The solution follows several steps:

1. Obviously, p=0.5 is the probability that makes the bet monetarily fair.
2. Next, we need to find the probability that makes the bet acceptable utility-wise. We can proceed by following these thoughts:
   - The decision maker can end up having: 0, 50 or 100. Utility of 0, 50 and 100 is 0.000, 0.562 and 0.659 (see Figure 5.5 for the decreasing risk aversion).
   - Expected utility of bet is therefore $0.5 \times 0.659 + 0.5 \times 0.000 = 0.330$
   - Expected utility of not betting is: 1×0.562
   - Since $0.562 > 0.330$, it is therefore advised that he does not bet.
   - Need to set $0.659 \times P = 0.562 => P = 0.85$

3. Therefore, the probability premium is $P - p = 0.85–0.50 = 0.35$. The probability of winning needs to increase by at least 0.35 (from 0.5), and then the decreasing risk-averse person will be willing to bet.

5.1.5 Monetary premium

Example. Insurance. Buying (or not) insurance is the same as taking a bet. Here is a generic table for house insurance decision-making. We need to decide whether or not to buy insurance to protect ourselves against a calamity.

	States	
	θ_1: Calamity	θ_2: No calamity
d_1: Insurance	Inconvenience	Loss of premium
d_2: No insurance	Loss of house	Status quo

TABLE 5.10: Insurance example with the list of all consequences.

Next, we assume that C is the value of the house, m is the price of the insurance i.e. the premium, and h is the cost of the calamity event. We can put all these values into the table to get Table 5.11. This table is just like the investment tables before (e.g., 5.8), except this time the table is inverted: the "do something" decision is the non-risky one and the "do nothing" decision is the risky one.

	States	
	θ_1: Calamity	θ_2: No calamity
d_1: Insurance	C-m	C-m
d_2: No insurance	C-h	C

TABLE 5.11: Insurance example with the values of all consequences.

Suppose, $C = 5000$ Euros, calamity is a total loss of the house, i.e. $h = 5000$ Euros. The decision maker (here the home-owner) is decreasingly risk-averse.

We can scale the utility graph by a factor of 100. Thus we get $u(50) = 0.562$. Hence,

$$E[u_1] = u(C - m)$$
$$E[u_2] = 0.562(1 - p)$$

The break-even point of the homeowner is when the expected utilities equal each other. This happens when

$$u(C - m) = 0.562(1 - p)$$

Hence for p=0.01 we have $u(50 - m) = 0.562 \times (1 - 0.01) = 0.556$. Next, we need to use the graph backwards (5.5) to find m such that it gives a utility of 0.556. Such m is equal to 49.

Next, we work with money values:

$$5,000 - m = 5,000 \times (1 - 0.01)$$

Hence m=50 Euros.

Why is there a difference? The insurance company has much more assets than the homeowner. They might have the same utility curve, but they make their decision much further to the right on the utility curve, where the curve is almost flat. So the insurer can make money with at least a £50 premium and the homeowner can choose to insure with at most a £49.

A further note on portfolio management: It is sometimes said when investing to leave some money in safe investments and some in riskier investments. If an investment is shown to be good, shouldn't we put all our money into it? No, because of diminishing marginal utility. Can actually find the optimum amount to invest: analytically, via finding a maximum, or numerically, via trying different values.

5.2 Making decisions under imprecise risk

In this section, we look at how to find optimal decisions in a situation of total probabilistic uncertainty, i.e. we do know the probabilities.

5.2.1 General strategy when probabilities are not known

If we do not know anything about the probabilities, then we face a situation of total uncertainty about probabilities. We can still assess and risk-manage the situation through worst-case scenarios, risk transfer, and so on. For example: in 2003, reinsurance companies and banks began to issue financial instruments with returns linked to the aggregate longevity of specified populations, though the market for these instruments is still very immature (see the book Essentials of Risk Management [17], page 10).

In some life situations, it is possible that we know the probabilities to some extent. For example, we may be 95% confident that the probability of 1 million profit is between 5 and 25%. This means we are not certain what the probability is, but we have some knowledge. We will not discuss such cases here in the next sections.

Obviously, we cannot find the optimal decision by maximising the expected utility any more. This is because the expected utility calculations involve the use of probabilities. Since we do not know the probabilities, we need some alternative decision criteria, which we will learn in the next section.

5.2.2 Alternative decision criteria

In this section, we are going to discuss what happens when we do not have probabilities for the outcomes of the decisions. How do we make a decision? Or how do we recommend a decision to the decision maker? We cannot use the method of maximising the expected utility from the previous section as it requires knowledge of probabilities. Instead, a different set of criteria can be used, which are often named as *alternative decision criteria.*

We will learn the alternative decision criteria. We will also learn that such criteria are often incoherent except in certain special cases. This should not be surprising, as such criteria do not use probabilities and hence should only be used with caution and as a last resort.

Practically all economic theories about behaviour in the absence of complete information use risk tools rather than uncertainty (wrongly, of course). Furthermore, decision science has little guidance to offer managers making decisions when they have no idea about the likelihood of various states of nature occurring. This should not be too surprising, given the nebulous nature of uncertainty. We will, however, present five rather simple decision rules that can help managers make decisions under uncertainty:

- Laplace criterion
- Max-min criterion
- Max-max (Wald) criterion
- Hurwitz criterion
- Min-max regret criterion.

We explain all criteria via an example.

Example. Food warehouse. Filip runs a food warehouse. One particular item costs £40 per crate and sells for £100. Orders are made once a month, and at the end of the month, unsold items are thrown away. Past experience suggests that never more than 4 are sold. Assume we do not have any probabilities for the amounts being sold. So given the information above, we can construct the table of Filip profits (Table 5.12). Next, Filip gives us his utilities for each of the possible 25 outcomes in Table 5.13.

Monetary profits		Amount demanded				
		0	1	2	3	4
Amount bought	0	0	0	0	0	0
	1	-40	60	60	60	60
	2	-80	20	120	120	120
	3	-120	-20	80	180	180
	4	-160	-60	40	140	240

TABLE 5.12: Filip's monetary profits in food warehouse example.

Utilities		Amount demanded				
		0	1	2	3	4
Amount bought	0	0.7	0.7	0.7	0.7	0.7
	1	0.6	0.85	0.85	0.85	0.85
	2	0.45	0.75	0.92	0.92	0.92
	3	0.3	0.65	0.87	0.95	0.95
	4	0.0	0.53	0.8	0.93	1.0

TABLE 5.13: Filip's utilities in the food warehouse example.

Laplace criterion. If we do not know what the probabilities are, the Laplace criterion advises us to assume that they are equiprobable and then choose the decision that gives the maximum expected utility. It helps to organise our calculations into a table with expected utilities (see the last column in Table 5.14). We see that the decision to buy 1 crate has the highest value of expected utility. Hence, according to Laplace's criterion, we shall advise Filip to buy 1 crate.

Laplace		Amount demanded					E[ ]
		0	1	2	3	4	
Amount bought	0	0.7	0.7	0.7	0.7	0.7	0.7
	1	0.6	0.85	0.85	0.85	0.85	0.8
	2	0.45	0.75	0.92	0.92	0.92	0.792
	3	0.3	0.65	0.87	0.95	0.95	0.744
	4	0.0	0.53	0.8	0.93	1.0	0.652
Probability	P()	0.25	0.25	0.25	0.25	0.25	

TABLE 5.14: Utilities of Filip in the food warehouse example and the calculations for Laplace criterion.

Max-min criterion (also called Wald criterion). If we do not know what the probabilities are, the Max-min criterion advises to push all the probability mass onto the minimum utility for each decision (i.e. for each decision,

we multiply the minimal utility with the probability 1 and we multiply the rest of utilities with probability zero; this is the same as giving weight 1 to the minimal utility and giving weight 0 to the rest of utilities). Then we choose the decision that gives the maximum expected utility.

The Wald criterion is viewed as a pessimistic choice criterion because for each decision it looks for the worst-case utility scenario and then finds the decision that gives the best utility among the worst-case scenarios. It helps to organise our calculations into a table with expected utilities (see the last column in Table 5.15).

Max-min (Wald)		Amount demanded 0	1	2	3	4	E[ ]
Amount bought	0	0.7×1	0.7×0	0.7×0	0.7×0	0.7×0	0.7
	1	0.6×1	0.85×0	0.85×0	0.85×0	0.85×0	0.6
	2	0.45×1	0.75×0	0.92×0	0.92×0	0.92×0	0.45
	3	0.3×1	0.65×0	0.87×0	0.95×0	0.95×0	0.3
	4	0.0×1	0.53×0	0.8×0	0.93×0	1.0×0	0.0

TABLE 5.15: Filip's utilities in food warehouse example and the calculations for Max-min criterion (also called Wald criterion).

Max-max criterion. If we do not know what the probabilities are, the Max-max criterion advises pushing all the probability mass onto the maximum utility for each decision. This means that for each decision, we multiply the maximal utility with the probability 1 and we multiply the rest of the utilities with probability zero. Then we choose the decision that gives the maximal expected utility. It helps to organise our calculations into a table with expected utilities (see the last column in Table 5.16). According to the Max-max criterion, we shall advise him to buy 4 crates because such a decision gives the highest expected utility.

Max-max		Amount demanded 0	1	2	3	4	E[ ]
Amount bought	0	0.7×0	0.7×0	0.7×0	0.7×0	0.7×1	0.7
	1	0.6×0	0.85×0	0.85×0	0.85×0	0.85×1	0.85
	2	0.45×0	0.75×0	0.92×0	0.92×0	0.92×1	0.92
	3	0.3×0	0.65×0	0.87×0	0.95×0	0.95×1	0.95
	4	0.0×0	0.53×0	0.8×0	0.93×0	1.0×1	1.0

TABLE 5.16: Utilities of Filip in food warehouse example and the calculations for Max-max criterion.

Hurwicz criterion. If we do not know what the probabilities are, the Hurwicz criterion advises calculating weighted averages of the minimum and maximum utilities, where a weight α is given to the maximal utility and weight $(1-\alpha)$ is given to the minimal utility. Then we choose the decision that gives the maximal expected utility. In the food warehouse example, let us assume $\alpha = 0.75$. Then we do the calculations of the expected utilities (see the last column, in Table 5.17). According to Hurwicz's criterion with $\alpha = 0.75$ we shall advise Filip to buy 2 crates, as it maximizes the expected utility.

Hurwicz		Amount demanded					E[ ]
		0	1	2	3	4	
Amount bought	0	0.7×0.25	0.7×0	0.7×0	0.7×0	0.7×0.75	0.7
	1	0.6×0.25	0.85×0	0.85×0	0.85×0	0.85×0.75	0.788
	2	0.45×0.25	0.75×0	0.92×0	0.92×0	0.92×0.75	0.803
	3	0.3×0.25	0.65×0	0.87×0	0.95×0	0.95×0.75	0.788
	4	0.0×0.25	0.53×0	0.8×0	0.93×0	1.0×0.75	0.75

TABLE 5.17: Filip's utilities in food warehouse example and the calculations for Hurwicz criterion.

Min-max criterion (also called Regret criterion, or Minimax regret). The next criterion is similar to the Max-min (Wald) criterion. If we do not know what the probabilities are, the Min-max criterion advises the following two steps. The first step is to calculate the table of regrets. We do this by calculating the loss for each decision and outcome. We do it by taking the difference: the maximum utility for that outcome minus the utility. This is the regret, the opportunity loss through having made the wrong decision, i.e., the loss for the outcome. The second step is to choose the decision that minimises the maximal regret. This means we choose the decision that minimises the maximum loss for a particular outcome.

Using the table of utilities (Table 5.13), we create a table of regrets (Table 5.18. For example, for "amount bought 3" and "amount demanded 2" the regret is 0.05 because Filip bought 3, and he regrets he did not buy 2. If he bought 2 his utility would be 0.92 after he would sell them. But he bought 3, and his utility is only 0.87 since he can only sell 2. So the value of the regret is 0.92-0.87=0.05. Another example, for "amount bought 2" and "amount demanded 4", the regret is 0.08. This is because Filip bought 2, and he regrets he did not buy 4. If he bought 4 his utility would be 1.0 after he would sell them all. But he bought 2 and can sell 2 only, so his utility is only 0.92. So the value of the regret is 1.0-0.92=0.08.

Then we put all probability mass (hence weight) onto the maximum regret. This means that for each decision, we multiply the maximum regret with probability 1, and use zero probability elsewhere (Table 5.19). According to

Regrets		Amount demanded				
		0	1	2	3	4
Amount bought	0	0.00	0.15	0.22	0.25	0.30
	1	0.10	0.00	0.07	0.10	0.15
	2	0.20	0.10	0.00	0.03	0.08
	3	0.40	0.20	0.05	0.00	0.05
	4	0.70	0.32	0.12	0.02	0.00

TABLE 5.18: Filip's regrets in food warehouse example. A regret is the difference between the maximal utility of an outcome and the utility of the decision at that outcome.

the Min-max regret criterion, we shall advise him to buy 1 crate because it minimizes the expected regret.

Regret criterion		Amount demanded					E[ ]
		0	1	2	3	4	
Amount bought	0	0.00×0	0.15×0	0.22×0	0.25×0	0.30×1	0.30
	1	0.10×0	0.00×0	0.07×0	0.10×0	0.15×1	0.15
	2	0.20×1	0.10×0	0.00×0	0.03×0	0.08×0	0.25
	3	0.40×1	0.20×0	0.05×0	0.00×0	0.05×0	0.40
	4	0.70×1	0.32×0	0.12×0	0.02×0	0.00×0	0.70

TABLE 5.19: Filip's regrets in food warehouse example and Regret criterion. For each decision, we put all probability mass (hence weight 1) onto the maximal regret, and we put zero probability mass (hence weight 0) on all other regrets.

5.3 Tips to think and act like a risk expert

Here we will give tips and tricks, for which we will bring two case studies.

5.3.1 Be aware of the criticism of the alternative criteria

The alternative criteria can often lead to inconsistency. We explain by a simple example. Imagine we have £1. We can bet that £1 for a payoff of £u. Let d1 be a decision to bet, and d2 is not betting. Let θ_1 be a scenario when we win the payoff, and θ_2 is a scenario when we do not win. We are going to use the Max-min criterion (see Figure 5.20). In this case, d_2 will be always picked as the

optimal decision, even if $u = 1,000,000$, which is an illogical recommendation.

	θ_1	θ_2	E[]
d_1	u	0	0
d_2	1	1	1

TABLE 5.20: Table of utilities to illustrate criticism of Max-min criteria.

Min-max avoids some of the problems of Max-min (Wald) but can still be inconsistent. We show on a simple example. In Table 5.21 decision d_1 will be chosen, according to the Min-max.

	Utilities		Loses		
	θ_1	θ_2	θ_1	θ_2	E[ ]
d_1	8	0	0	4	4
d_2	2	4	6	0	6

TABLE 5.21: Table of two decisions to illustrate Min-max. It leads to choosing the optimal decision d_1.

Next, what about if we add an extra decision into Table 5.21, thus getting a new Table 5.22? According to the Min-max criterion, the decision d_2 is the best decision. This makes no sense: when comparing d_1 and d_2, we choose d_1 as the best decision. But then when we add extra decision d_3 we find d_2 to be the best, i.e. we consider d_2 as better than d_1. This is one of the critiques of the Min-max criteria, that sometimes it leads to an illogical choice of the best decision.

	Utilities		Loses		
	θ_1	θ_2	θ_1	θ_2	E[ ]
d_1	8	0	0	4	4
d_2	2	4	6	0	6
d_3	1	7	7	0	7

TABLE 5.22: Criticism of Min-max criteria, when adding another decision to what we had in Table 5.12. Min-max leads to optimal decision d_2, i.e. it suggests that now d_1 is worse than d_2.

5.4 Summary

We learned in this chapter:

1. How to identify if the decision maker is risk averse or risk loving, by looking at his/her plot of the utility function.
2. How to find an optimal decision via maximising the expected utility.
3. How to calculate the probability premium.
4. How to use several alternative criteria that can be used in situations where there is no way to get reasonable estimates of the probabilities of scenarios. We also learned that often alternative decision criteria are inconsistent, and we looked at several examples.

5.5 Further reading

In our chapter, we were inspired by several resources. Here we mention them, as well as recommend further resources:

1. For more insights about decisions under imprecise risk in applied in economics, we recommend the book *Managerial economics, Foundations of business analysis and strategy* by Christopher R. Thomas and S. Charles Maurice [59] (especially their Chapter 15).
2. For a comprehensive reading on decision analysis, we recommend the book *Foundations of decision analysis* by Ronald A. Howard and Ali E. Abbas [33].
3. For more insight into utility functions in finance, we recommend the book by Rob Kaas, Marc Goovaerts, Jan Dhaene and Michel Denuit *Modern actuarial risk theory using R* [35]
4. For further insight into cognitive psychology and behavioural economics of how humans make decisions under imprecise risk, we recommend starting with the groundbreaking work of Tversky and Kahneman [60]. They made discoveries of systematic human cognitive bias and handling of risk. In October 2022, Kahneman was awarded the Nobel Memorial Prize in Economic Sciences for their work in applying psychological insights to economic theory, particularly in the areas of judgment and decision-making under uncertainty. He has done this work with Amos Tversky, who died in 1996. While Tversky was acknowledged in the announcement, the Royal Swedish Academy of Sciences does not award prizes posthumously.

5. It is interesting to read the paper by Volz and Gigerenzer [62] where they point to an MRI study they conducted showing that an additional area of the brain is involved when humans are making decisions under imprecise risk (they call it decision under uncertainty) as opposed to under precise risk (they call it decision under risk).

6. We also recommend reading the latest research about uncertainty in the Journal of Approximate Reasoning.

5.6 Exercises

Solve the following exercises by using pen, paper and calculator.

1. [**Purpose: Practicing how to find probability premium.**] A decision maker with decreasing risk aversion depicted in figure from the lecture has assets of £20 and contemplates a gamble which may win her £20 or lose her £10. It is therefore actuarially fair if the chance of winning is 1/3. Determine her probability premium. Do the same for the constant risk-averse decision maker (use the Figure from the lecture. Find the probability premium for the first decision maker when her assets are £200.

2. [**Purpose: Practicing how to tell where the utility function is risk averse, and practising how to find probability premium.**] A decision maker has the following utilities for money:

Money	0	100	200	300	400	500	600	700
Utility	0	0.8	2.6	5.4	10.0	13.6	14.8	15.7

   - By sketching a graph, shows that she is risk-averse for assets above about £400 but not below this amount.
   - Consider a gamble that might win or lose £100, first when her assets are £200, and then when they are £500. In each case, determine the probability premium.

3. [**Purpose: Practicing alternative decision criteria.**] Assume the following utility table:

Money	θ_1	θ_2	θ_3
d_1	5	-1	2
d_2	0	2	3
d_3	-5	3	0

   Determine the decisions that would be reached under the following criteria:

   - Wald,
   - Max-max,
   - Laplace,
   - Hurwicz with $\alpha = 0.7$,
   - Minimax criteria.

4. [**Purpose: Practicing alternative decision criteria.**] Given the following utility table:

Money	θ_1	θ_2	θ_3
d_1	110	30	10
d_2	80	50	100
d_3	40	115	60

   (a) Determine the decisions that would be reached under the following criteria: Wald, Max-max, Laplace, Hurwicz with $\alpha = 0.6$, Minimax criteria.

   (b) Now assume that d_1 is no longer available. Recalculate under this new table and comment on the answers you got.

5. [**Purpose: Practicing alternative decision criteria.**] An electric store is considering extending the range of items which it intends to stock. The utility table is shown below. Calculate the optimal decisions according to Laplace, Max-min (Wald), Max-max and Minimax (Regret) criteria.

Utilities		Demanded		
		High	Average	Low
Amount bought	Microwaves	20	40	30
	Home security	80	70	-10
	Satelite TVs	90	10	-20
	Large TVs	10	100	40

6

Decision trees

Monika Kovacova

Gabriela Czanner

CONTENTS

6.1 Decision tree 216
6.1.1 Typical decision tree format 216
6.1.2 The five steps of building a decision tree 216
6.2 When probabilities are in a convenient format 218
6.2.1 Example: Manufacturer Mahiro 218
6.2.2 Solution step by step 219
6.3 When probabilities are *not* in a convenient format 225
6.3.1 Example: Investor Iveta 225
6.3.2 Solution step by step 227
6.4 Tips to think and act like a risk expert 235
6.4.1 Sensitivity analysis for decision trees 236
6.4.2 Value of information 237
6.5 Summary 237
6.6 Further reading 238
6.7 Exercises 239

In this chapter, we are looking again at how to make decisions under risk, but now we look into how to decide in sequential decision problems, and for that, we will use Decision Trees.

In Section 5.1.2 we worked with tables of utilities, and we learned how to make a decision. We assumed that only one decision is made, e.g. how much stock to buy. However, in reality, we often make several decisions that are not isolated; rather, they are done in sequence, where the later decisions are affected by the outcome of previous decisions. Making several decisions in a sequence is also called a **multistage decisions**, **sequential decisions** or **phased decisions**. We have a multistage decision problem if:

- we have to make multiple decisions, one after the other;

- and the decision taken later will depend on the outcomes of the earlier decisions.

When making two or more decisions in sequence, it is challenging to represent the decisions and outcomes in a table as we did in Chapter 5. We will see that it is much easier to represent the problem using a decision tree. Decision trees provide a visual display of the sequential decision processes and subsequent consequences. Decision trees can help in organising the computational work and hence in arriving at a recommended most beneficial decision. It also helps in communication with stakeholders, so they see all information displayed on one figure and all outcomes.

Learning objectives

- Learn what decision trees are.
- Learn how to draw a decision tree and do all computations to derive the best sequence of decisions.
- Explore how to give a recommendation to a stakeholder who needs to make several multistage decisions.

6.1 Decision tree

Here we explain the components of the decision tree and the five steps of building a decision tree.

6.1.1 Typical decision tree format

A typical tree normally goes from left to right and has two types of nodes (see Figure 6.1).

- Decision node: this is where the decision maker is in control. He/she makes a decision.
- Random node: this is where things are out of the control of the decision maker. This is where we put the value of the outcome that could happen. This node is also called chance node or probability node

A decision tree is a single-decision tree if it includes only one decision node along any given path. A decision tree is a sequential decision-making problem tree if there is at least one path that contains at least two decision nodes.

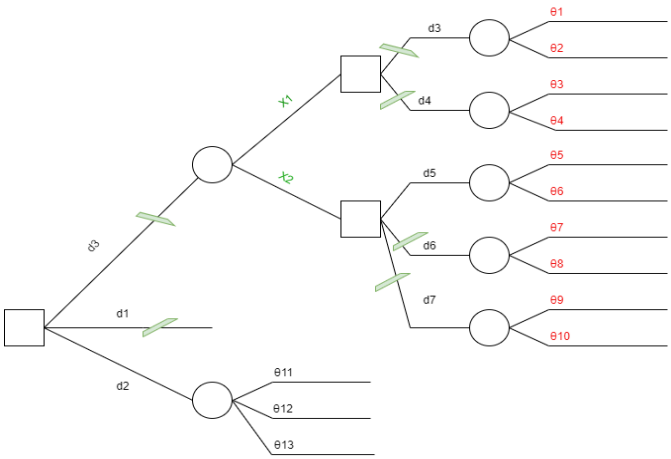


FIGURE 6.1: Decision tree format. Squares = Decisions nodes. Circles = random (chance) nodes. The last node, where no decision and no chance happens, is called the final points or leaves. The lines in the tree are called branches.

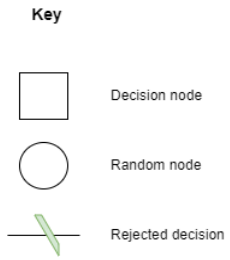


FIGURE 6.2: Decision tree notation.

6.1.2 The five steps of building a decision tree

It is important to note that one of the most difficult tasks of using decision trees is drawing them from a written description of the problem. Once that has been accomplished, deriving the recommendation for a decision is relatively straightforward. It might help when drawing a decision tree if we pretend we are the relevant stakeholder and ask ourselves, "What happens next?" at each node of the tree as we draw it. Here a relevant stakeholder is the person who needs to make a decision.

We will build the decision tree and do all the calculations in 5 steps. In Steps 1 to 3, we move from left to right; in Step 4, we move from right to left. In Step 5, we make a recommendation. The five steps are:

- Step 1: Create a decision tree structure from left to right in chronologi-

cal order. For each decision node, we need to attach a description of the branches coming from such nodes, we also attach costs of such decisions.

- Step 2: For random nodes, we attach probabilities to branches coming from such nodes. These probabilities will always be conditioned on what has occurred before.
- Step 3: Attach utilities to the terminal points.
- Step 4: Going from right to left, take the expectations of the utilities at random nodes and maximise utilities at decision nodes.
- Step 5: Make recommendations about decisions.

6.2 When probabilities are in a convenient format

A decision tree helps to find the optimal sequence of decisions in a risk situation. The risk is expressed in probabilities. Sometimes the probabilities are provided to use conveniently, so all further calculations are straightforward. We explain this in one example: explaining all five steps, from building the tree to finding optimal decisions.

6.2.1 Example: Manufacturer Mahiro

Manufacturer Mahiro faces a decision concerning a product (code-named M997) developed by one of his research laboratories. He has to decide whether to proceed to test market M997 or whether to drop it completely. Here is some information:

- It is estimated that test marketing will cost £100,000.
- Past experience indicates that only 30% of products succeed in the test market.
- If M997 is successful at the test market stage, the company faces a further decision relating to the size of the plant set up to produce M997.
- A small plant will cost £150,000 to build and produce 2,000 units annually.
- A large plant will cost £250,000 to build and produce 4,000 units annually.
- The marketing department has estimated that there is a 40% chance that the competition will respond with a similar product and that the price per unit sold (in £) will be as follows (assuming all production sold):

Example M997 information on competition

	Large plant	**Small plant**
Competition responds	20	35
Competition do not respond	50	65

TABLE 6.1: Manufacturer Mahiro's information on competition

- We assume that the life of the market for M997 is estimated to be seven years.
- Yearly running costs are £50,000 (both plant sizes).
- We will assume that the utility is directly measured via monetary profits.

Should the company go ahead and test the market for product M997? This is a multistage decision problem, and we will solve this problem by building a decision tree in the following sections. Although this example is somewhat simplified, it represents the type of decision that often has to be made about new products. In particular, you should note that we cannot separate the test market decision from any decisions about the future profitability (if any) of the product if test marketing is successful.

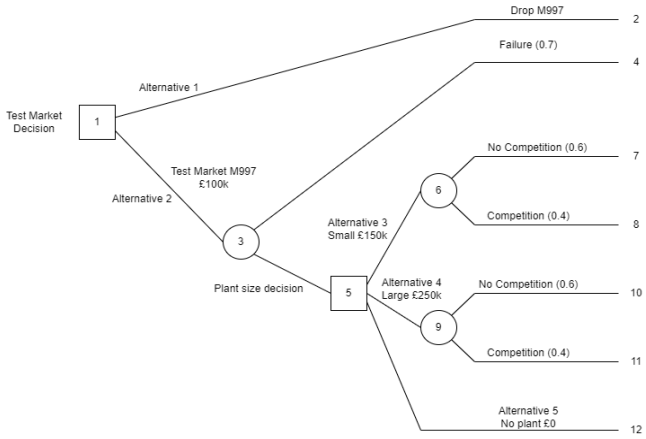


FIGURE 6.3: Manufacturer Mahiro's decision tree, after we did steps 1 and 2.

6.2.2 Solution step by step

Steps 1 and 2 Creation of the decision tree structure and attaching the probabilities to random nodes. Figure 6.3 shows the decision tree after completing Steps 1 and 2 for the M997 problem. We must always add a "do nothing" alternative at every decision node. So, for example, there is a "no plant" alternative at the plant size decision node. This is necessary because it may not be profitable to build any plant (large or small) even if the product is successful in the test market. In any decision tree, we must include all possible alternatives (in action notes). It is very common in decision tree problems to find that at decision nodes, there is a "do nothing" alternative which is an implicit decision which can be taken.

We also need to ensure a unique path in the tree from the initial node to each terminal node. This is very important. This must be assured when drawing the branches and nodes.

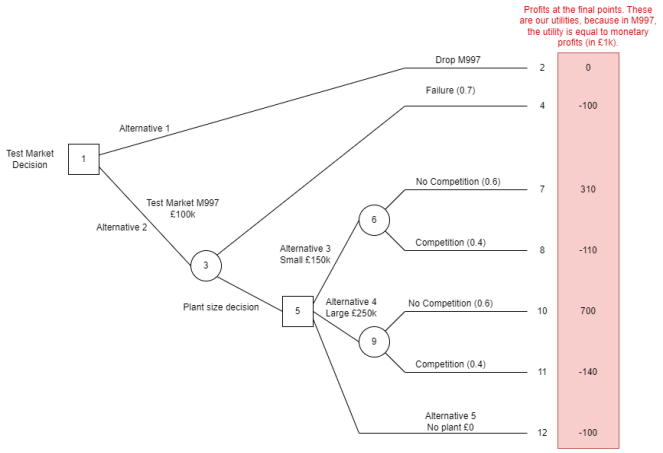


FIGURE 6.4: Manufacturer Mahiro's decision tree, after we did steps 1-3.

Step 3: Attaching the utilities to the terminal points. To calculate the utility for terminal point 2 is simple:

- We drop M997 (see Alternative 1 in Figure 6.3).
- Total Revenue = 0
- Total Cost = 0
- Hence Total Profit = 0

Note that we are ignoring in every case any money already spent on developing M997 since that is a sunk cost, i.e., a cost that does not matter for our future decisions; such cost cannot be altered, so it has no part in deciding future decisions. Calculating the utility of terminal point 7 is more complex. It helps to look at the decision tree to remind ourselves of the path from the first decision node to the terminal point 7 (Figure 6.3). Then going from the left to the right, we see the costs and profits along the path: We test market M997 (cost £100k), find it successful, build a small plant (cost £150k) and find we are without competition (revenue for seven years at 2,000 units a year at £65 per unit = £910k):

- Total Revenue = 910k
- Total Cost = 100k + 150k + (7 x 50k)[running costs] = 600k
- Total Profit = 910k - 600k = 310k

By doing the same calculations, we get the profits at all the terminal points, see Table 6.2.

Terminal node	**Profit**
2	0
4	-100
7	310
8	-110
10	700
11	-140
12	-100

TABLE 6.2: Manufacturer Mahiro's product M997 decision tree's terminal node profits. These profits are those to be attached to the corresponding final points of the decision tree. Note, in this example, it was agreed that the utility is directly equal to the monetary value, i.e. the monetary profit.

Step 4: Taking expectations at the random nodes and maximising utilities at decision nodes, while going from the right to the left. So far, we have ignored the probabilities in the M997 example. We assigned probabilities to the tree (in step 2), but we have not used them yet. We use them in the next step: in step 4. We work from the right-hand side of the tree back to the left-hand side. Going from right to left:

- We take the expectations of the utilities at random nodes. The random nodes are those circled: nodes 6, 3 and 9.
- We maximise utilities at decision nodes. The decision nodes are those squared: nodes 1 and 5.

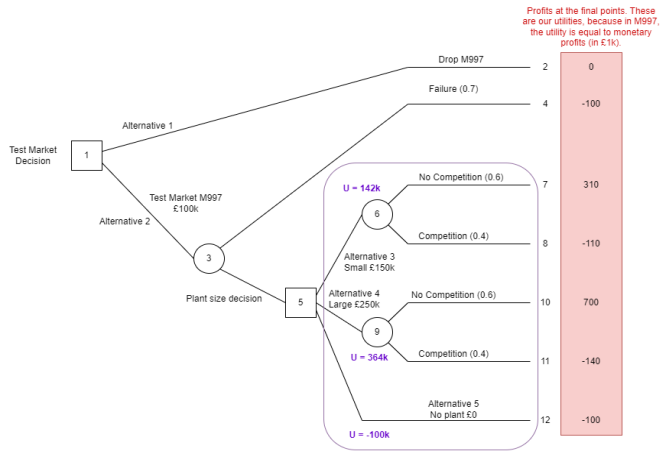


FIGURE 6.5: Manufacturer Mahiro's decision tree, processing nodes 6 and 9. Steps 1-3 were done; now as part of step 4 we are processing nodes 6 and 9. The value U=142k is EMV for node 6, the value U=364 is EMV for node 9 the value U=-100k is EMV for final node 12. The calculations are in the text.

The order in which we do the calculations is crucial: we proceed from right to left. So we will do the calculations in this order: 6, then 9 (or 9 and then 6), then 5, then 3, and finally 1.

Next, we consider the chance node 6. Since it is a chance node, we need to calculate its expected utility. Since, in this example, the utility is equal to the monetary value (, the company decided that their utility is directly measured via monetary profit), we need to find the expected profit. The work is shown in the following paragraphs. Node 6 has branches to terminal points 7 and 8. The branch to terminal point 7 occurs with probability 0.6 and has a total profit of £310k, while the branch to terminal point 8 occurs with probability 0.4 and has a total profit of -£110k. The expected monetary value (EMV) of this chance node number 6 is, therefore, given by:

$$0.6 \times 310 + 0.4 \times (-110) = 142(k)$$

Essentially this value represents the expected (or average) profit from this chance node: 60% of the time, we get £310k and 40% of the time, we get -£110k, so on average we get £142k. Then we can put the utility into Figure 6.5.

Next, we consider the chance node 9. By analogy with chance node 6, it can be shown that chance node 9 has the expected utility of 364 (£k). Then we can put the utility into Figure 6.5.

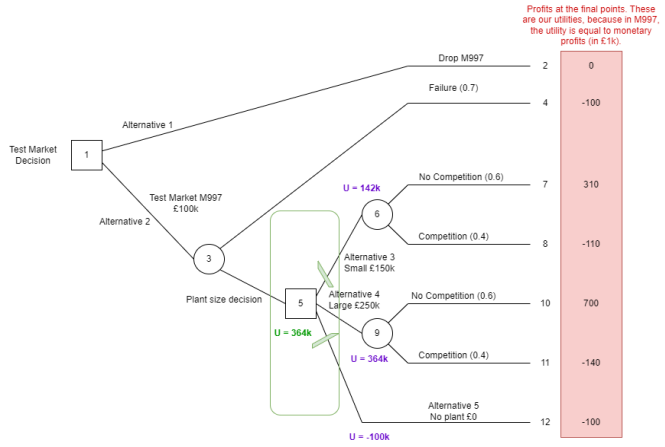


FIGURE 6.6: Manufacturer Mahiro's decision tree, processing node 5 as part of step 4. Node 5 is a decision node with a utility value of U=364 (£k).

Next, we consider the decision node 5. Since it is a decision node, we need to calculate and maximise the utility of this node. The work is as follows. At the plant decision node 5 we have three alternatives:

- Alternative 3: build a small plant, giving U = EMV = 142k
- Alternative 4: build a large plant, giving U = EMV = 364k
- Alternative 5: build no plant, giving U = EMV = -100k

It is clear that if the decision maker ends up being in the decision node 5 Alternative 4 is the most attractive to take. This is because Alternative 4 offers the highest expected monetary value (remember that in this example we measure the utility directly via money). Thus we can discard the other alternatives, giving the revised decision tree Figure 6.6.

Next, we can continue in the process for node 3. Since this is a random node, we need to calculate its expected utility. Since, in this example, the utility is equal to the monetary value (simply, the company decided that their utility is directly measured via monetary profit), we need to find the expected profit. The EMV for chance node 3 representing whether M997 is a success in the test market or not is given by:

$$0.3 \times 364 + 0.7 \times (-100) = 39.2(k)$$

The revised decision tree is shown in Figure 6.7.

Step 5: Making a recommendation for decisions. Finally, we move

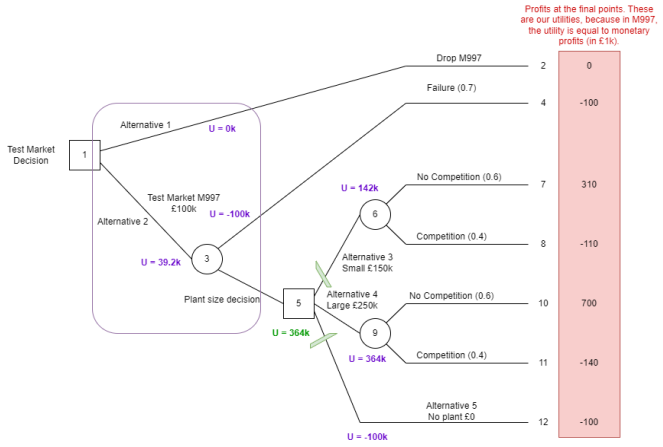


FIGURE 6.7: Manufacturer Mahiro's decision tree, processing node 3 as part of step 4. Node 3 is a random node, and it has a utility value of U=39.2 (£k).

to node 1. It is a decision node. So here, we need to maximise the utility. Hence the decision node (node 1) represents whether to test market M997 or not; we have the two alternatives (see also Figure 6.7):

- Alternative 1: drop M997, this yields U=EMV=0
- Alternative 2: test market M997, this yields U=EMV=39.2k

It is clear that, in monetary terms, Alternative 2 is preferable, and so we should advise the decision maker to test the market for M997. The updated (and final) decision tree is in (Figure 6.8).

The recommendations for the M997 example are:

- We should test market M997 and this decision has an expected monetary value (EMV) of £39.2k
- If M997 is successful in the test market then we anticipate, at this stage, building a large plant (recall the alternative we chose at the decision node relating to the size of the plant to build).
- However, it is clear that in real life we will review this once test marketing has been completed.
- The worst possible outcome (-£140) corresponds to terminal point 11. It is the downside of the decision to test market M997.
- The best possible outcome (£700) corresponds to terminal point 10. It is the upside of the decision to test market M997.

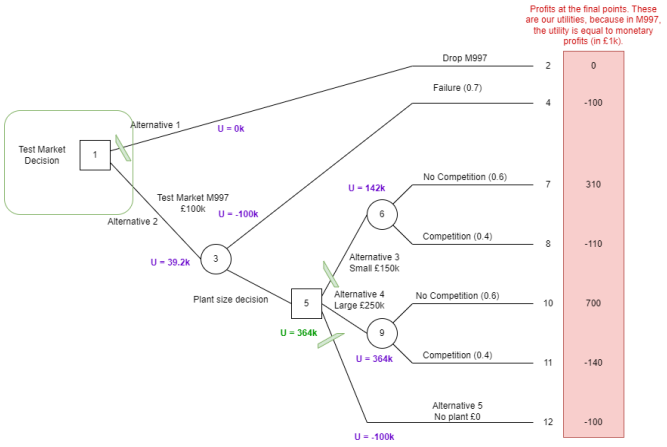


FIGURE 6.8: Manufacturer Mahiro's decision tree, processing node 1 as part of step 4. Node 1 is a decision node and has a utility value of U=39.2 (£k).

6.3 When probabilities are *not* in a convenient format

In the previous section, we did built a decision tree in Example M997, while all probabilities were provided in a convenient format. We did not need to do any further calculations on probabilities. We simply put all probabilities into the tree.

The company did have to make two decisions, where the second decision depended on the outcome of the first decision. We also assumed that the utility is directly given by the amount of money lost or gained (i.e. U(money)=money). In other words, we assumed that for the decision maker, the given utility values may be interpreted directly as £s. we also learned how to incorporate costs in our analysis.

Here, we continue with decision trees, where the probabilities are not in a convenient format. We will do one example and again we will assume that the utility is directly given by the amount of money lost or gained (i.e. U(money)=money, i.e. the utility is equal to the capital value). However, we will not be given the probability for the random branches in the required format. We will have to do some extra calculations and for that, we will use the law of total probability and Bayes Theorem (Section 2.3.7). We will also look at how the choice of the optimal decision depends on costs.

6.3.1 Example: Investor Iveta

We start with a single-stage decision problem example. Investor Iveta has £5000 capital. She can decide whether to invest or not.

- If investment is good (θ_1) she will gain £100
- If investment is bad (θ_2) she will lose £100
- $P(\theta_1) = 0.6, P(\theta_2) = 0.4$
- We assume that the utility is directly given by the amount of money lost or gained.

	State	
Decision	θ_1	θ_2
d_1: Invest	5100	4900
d_2: Leave	5000	5000
Probability	0.6	0.4

TABLE 6.3: Investor Iveta's profits. We assume that the utility is directly given by the amount of money lost or gained. So this can also be called the table of utilities.

If this (Table 6.3) is all the information that the investor uses, and if the investor's utility is equal to the capital value, what decision should he/she do? The investor can calculate the average or expected utility table because this is a single-stage decision problem (Table 6.4) or alternatively she can visualise it on a decision tree (Figure 6.9). According to Table 6.4, the investor should invest i.e. should decide d_1 as it has the highest expected monetary value (expected utility).

	State		
Decision	θ_1	θ_2	**Expected utility**
d_1: Invest	5100	4900	$0.6 \times 5100 + 0.4$x$4900 = 5020$
d_2: Leave	5000	5000	$0.6 \times 5000 + 0.4$x$5000 = 5000$
Probability	0.6	0.4	

TABLE 6.4: Investor Iveta's expected utilities. In this example, we assume that the investor's utility is equal to the monetary profits.

Next, we additionally assume that the investor Iveta could use a broker, at a cost of f pounds, who will advise on an investment. Here are the further assumptions:

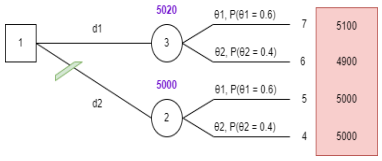


FIGURE 6.9: Investor Iveta's decision tree. Here we need to solve a single-stage decision problem: this tree represents the payoff decision Table 6.4.

- The broker can advise to invest (we will denote it X_1) or to leave (which we will denote as X_2).
- The broker is a bit better at getting winners than losers, specifically:

$$P(X_1|\theta_1) = 0.8, P(X_2|\theta_2) = 0.7$$

What do these two above probabilities mean? The $P(X_1|\theta_1)$ is the probability of advising to invest (i.e. X_1) given that the investment is good (θ_1). In 80% of good investments, the broker will correctly advise to invest, and in 20% of good investments, the broker will mistakenly advise not to invest. The second term is the probability of advising to leave (i.e. X_2) given that the investment is bad (θ_2). Of course, we do not know if the investment is good or bad until we wait to see the return. However, we know that in 70% of all bad investments, the broker will correctly advise not to invest, and in 30% of bad investments, the broker with mistakenly advise to invest.

What course of action should we recommend and why? What is the most rational thing to do for the investor, given all the information above? This is a *sequential decision-making problem*: first, the investor needs to decide if to consult a broker, and then she needs to decide if to invest or not, while her second decision depends on the advice of the broker. To advise her, we cannot use utility tables, as those are suitable for situations where a single decision is to be made. In a sequential decision-making problem, we need to construct a Decision tree and use it to find the optimal sequence of two decisions.

Note that in this example, we assume that the utility is directly given by the amount of money lost or gained.

6.3.2 Solution step by step

Step 1: Create a decision tree structure from left to right in chronological order. We plot the three from left to right, including all the random and decision nodes, as well as terminal nodes. We take the tree from Figure 6.9 and add the branches that represent the decision to ask a broker for advice, thus yielding the tree in Figure 6.10. For each decision node, we need to

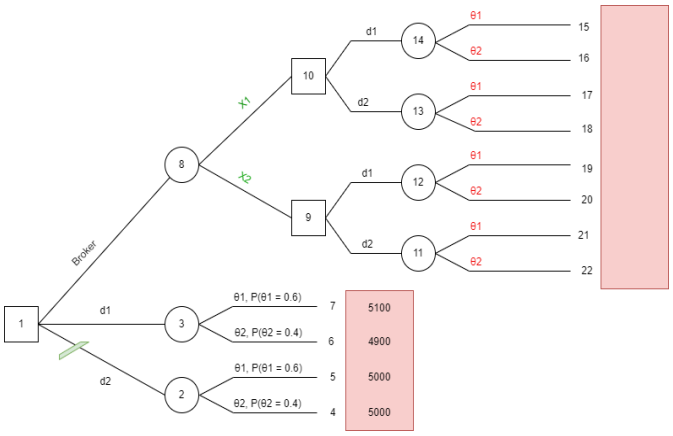


FIGURE 6.10: Investor Iveta's decision tree for the two sequential decision problems after Step 1 has been completed.

attach a description of the branches coming from such nodes, we also attach costs of such decisions.

Step 2: For random nodes, we attach probabilities to branches coming from such nodes. We need to attach the probabilities to branches coming from random nodes. These probabilities are always conditioned on what has occurred before. This step is NOT easy in the investor example. Why not easy? Because we are not told the probabilities at the branches coming from random nodes (circles). We need to calculate those probabilities. We will start by calculating the probability on the right side of the tree.

What is the probability of being on the branch from node 14 to node 15, in Figure 6.11? What is the probability of that we need to calculate? Below is a set of probabilities, but only one is correct. Which of these probabilities do we need to calculate?

- $P(\theta_1|d_1)$
- $P(\theta_1)$
- $P(d_1)$
- $P(X_1)$
- $P(X_1|\theta_1)$
- $P(\theta_1|X_1)$

Here is a hint: The node 14 is a random node. Imagine, the investor is sitting on that random node and she is asking herself: "What is my best guess

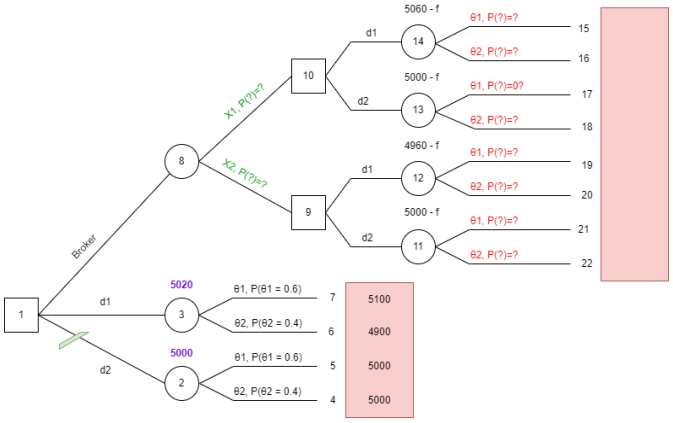


FIGURE 6.11: Investor Iveta's decision tree, after step 1 has been completed and after we added all the information for the subtree on the left. The subtree on the left has all the information (the probabilities and payoffs) because we calculated it previously. Note: θ_1 = investment is good, θ_2 = investment is bad, d_1 = decision to invest (decision of the investor), d_2 = decision not to invest, X_1 = advice from broker to invest (this advice from the broker is random to the investor), X_2 = advice from the broker not to invest.

of the probability of ending up on node 15? What information do I have to make that guess? I paid the broker and he advised me to invest and then I decided to invest. So the information I have is that the investor advised me to invest. My initial guess about the goodness of the investment did alter after I met the broker. What is now my belief about the investment to be good? In other words, what is now my probability estimate of the investment to be good?" See also the subjective approach to probability that we discussed in Chapter 2.

The correct answer is: Iveta's degree of belief of the investment to be good is $P(\theta_1|X1)$. It is a conditional probability. Her belief about investment changed after she spoke to the broker. The probability is conditioned on what has occurred before: on her talking to the broker.

Are you able to calculate all the probabilities on the branches coming from nodes 14, 13, 12 and 11? The solutions are on Figure 6.13.

The decision tree Figure 6.11 shows another probability that we need to calculate: from node 8 to node 10. What is the probability that we need to calculate? Here is a list of options. Only one is correct. Which is correct?

- $P(\theta_1|d_1)$

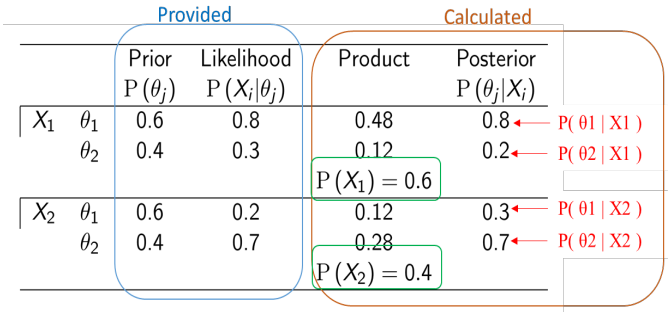


FIGURE 6.12: Investor Iveta's probabilities calculation.

- $P(\theta_1)$
- $P(d_1)$
- $P(X_1)$
- $P(X_1|\theta_1)$
- $P(\theta_1|X_1)$

Here is a hint. The node 8 is a random node. Imagine the investor is sitting on that random node 8 and she is asking herself: "What is my best guess of the probability of ending up on node 14? What information do I have to make that guess? I paid the broker, but he still needs to advise. So the information I have is the same as at the beginning. My belief about the broker advising 'invest' is the same as at the beginning, i.e. nothing has altered my belief. What is now my belief about the broker advising me to invest? In other words, what is now my probability estimate of the broker advising me to invest?

The correct answer is: The probability that we need is $P(X1)$, i.e. the probability that the investor is advised by the broker to invest. Why? This probability is always to be conditioned on what has occurred before (as the hint says), but nothing has happened before that would give the investor useful information. Alternatively, we look at it like this: What happened before node 8? The investor decided to ask the broker for advice. So she can calculate a conditional probability P(broker advises to invest|investor asked for advice). However, in this moment, P(broker advises to invest|investor asked for advice) = P(broker advises to invest) because the broker's advice does not depend on whether he is asked for an opinion or not. So, here, at this branch, the conditional probability equals to the marginal probability (Chapter 2).

Before we do calculations, it is good to remind ourselves what information we have. We know this:

- $P(X_i|\theta_j)$ = Prob of being advised X_i given that the true state θ_j.
  - $P(X_1|\theta_1)$ is the probability of being advised to invest given that the investment is good. This is a conditional probability.
  - $P(X_2|\theta_2)$ is the probability of being advised not to invest given that the investment is not good. This is, again, a conditional probability.
  - Analogically, $P(X_2|\theta_1)$ is the probability of being advised not to invest, given that the investment is good. This is, again, a conditional probability.
  - $P(X_1|\theta_2)$ is the probability of being advised to invest, given that the investment is not good. This is, again, a conditional probability.
- $P(\theta_j)$ = probability of the state θ_j.
  - $P(\theta_1)$ is the probability of the investment being good. This is a marginal probability.
  - $P(\theta_2)$ is the probability of the investment being bad. This is a marginal probability.

However, we need to get these probabilities:

- $P(\theta_j|X_i)$ = Prob of the trues state θ_j given that the investor is advised X_i.
  - For example, $P(\theta_1|X_2)$ = probability of the investment is good given that the broker advises not to invest.
  - We need four values: $P(\theta_1|X_1)$, $P(\theta_1|X_2)$, $P(\theta_2|X_1)$ and $P(\theta_2|X_2)$.
- $P(X_i)$, $i = 1, 2$ is the probability of advice.
  - $P(X_1)$ = probability of being advised to invest. This is a marginal probability.
  - $P(X_2)$ = probability of being advised to not invest. This is a marginal probability.

We recognise that the probabilities that we need are different from what we have: we know $P(X_i|\theta_j)$, and we need $P(\theta_j|X_i)$. We will employ Bayes' rule to calculate the conditional probabilities:

$$P(\theta_j|X_i) = \frac{P(X_i|\theta_j)P(\theta_j)}{P(X_i)} \tag{6.1}$$

where $i = 1, 2$ and $j = 1, 2$. Also we know $P(\theta_j)$ and we need $P(X_i)$. We will use the law of total probability to get the marginal probabilities that we need (see Chapter 2). The probabilities that we have are summarised in Table 6.5. This is our starting point.

		Prior $P(\theta_j)$	**Likelihood** $P(X_i\|\theta_j)$
X_1	θ_1	0.6	0.8
	θ_2	0.4	0.3
X_2	θ_1	0.6	0.2
	θ_2	0.4	0.7

TABLE 6.5: Investor Iveta's provided probabilities.

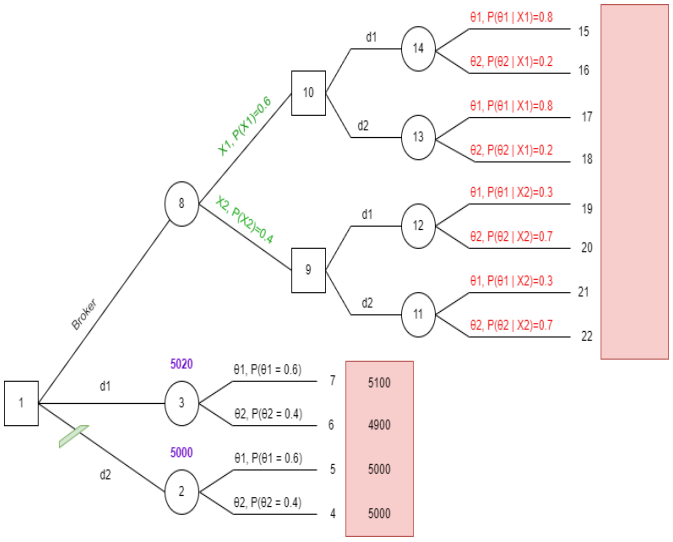


FIGURE 6.13: Investor's decision tree after Steps 1 and 2 have been completed. All the calculated probabilities are shown.

Table 6.5 summarises all provided probabilities. We note that the probability that the broker advises X_1 strategy, given that the investment is θ_2, is

$$P(X_1|\theta_2) = 1 - P(X_2|\theta_2) = 1–0.7 = 0.3$$

Similarly (from Table 6.5) the probability that the broker advises $X2$ strategy given that the investment is θ_1 is

$$P(X_2|\theta_1) = 1 - P(X_1|\theta_1) = 1–0.8 = 0.2.$$

Next, using the law of the total probability

$$P(X_1) = P(X_1|\theta_1)P(\theta_1)+P(X_1|\theta_2)P(\theta_2) = 0.8\times0.6+0.3\times0.4 = 0.48+0.12 = 0.6$$

$$P(X_2) = P(X_2|\theta_1)P(\theta_1)+P(X_2|\theta_2)P(\theta_2) = 0.6\times0.2+0.4\times0.7 = 0.12+0.28 = 0.4$$

or simply

$$P(X_2) = 1 - P(X_1) = 1 - 0.6,$$

since X1 and X2 are two mutually exclusive events, and they create the whole space of events. Next, using Bayes' rule

$$P(\theta_1|X_1) = \frac{P(X_1|\theta_1)P(\theta_1)}{P(X_1)} = \frac{0.8 \times 0.6}{0.6} = 0.8$$

Similarly, we get:

$$P(\theta_2|X_1) = 0.2, P(\theta_1|X_2) = 0.3, \text{and} P(\theta_2|X_2) = 0.7$$

We can organise all the probabilities into Table 6.12, and we put them all into the decision tree in Figure 6.13.

Step 3: Attaching utilities to the final nodes. (the leaves), see Figure 6.14. Note, that in this example, we assume that the utility is directly given by the amount of money lost or gained. The money lost is the money paid for the broker's advice, which is f.

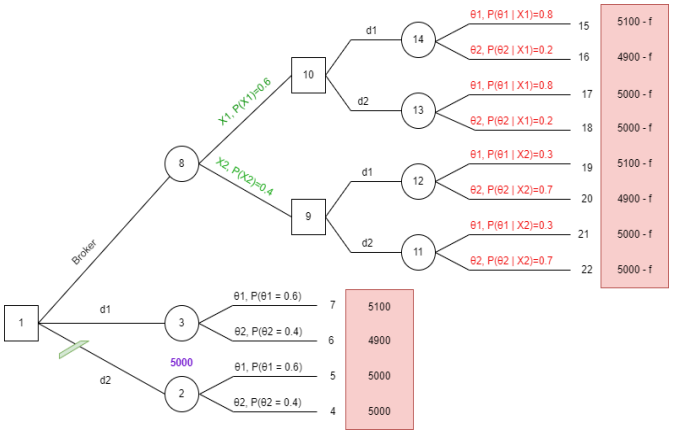


FIGURE 6.14: Investor Iveta's decision tree after Steps 1-3 have been completed. All the calculated probabilities are included in the tree, as well as the utilities at the terminal nodes.

Next, in **Step 4**, of the decision tree building process, we go from right to left:

- We take expectations of utilities at random nodes, and
- We maximise the utilities at decision nodes.

We go back to the Investor example. First, we calculate utilities at the nodes that are the closest to the leaves, see the brown rectangle in Figure 6.15.

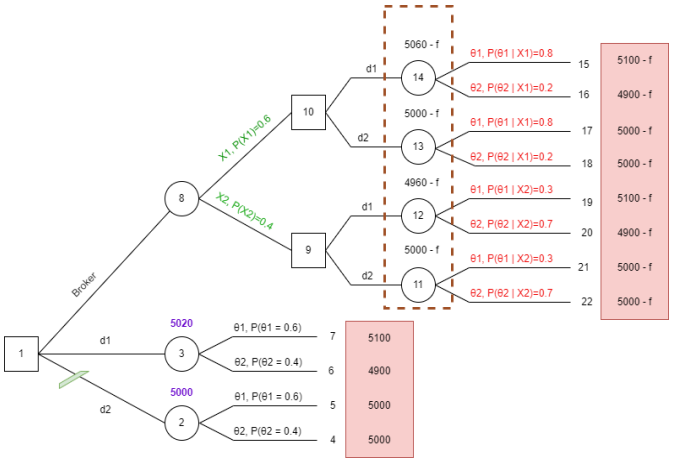


FIGURE 6.15: Investor Iveta's decision tree while doing Step 4 (the falling back). Here we need to find utilities of the random nodes in the brown rectangle.

Next, we move one node to the left. We are now at two decision nodes (see the green rectangle in Figure 6.16). What is the best decision at the top node in the green rectangle? Hint: We need to maximise the utility.

What is the best decision for the bottom node in the green rectangle? We cut away the rejected decisions in the tree. This is done by drawing // on the corresponding branch of the tree (see Figure 6.17). Next, we move one node to the left. We are now at random nodes (see the brown rectangle in Figure 6.17). We calculate utilities at these nodes.

Question: How did we get all the utilities (above in the brown rectangle)? Hint: This is done similarly to calculations in Figure 6.17.

Step 5: Making recommendations about decisions at the initial decision node. How do we decide? See Figure 6.17. We need to decide so we maximise utility.

What course of action would you recommend and why? Answer:

- Decision d_2 is the worst strategy, i.e. it gives the smallest expected utility (here it is profit).
- Decision to use the broker will be the best strategy if he charges fees $f < 16$ $(16 = 5036 - 5020)$. If the broker charges ¡16, then the investor should use the broker. Then if the broker's advice is to invest, then the investor should invest. If the broker's advice is not to invest, then the investor should not

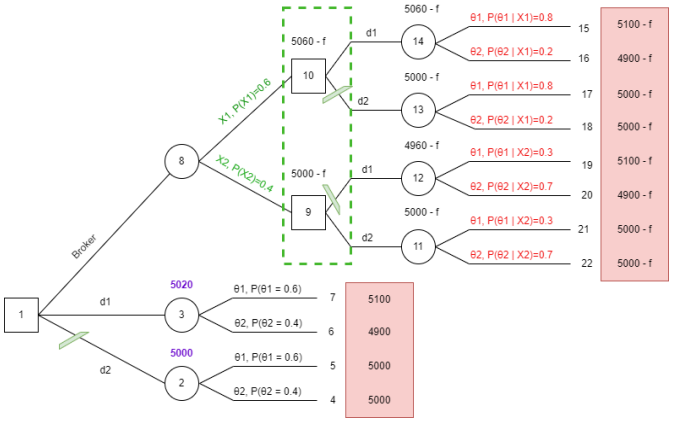


FIGURE 6.16: Investor Iveta's decision tree while working on decision nodes of step 4. We need to find the utilities of the random nodes in the green rectangle.

invest. This way, the investor's expected utility is maximised, which means the expected profit is maximised (since here the utility value is directly measured via money).

- Decision to not use the broker will be the best strategy If he charges $f > 16$. So if the broker charges $f > 16$, the investor should not use the broker. After that decision is made, the investor should invest. This way, the investor's expected utility is maximised, which means the expected profit is maximised (since here the utility value is directly measured via money).

Warning: this recommendation is based on maximising the expected utility (or expected monetary value), and hence it should be used when the decision maker is doing repeated decisions (i.e. has several investments of the same payoff table). If the investor is deciding about one investment, then the payoff she/he get is not equal to the mean, but rather to the values in the leaves, i.e. there is a real risk of losing money.

Note that here we rejected one decision only (see the // notation for d_2 branch coming from the main decision node, in Figure 6.17). The choice of the other two decisions depends on the broker's fee f.

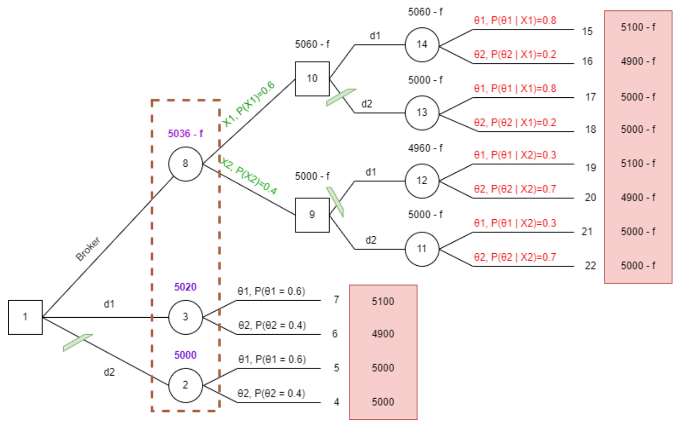


FIGURE 6.17: Investor Iveta's decision tree, while working on step 4. Here we take care of the nodes in the brown rectangle.

6.4 Tips to think and act like a risk expert

Here we briefly give two important tips for risk analysts.

6.4.1 Sensitivity analysis for decision trees

In the examples above, we assumed that we know the probabilities. For example, we assumed precise probabilities of competition and market success in the Manufacturer Mahiro example. We assumed we knew the other values too. In Mahiro's example, we assumed we know precisely the cost of the small plant, of the large plant, of the number of products produced, and the price at which he will sell the products. In real life, we do not know them precisely, but hopefully, we have some imprecise answers. For example, Manufacturer Mahiro may estimate that the cost of the construction of the small plant is between £130 to £160. In other words, Mahiro is facing uncertainty, as he does not have the precise numbers to put into the decision tree. How can he resolve this problem? A way to resolve it is by conducting a sensitivity analysis.

Sensitivity analysis is an important tool of risk analysis (not just decision tree analysis). It aids in reducing uncertainty by identifying high-impact parameters (such as probabilities). This can help in finding out which data (information) to acquire to reduce uncertainty on said parameters.

In decision trees, it is crucial to do sensitivity analysis. It can help in finding out which parameters are not needed and hence reduce the complexity of the decision tree. It can also help to see how the result (the recommended set of actions) depends on the specification of the probabilities or on outcome values. For example, Manufacturer Mahiro can run a decision tree analysis with the

£130 cost of a small plant, and then another analysis with the £160 cost of the small plant and compare the recommended optimal decisions.

6.4.2 Value of information

The decision tree in the Investor Iveta example helps us to calculate a so-called expected monetary value (EMV) of information. In Investor Iveta's example, the EMV when the advice was not sought was £5,200 (Figure 6.9). After the investor sought the advice the EMV was £5,036-f. This means that the EMV value of the information from the broker is £16. Iveta should pay less than £16 for the broker's advice.

6.5 Summary

We learned in this chapter:

1. We learned the five steps for constructing decision trees. In some decision trees, we need to use Bayes' theorem to get the probabilities needed for the tree. Then we learned that careful communication of the results from the decision tree is crucial so that the decision-maker understands us and so he/she can make an informed decision.

2. A decision tree can be used as a model for a sequential decision problem where the stakeholder faces risk. The advantages of using decision trees are: (1) Constructing a decision tree may reveal alternative courses of action that had not previously been identified or thought of. (2) A decision tree formalises the decision-making process and makes the process more objective and logical. (3) The decision tree offers a visual presentation of the options. (4) The requirement of numerical values makes the results of decision-making better. (5) The Management is forced to consider risk.

3. Disadvantages of decision trees are: (1) Probabilities will usually have to be estimated, and might, themselves, be subject to uncertainty (expressed via e.g. variability). The estimated probabilities (and their variance) are as accurate as the data that we use for the estimation. So the quality of data is important to investigate and to communicate to the decision maker too. This is relevant to the strength of knowledge (see SoK in Figure 1.3). (2) Not all factors can be given numerical values. Consequently, the results are purely quantitative and ignore more important qualitative factors. (3) Data may be out of date when a decision is actually taken. (4) The process can be time-consuming and expensive. It may not be worth doing a decision tree. Possible options may be overlooked. (5) Expected values are weighted averages of outcomes and are unlikely to relate to the actual outcome.

6.6 Further reading

We recommend several resources for further reading:

1. A recommended reading on decision trees, with examples from business, is the book called "Decision Behaviour Analysis and Support" by Simon French, John Maule and Nadia Papamichail [24]. They give many real-life examples from their research which is mostly in business such as the Airline purchasing problem on pages 239-249. We also recommend "Decision Theory: Principles and Approaches" by Giovanni Parmigiani, Lurdes Y T Inoue, Hedibert F Lopes [45]. They provide more business examples such as a Travel Insurance example on pages 126-131.

2. Decision trees are very useful and popular in a field called **health economics**. The tree in Exercise 3 (below) is one of the basic trees from health economics. One method used in health economics is decision tree modelling, which extrapolates the cost and effectiveness of competing interventions over time. Such decision tree models are the basis of reimbursement decisions in countries using health technology assessment for decision-making. A good source to start reading about decision trees in health economics is the review by Rauntenberg and colleagues [48] and a case study in [50].

3. To read more about sensitivity analyses we recommend [19] (pages 208-224).

4. The term Decision tree is adopted in **machine learning** (ML) and Artificial Intelligence (AI) community, however, it has a different meaning from the classical decision trees approach shown in this chapter. ML decision trees share the main idea of the decision trees from this chapter: the tree structure. However, the process of building an ML decision tree and the goal of an ML decision tree is different from what we showed in this chapter. In ML, a decision tree is a type of model that uses a set of predictor variables to build a decision tree (to do the branching) that predicts the value of a response variable (at the leaves). For example, we can have a set of data from a patient (age, gender, blood test, image of retina) and we can build a decision tree to predict if the patient has a sight-threatening diabetic retinopathy disease. Many ML books have a chapter on ML decision trees. Such ML decision trees are not the aim of this book.

6.7 Exercises

Solve the following exercises by using pen, paper and calculator.

1. **[Purpose: Practicing decision trees using an oil prospering example.]** A company has to decide whether or not to drill for oil in a particular spot. It costs c units to make a seismic test, the result of which will be a 'good', 'fair', or 'bad' prospect of oil. The actual drilling operation costs 75 units. There are three possible results of drilling: a high yield of oil which can be sold for 200 units, a moderate yield of 100 units, or no oil. The company's data for previous places of this type are as follows:

	High	**Moderate**	**None**	**Totals**
Good	20	10	10	40
Fair	9	9	12	30
Bad	3	12	15	30

   In addition, drilling was not carried out at places with the following seismic records
   Good: 0, Fair: 10; Bad: 20
   Had drilling been carried out it is believed that the results would have been similar to those in the present table. In the past, in places of this type, a seismic test has always been carried out. Answer the following questions:

   (a) What is the maximum value of c to make a seismic test worthwhile?

   (b) If the actual seismic test costs 5 units less than this value, determine the optimum interpretation of the test in terms of whether or not to drill. What is the expected profit?

2. **[Purpose: Practicing decision trees using a manufacturing defects example.]** A part of an aircraft engine can be given a test before installation. The test has only a 75% chance of revealing a defect if it is present, and the same chance of passing a sound part. Whether or not the part has been tested it may undergo an expensive rework operation, which is certain to produce a part free from defects. If a defective part is installed in the engine, the loss is L utiles (loss in utility). The rework operation costs $L/5$ utiles, and 1 in 8 of the parts are initially defective.

   (a) Create the decision tree structure for this problem (i.e. step 1).

   (b) Calculate and add utilities to the tree structure.

   (c) Calculate and add the random node probabilities to the tree structure.

(d) Calculate how much you could pay for the test and determine all the optimum decisions.

3. [**Purpose: Practicing decision trees using a medical example.**] A doctor has the task of deciding whether or not to carry out a dangerous operation on a person suspected of suffering from a disease. If he has the disease and does operate, the chance of recovery is only 50%, without an operation the similar chance is only 1 in 20. On the other hand, if he does not have the disease and the operation is performed there is a 1 chance in 5 of his dying as a result of the operation, whereas there is no chance of death without the operation. Assume there are only two possibilities: death or recovery. What is the optimal decision that a doctor should make?

4. [**Purpose: Practicing decision trees using a property company example.**] A property owner is faced with a choice between three decisions:

   - Decision A: A large-scale investment to improve her flats. This could produce a substantial pay-off in terms of increased revenue net of costs but will require an investment of £1,400,000. After extensive market research, it is considered that there is a 40% chance that a pay-off of £2,500,000 will be obtained, but there is a 60% chance that it will be only £800,000.
   - Decision B: A smaller scale project to re-decorate her premises. At £500,000 this is less costly but will produce a lower pay-off. Research data suggests a 30% chance of a gain of £1,000,000 but a 70
   - Decision C: Continuing the present operation without change. It will cost nothing, but neither will it produce any pay-off. Clients will be unhappy and it will become harder and harder to rent the flats out when they become free.

   (a) Create a decision tree to help make the decisions.

   (b) What choice will you recommend to the property owner?

Part V

COMMUNICATION OF RISK

7

Communication of risk

Gabriela Czanner

Silvester Czanner

CONTENTS

7.1 Lost in translation: Story 1 244
7.2 Lost in translation: Story 2 245
7.3 Objectives of the risk communication 246
7.4 Stakeholders 247
7.5 Brief foundations of risk communication 248
7.5.1 Factors influencing risk perception 248
7.5.2 Cognitive biases and heuristics 249
7.5.3 Communication theories relevant to risk communication 250
7.6 Toward effective risk communication 252
7.6.1 Practical elements of effective risk communication 252
7.6.2 Ethical and legal considerations in risk communication 253
7.6.3 Evaluation and improvement of risk communication ... 253
7.7 Tips to think and act like a risk expert 254
7.7.1 On risk communication in finance 254
7.7.2 On risk communication in medicine 254
7.7.3 On risk communication in Artificial Intelligence 255
7.8 Summary 256
7.9 Further reading 256

Communication of risk is crucial. It should not be merely a statement of risks, but rather a dialogue between the risk expert and the stakeholders. Effective risk communication encourages open dialogue and two-way communication between experts and stakeholders. It creates opportunities for individuals to ask questions, express concerns, provide feedback, and contribute to the decision-making process. Such two-way communication enables a better understanding of risk among stakeholders, it enables a better understanding of perceptions, needs, and values among the risk experts too. This then leads to more effective

risk management strategies.

Learning objectives

1. Learn what the stakeholders are and get a gentle introduction into what shapes their perception of risk and, thus, their behaviour. We will briefly discuss several cognitive biases related to the risk that people have. By people, we will mean the stakeholders.
2. We will look into strategies for effective risk communication. We will also briefly discuss the ethical and legal considerations.
3. We will learn how to evaluate and improve our communication of risks.

7.1 Lost in translation: Story 1

This is a story about Ariel, an experienced risk analyst and AI developer, who failed to communicate risk when presenting the AI prototype software developed by her research team. For a new hypothetical patient, the AI was estimating if the patient has sight-threatening diabetic retinopathy based on seeing the patient's colour fundus image of the retina. She was talking to her collaborators: one expert in cybernetics and education and one expert in glaucoma (inspired by our own experience, the name was changed).

She was challenged by both collaborators about the meaning of the estimated probability. The probability was 91% with a 95% credible interval of 75% to 93%. While showing these numbers to the collaborators, the clinician asked her: "What do these probabilities mean?" And the expert in cybernetics and education asked her, "Are these probabilities an objective examination of the patient's risk?".

They were good questions. The team planned to spend a large amount of time negotiating with investors and persuading them to buy the prototype and a large amount of money to pay for a patent application. It was already established that the AI prototype is accurate enough via metrics such as sensitivity, specificity, positive predictive value and negative predictive value (see Section 2.4.2). The problem was that Ariel could not give the collaborators a good answer. She was the risk expert but did not have clear and convincing answers to these basic questions, also addressing how to understand uncertainties related to these numbers. She was not able to communicate the results of the AI-estimated risk in a way that was trustworthy.

The clinician was the main decision maker, and he had to decide whether to talk to investors. He needed to be sure how to answer the questions above to potential investors and regulators. He was informed about the probability and the credible interval, but there was a lack of clarity on what these numbers

meant, which made him question how such an AI prototype would be used in real life. The questions are about interpretation. Ariel was answering in academic terms, and she kept answering, "It is the probability of the patient having sight-threatening diabetic retinopathy" and "No, the probabilities are not objective; they are subjective estimation or guess made by AI". The problem with the first answer is that patients do not generally use probabilities to make decisions; they use natural frequencies. The problem with the second answer is that the clinician and the cybernetics expert did hope for an objective answer; they were not ready to accept a subjective answer; they felt AI "ought to be objective". Ariel felt that AI is "subjective" since it learns from certain data. (As she found later, her logic was wrong.)

The long discussion ended with the collaborators telling the risk analyst she was "biased in her knowledge of probability". At that point, the risk analyst was annoyed. After all, she has trained all her life not to be biased.

This example shows how even a seemingly simple case can turn into confusion. Without understanding the estimated risk, a patient and clinician would not know how to make a decision. Another non-medical example of miscommunication can be found in [9] (see their Chapter 7).

7.2 Lost in translation: Story 2

This is a story about Danica, a retired former research nurse. She was failed by a sequence of misfortunate communications in the health system, causing her to lose sight in her left eye irreversibly (inspired by a real story we heard from a patient we interviewed for our project, the name has been changed).

During Covid-19 time, in early 2021, she went to her optometrist to check her left eye. Her vision was blurred. The optometrist used high-end expensive eye cameras to check the back of the eye as well as the front of the eye. The optometrist then told her that this was a likely cataract, which is treatable and that she did not need to hurry to the hospital. He said it is not glaucoma disease. He said that it is OK to wait six months till hospitals will be more free and the pandemic calms down. He also referred her to see her general practitioner. She saw her general health practitioner, who read the message from the optometrist while forgetting to tell her that her intraocular pressure was a bit too high (which is a risk factor for glaucoma). Another four months later, she lost sight in her left eye due to glaucoma, which is irreversible as there is no treatment. However, there is a medical treatment to slow down or stop the progression of glaucoma.

What went wrong? The problem was in the risk communication. Danica trusted the optometrist. She saw the optometrist operating the same cameras as she once saw in hospitals; thus, she concluded that the optometrist is trained to diagnose various eye diseases, including glaucoma. The optometrist did not

express any words of uncertainty. He sounded very sure and competent. Danica trusted her general health practitioner. However, the practitioner was unaware (or forgot) that an increase in intraocular pressure is a risk factor for glaucoma (though there are glaucoma patients who never had an increased intraocular pressure).

This example shows how risk communication can turn into a wrong decision. Danica would probably act differently had she known that there is even a small chance of having glaucoma. Danica might have acted differently if the optometrist had said that he was unsure about what he saw and if he told her to come to see him again in two months. Communication of the uncertainty of the optometrist was not done here. There were two sources of uncertainty: cataract makes it hard and sometimes impossible to see the back of the eye, and thus to diagnose glaucoma, the optometrists do not have training for diagnosis of glaucoma as it is too complex and thus it is done by glaucoma specialists at the hospitals. None of these sources of uncertainty were communicated to Danica. This example highlights the importance of clear risk communication, the complexity of glaucoma disease, and the need to do more research on glaucoma detection, including developing useful AI tools to aid eye professionals in preserving people's sight.

7.3 Objectives of the risk communication

Generally, the aim of risk communication is to inform the stakeholders about the risks so that they understand them and make the best possible decision. In what follows, we describe several specific objectives of risk communication.

Firstly, we inform the stakeholders and provide accurate, relevant, and timely information about risks to individuals, communities, and other stakeholders. It must be ensured that people have a clear understanding of the nature of the risk, its potential consequences, and the actions they can take to mitigate or respond to it.

We educate the stakeholders about the causes, factors, and underlying science behind the risk, as well as the potential health, environmental, or societal impacts associated with it. By improving stakeholders' knowledge, they can make informed decisions and take appropriate precautions. This way, we empower the stakeholders to make informed choices and take action to protect themselves from risks. It provides them with the necessary tools, resources, and guidance to assess their own vulnerabilities, understand their options, and implement risk reduction or mitigation measures.

We build trust with the stakeholders, individuals, communities, and businesses by providing transparent, honest, and consistent information. Often communication starts with not trust. Building trust enables effective communication and facilitates a sense of collective responsibility in managing risks.

Sometimes we need to influence the behaviour of individuals and motivate them to adopt protective measures (such as not socialising during the Covid-19 pandemic lockdown). By influencing behaviour, we seek to minimise the potential impacts on individuals and communities.

We facilitate the preparedness and resilience of people in the face of risks to be ready in case of future emergencies, disasters, or ongoing risks. It is important to do it in such a way that it does reduce uncertainty and anxiety.

We create stakeholder engagement and participation because the stakeholders are involved in shaping the risk management processes. The stakeholders are encouraged to say what their fears are and what their hopes are about risk communications. This enhances the legitimacy and effectiveness of risk management efforts and improves the collective feeling.

7.4 Stakeholders

There are various stakeholders in risk communication, each having different responsibilities and interests. Each risk situation involves several or many stakeholders, where some are decision-makers, and some are affected. In the Covid-19 pandemic, everyone was a stakeholder, from governments, doctors, nurses, cleaners, and researchers to public individuals. In what follows, we briefly list the types of stakeholders.

Government agencies include public health departments, environmental protection agencies, and regulatory bodies. They are responsible for assessing and managing risks, providing accurate information to the public, and coordinating emergency response efforts. An example was Covid-19 and health departments and government representatives needing to decide what action to take concerning the lockdown and what to say to the public.

Scientific and technical experts and subject matter (e.g., experts in virology). They analyse data, conduct risk assessments, and provide evidence-based information to inform risk communication strategies. These experts often collaborate with government agencies, industry representatives, and other stakeholders to ensure accurate and up-to-date information is conveyed to the public.

Industry and business entities that produce or handle potentially risky products (such as toxic waste) or processes (such as AI-related risks where AI wrongly classify a person as healthy) have a responsibility to communicate the associated risks. They collaborate with government agencies to adhere to regulations, provide necessary warnings or precautions, and ensure the safety of their products or services. Industry stakeholders also play a role in crisis communication during incidents or accidents related to their operations.

Industry and business entities also want to ensure that their businesses are profitable and not victims of risks too caused by, e.g. digital attacks, the

arrival of a competing product, or the effect of inflation. They want to discuss their projects with risk analysts as well as a domain expert relevant to their projects. Typically they want to diversify the risks into a portfolio of various projects, some less risky, some more risky.

Non-governmental organisations (NGOs), such as advocacy groups, consumer organisations, and public interest groups, often engage in risk communication to represent the concerns and interests of specific communities or populations. They play an important role in ensuring that public opinions, needs, and perspectives are considered. NGOs can also provide additional expertise, mobilise community engagement, and hold other stakeholders accountable for their risk communication efforts.

Media and journalists are vital in disseminating risk-related information to the public. Journalists act as intermediaries between experts, authorities, and the public, translating complex information into accessible formats. Usually, journalists work with risk experts to ensure the information is clear and not too simplified.

Health and safety professionals in public health, occupational health and safety, and emergency management are instrumental in risk communication. They work within organisations, government agencies, or independently to assess, manage, and communicate risks related to health, safety, and emergencies. These professionals provide guidance, develop protocols, and support risk communication efforts in their respective fields.

Community leaders and local organisations are crucial in risk communication, particularly in engaging and mobilising specific communities.

The general public is an important stakeholder in risk communication. Individuals and communities must be informed about risks, understand their implications, and know how to protect themselves. Active participation, engagement, and compliance with risk communication messages are vital for effective risk management.

7.5 Brief foundations of risk communication

Here we briefly learn about risk perception and psychology: the factors influencing risk perception and cognitive biases and heuristics.

7.5.1 Factors influencing risk perception

People perceive risk differently. Various factors influence them. Personal experience is the first important factor. If someone has personally experienced a negative outcome or harm related to a specific risk, he/she may perceive it as more significant and alarming.

Media plays a crucial role in shaping risk perception. How risks are por-

trayed in the media can influence how people perceive them. Sensationalised or exaggerated media coverage can lead to overestimating risks, while limited coverage or downplaying of risks may result in underestimation.

Cultural and social factors can shape risk perception. Different cultures and societies have varying values, beliefs, and norms that influence how they perceive risks. Factors such as collective responsibility, trust in institutions, and social influence can impact individual risk perceptions. If individuals have confidence in these entities, they may perceive risks as better managed and controlled, leading to lower levels of concern.

The degree of control individuals feel over risk can influence their perception of it. If someone believes they have control over risk, they may perceive it as lower and more manageable. On the other hand, if they perceive a lack of control, the risk may be perceived as higher and more threatening.

The level of knowledge and information individuals have about a particular risk can shape their perception of it. Understanding the nature of a risk, its potential consequences, and the likelihood of occurrence can influence risk perception.

Emotions can play a significant role in risk perception. Risks that evoke strong emotions, such as fear, anger, or disgust, are often perceived as more significant and threatening. Emotional responses can sometimes override rational assessments of risks.

The availability heuristic is another factor influencing risk perception. It refers to the tendency to judge the likelihood and severity of risks based on how easily examples or instances of those risks come to mind. If people can easily recall vivid or memorable instances of a risk, they may perceive it as more common or severe.

Social amplification of risk occurs when public perception and concern about risk are magnified through social interactions and communication. If people perceive that others around them are highly concerned about a particular risk, it can amplify their risk perception.

It's important to note that these factors interact and can influence each other, leading to complex and nuanced risk perceptions among individuals and communities.

7.5.2 Cognitive biases and heuristics

Several cognitive biases and heuristics are relevant to risk perception and decision-making [60], [36]. Here are a few key ones:

Availability heuristic bias occurs when people judge the likelihood or frequency of an event based on how easily they can recall or remember similar events. If a particular risk or event is more memorable or vivid, it tends to be perceived as more likely or prevalent than it is.

Anchoring and adjustment heuristic bias involve relying heavily on an initial piece of information (the anchor) when making decisions or estimations.

Subsequent judgments are then adjusted from this initial anchor. In the context of risk perception, if individuals are presented with a specific reference point or starting point, it can influence their perception of the associated risk.

Overconfidence bias is the tendency to overestimate one's abilities, knowledge, or accuracy in making judgments or decisions. People often have excessive confidence in their ability to handle risks, leading them to underestimate the potential negative consequences.

Optimism bias is characterised by individuals believing that they are less likely to experience negative events or risks compared to others. This bias can lead to underestimating risks and overestimating the likelihood of positive outcomes.

Loss aversion is the tendency to weigh potential losses more heavily than equivalent gains. People are generally more sensitive to losses and take risks to avoid losses than potential gains. This bias can influence risk perception and decision-making when individuals focus more on avoiding losses than maximising potential gains.

Confirmation bias occurs when individuals seek or interpret information in a way that confirms their preexisting beliefs or expectations. In the context of risk perception, people may selectively pay attention to information that supports their existing views on risk and ignore contradictory evidence, leading to biased perceptions and decisions.

Social influence bias occurs when individuals are influenced by the opinions, beliefs, and behaviours of others in their social groups. People may adjust their risk perception and decisions based on what they perceive as the norm or socially desirable behaviour within their social circles.

Understanding these cognitive biases and heuristics can help individuals become more aware of their potential influence on risk perception and decision-making. By recognising these biases, individuals can make more informed and rational assessments of risks and improve their decision-making processes.

7.5.3 Communication theories relevant to risk communication

We mention here three theories: the social amplification of risk framework, the social cognitive theory and the diffusion of innovations theory.

The Social Amplification of Risk Framework (SARF) is a theoretical framework developed by social scientists to explain how social processes can amplify or attenuate the perception, communication, and response to risks. SARF recognizes that risk perception is not solely determined by the objective characteristics of a risk but is also shaped by social, cultural, and psychological factors. The framework proposes that risk events undergo a series of amplification or attenuation processes as they pass through various stages of communication and interpretation.

Social cognitive theory is also relevant to understanding risk perception and behaviour. It provides several insights into risk perception. Firstly, it highlights

the importance of observational learning. Individuals can acquire knowledge, attitudes, and behaviours related to risk by observing others, such as parents, peers, or media figures. Through observation, people can learn about the potential risks, how others perceive and respond to them, and the outcomes associated with different risk-related behaviours. Secondly, it believes in self-efficacy. Individuals that are self-efficient are more likely to perceive risks as manageable and have confidence in their ability to engage in risk-reducing behaviours. On the other hand, individuals with low self-efficacy may perceive risks as overwhelming and may be less likely to take appropriate preventive actions. Thirdly, the social cognitive theory emphasises the importance of outcome expectations in risk perception and behaviour. Positive outcome expectations, such as perceived benefits or rewards, can increase the likelihood of engaging in risky behaviours, while negative outcome expectations, such as perceived harms or punishments, can deter individuals from taking risks. Fourthly, the social cognitive theory sees importance in self-regulation processes of individuals, such as self-monitoring, goal setting, and self-reflection in risk-related decisions. Individuals who actively monitor and regulate their risk behaviours through goal setting and self-reflection are more likely to engage in adaptive risk management strategies and make informed decisions. Fifth, the social cognitive theory highlights the bidirectional relationships between individuals and their environment. Risk perception and behaviour are influenced by environmental factors such as social norms, cultural values, and media messages. At the same time, individuals can shape their environment by making choices and engaging in behaviours that affect their risk exposure and perception. By considering social cognitive theory, risk communication can be made so that it enables effective interventions, communication strategies, and educational programs aimed at promoting accurate risk perception, enhancing risk management skills, and facilitating behaviour change towards safer and more informed decision-making in the face of risks.

Diffusion of innovations theory was developed by sociologist Everett Rogers and is highly relevant to risk communication. The theory focuses on how new ideas, behaviours, or innovations spread and are adopted within a population. In what follows, we mention key principles. The first principle is the adoption of Risk-Mitigating behaviours. In risk communication, the goal is often to encourage individuals to adopt behaviours that reduce their exposure to risks. Diffusion of Innovations theory provides insights into the factors that influence the adoption of such risk-mitigating behaviours. These factors include the perceived benefits and relative advantages of the behaviour, its compatibility with existing values and norms, simplicity and ease of use, and the social influence of opinion leaders or influential individuals who have already adopted the behaviour. The second principle is the importance of choosing the right communication channels in disseminating information about risks and risk-mitigating behaviours. Different communication channels, such as mass media, social media, interpersonal communication, or community networks, can be utilised to facilitate the spread of risk infor-

mation and encourage behaviour change. The third principle is about opinion leaders. They play a crucial role in the diffusion of innovations, including risk-mitigating behaviours. These are individuals who are influential within their social networks and are early adopters of new ideas or behaviours. Leveraging opinion leaders in risk communication can help accelerate the adoption of risk-mitigating behaviours within a community by providing visible examples and reinforcing social norms. The fourth principle is the importance of social norms on risk perception and behaviour. Social norms refer to the shared expectations, values, and beliefs within a community that influence individuals' behaviour. Effective risk communication aims to influence social norms by highlighting the prevalence of risk-mitigating behaviours and emphasising their acceptance within the community. The fifth principle is the decision-making process. Diffusion of Innovations theory highlights the different stages of the decision-making process individuals go through when considering the adoption of new behaviours. These stages include knowledge, persuasion, decision, implementation, and confirmation. Risk communication efforts can address each stage by providing information, addressing concerns, offering incentives, and providing ongoing support to facilitate the adoption and maintenance of risk-mitigating behaviours.

The theories we briefly mentioned here are relevant to any stakeholder: from the public to business owners.

7.6 Toward effective risk communication

Here we discuss several ideas for effective risk communication: practical elements of the communication, ethical and legal considerations, and evaluation and improvement of risk communication.

7.6.1 Practical elements of effective risk communication

Here we briefly mention the elements of effective communication include:

1. Message development. The message needs to be clear and concise. It needs to be tailored to the audience. And it also needs to address uncertainty, e.g. when communicating the risk of oesophagal cancer, we need to say what data (information) we used to calculate the risk.

2. Channel selection. We need to choose the right mean of communication, e.g. verbal or written, media or newspaper. We need to remember to manage the expectations of our audience and think about how to become trustworthy.

3. Building trust and credibility. This is recommended by being transparent

and not hiding any important facts, by demonstrating competence and by trying to create a two-way communication.

4. Cultural and linguistic considerations. Here we must address cultural differences, overcome language barriers and incorporate cultural values and norms where possible.

7.6.2 Ethical and legal considerations in risk communication

When we communicate the risks, we need to consider the ethical principles and dilemmas. For example, when the risk of Covid-19 was communicated to the public in February 2020 (i.e. at the very start of the pandemic), it was important to find the balance between transparency and public panic. It was important to communicate so that it protects vulnerable populations (e.g. people with immunodeficiency problems). It is also important to avoid conflicts of interest.

When communicating risks, we also need to consider legal frameworks and regulations. For example, we need to be compliant with privacy and data protection laws, with advertising and marketing restrictions, and with health and safety regulations.

7.6.3 Evaluation and improvement of risk communication

Lastly, a risk expert (or risk team) must be doing a self-reflection. This means evaluating how the risk communication went and improving on it in future.

There are several strategies that a risk expert can use to evaluate the quality of risk communication. The first strategy is monitoring and evaluation metric. This means assessing audience understanding and perception, e.g. via a questionnaire or a dialogue. This also means analysing media coverage and public discourse. And it can also include a measurement of the behavioural change and adherence to the change (e.g. if a doctor recommends a diet to lower cholesterol, then there are ways to measure the cholesterol level several weeks after the recommendation).

There are some ideas on how a risk expert can engage in continuous improvement of his/her risk communication. The natural ideas incorporate feedback and lessons learned, adapting strategies based on emerging risks and incorporating new technologies and tools.

The risk analyst Ariel from the example at the beginning of this chapter, did reflect after the heated discussion. She read several papers and books on risk communication, and she consulted other risk experts. She arrived at answers that were satisfactory to the clinician and cybernetics specialist. The first question, "What do these probabilities (91% with a credible interval 75% to 93%) really mean?" she answered: "It means that if we have 100 patients with the back of the eye looking similar to the patient in front of us, then we expect that 91 of them do have a sight-threatening diabetic retinopathy,

however with probability 95% there can be from 75 to 93 patients having sight-threatening diabetic retinopathy". The second question "Are these probabilities an objective examination of the patient's risk?", she answered as "yes, these probabilities should be seen as objective (i.e. rational) beliefs as opposed to subjective. What is meant by objective belief? The degree of belief that is rational for a person (or AI) to hold, given the evidence available, is fixed, and in that sense is objective. But this objectivity exists at the level of knowledge. For AI the knowledge comes from the training data that we used to train the AI, and the type of the model we chose for AI was convolutional neural networks".

7.7 Tips to think and act like a risk expert

Here we discuss several tips on risk communication by bringing in several practical use cases.

7.7.1 On risk communication in finance

Crouhy, Galai and Mark, in their book "The essentials of risk management" (2006) [17] they mention a talk of Mervyn King, governor of the Bank of England, who pointed out the distinction between risk and uncertainty using the example of the pensions and insurance industries. These industries have used statistical analysis to develop products (life insurance, pensions, annuities, and so on) that are important to us all in looking after the financial well-being of our families. These products act to "collectivize" the financial effects of any one individual's life events among any given generation.

In his speech, Mervyn King set out two principles of risk communication for public policymakers. Such principles can be used by senior risk committees at corporations looking at the results of complex risk calculations:

1. Information must be provided objectively and placed in context so that risks can be assessed and understood.

2. Experts and policymakers must be open about the extent of our knowledge and our ignorance. Transparency about what we know and what we don't know, far from undermining credibility, helps to build trust and confidence.

7.7.2 On risk communication in medicine

This section is based on the blog María del Carment Climént [18] where she is proposing seven steps to communication. We summarise the steps in Table 7.1.

Step	Need to do
Step 1	Clarify what the risk is and who is affected
Step 2	Specify the time period the risk refers to
Step 3	Present relative and absolute risks
Step 4	Pay attention to the format of numbers
Step 5	Include graphics whenever possible
Step 6	Provide balanced information
Step 7	Explain uncertainties

TABLE 7.1: Seven steps of risk communication. (Adopted from [18]).

Next, we will discuss the first step, only. In Step 1, we need to clarify what can happen and clarify who is affected by it. For example, in the case of diabetic retinopathy-related harm, we need to specify if we talk about the risk of progressing to mild retinopathy, sight-threatening retinopathy, or losing sight. In the case of Covid-19 related harm, we need to say if we are talking about the risk of infection, the risk of hospitalisation or the risk of dying.

We need to clarify who is affected by the risk. In other words, we need to clearly say who is the affected stakeholder and if the research we did was on the same type of stakeholders. For example, do we talk about diabetic retinopathy in UK people who are 50-100 years old, or any age? Specifying the age and location is one example of a so-called stratification. If the risk is communicated about humans, then we should **stratify the group** we are referring to as much as possible, e.g. 50-100 years old in the UK with diabetes.

Caution! The risk of contracting the disease is not the same as the risk of dying from the disease. Contracting the disease and dying from the disease are two different outcomes. Hence we must specify what outcome we have in mind, i.e. what can happen.

Caution! The research done in humans, and evidence found in humans is not the same as the research evidence found in another species. The research results obtained in humans in Spain may not be relevant for affected people in Slovakia. Hence we should say where the research was done.

7.7.3 On risk communication in Artificial Intelligence

AI holds the potential to improve lives, by helping us in doing complex cognitive tasks. For example, there is research for AI to help optometrists to detect glaucoma from retinal images, but such an algorithm is still not used in real life. There are algorithms already used in real life, some are helping humanity,

but some have created terrible mistakes [21]. There are two main types of risk communication by AI

1. AI should be able to tell how accurate it is on a target population. We mentioned some metrics in Chapter Probability in Section 2.4.2). There are more metrics. Notably, sometimes AI makes mistakes which lead to AI-related risks.

2. AI should be able to tell how precise it is on a person (item). We mentioned one example earlier in this Chapter when we discussed Ariel's experience communicating her AI prototype to a clinician and cybernetics expert. The precision was measured by the 95% credible interval (the narrower the interval, the better the precision).

Both types are an intensive area of current research. Especially the second one requires a lot of research attention (as concluded by European Commission High-Level Expert Groups).

7.8 Summary

We learned in this chapter:

1. One communication of the risks can involve various stakeholders with conflicting needs or requirements.

2. A risk communication is not merely a state of providing information. It is a dialogue. Even when a risk about Covid-19 is put into the news as a one-way communication, it should be open for feedback regarding the format, content and clarity of the messages.

3. Effective risk communication requires knowledge of cognitive psychology and is an active area of research. The research must be ongoing and in close collaboration and participation of the stakeholders (this is called *participatory design*).

4. One emerging area of research is how to run focus groups to understand people's attitudes toward AI, how AI can gain the trust of people and what people feel is a safe AI.

7.9 Further reading

The chapter was inspired by our research, the research of others, guidelines of international societies and several books too. Here we list the monographs we were inspired the most, as well as we recommend resources for future reading:

1. Our chapter was mostly inspired by the monograph *Risk Science* by Aven and Thekdi [9], 2022, and their chapter on communication.

2. We were also inspired by the work of psychologist Gerd Gigerenzer, especially his 2002 book named *Reckoning with risk. Learning to live with uncertainty* [25] and his 2014 book named *Risk savvy. How to make good decisions* [26]. The book is written in lay language, and we highly recommend reading it. It provides examples of communication of risks in all areas of life, from criminal court cases to health screening programs.

3. We were also inspired by the paper of Tversky, A. and Kahneman, D Judgment under Uncertainty: Heuristics and Biases, 1974 [60].

4. We only mentioned several cognitive biases. A more comprehensive list (about 180 types) of cognitive biases can be found, e.g. in [38].

5. For a further reading we also recommend *Risk Assessment and Decision Analysis with Bayesian Networks*, by Fenton and Neil [22].

Bibliography

[1] Cambridge Dictionary Website: https://dictionary.cambridge.org/.

[2] Johnson's Dictionary of the English Language Website: https://www.britannica.com/topic/A-Dictionary-of-the-English-Language-by-Johnson.

[3] Oxford English Dictionary.

[4] The Stanford Encyclopedia of Philosophy: Risk, year = 2007 Website: https://plato.stanford.edu/entries/risk/.

[5] H. Akaike. Fitting autoregressive models for prediction. *Ann. Inst. Stat. Math.*, 21(-):243–247, 1969.

[6] H. Akaike. Information theory and an extension of the maximum likelihood principal. In *In 2nd Int. Symp. Inform. Theory*, pages 267–281. B.N. Petrov and F. Csake, eds. Budapest: Akademia Kiado, 1973.

[7] H. Akaike. A new look at statistical model identification. *IEEE Trans. Automat. Contr.*, AC-19(-):716–723, 1974.

[8] B.J.M. Ale. Risk analysis and big data. *Safety and Reliability*, 36(3):153–165, 2016.

[9] T. Aven and S. Thekdi. *Risk Science*. Routledge, New York, 1st edition., 2022.

[10] T. Bedford and R. Cooke. *Probabilistic risk analysis: foundations and methods*. Cambridge University Press, The Edinburgh Building, Cambridge CB2 8RU, UK, 2001.

[11] T. Blount. *Glossographia, Year = 1656*.

[12] R.G. Brown. *Statistical forecasting for inventory control*. McGraw-Hill Education, New York, 1959.

[13] M.A. Carlton and J.L. Devore. *Probability with Applications in Engineering, Science, and Technology*. Chapman and Hall/CRC; 2nd edition, 2018.

[14] T.M. Choi and J.H. Lambert. Advances in risk analysis with big data. *Risk Analysis*, 37(8):1435–1442, 2017.

[15] A. Coghlan. *A Little Book of R For Biomedical Statistics. Release 0.2.* Parasite Genomics Group, Wellcome Trust Sanger Institute, Cambridge https://a-little-book-of-r-for-biomedical-statistics.readthedocs.io/en/latest/, 2016.

[16] M.J. Crawley. *The R Book. Edition 2.* Wiley, 2012.

[17] M. Crouhy, D. Galai, and R. Mark. *The essentials of risk management.* McGraw-Hill, 2006.

[18] M. del Carment Climént. How to communicate risks reported in scientific articles in an understandable way Project Website: https://sciencemediacentre.es/en/how-communicate-risks-reported-scientific-articles-understandable-way.

[19] D. Halldorsson T. Jeger M.J. Knutsen H.K. More S. Naegeli H. Noteborn H. Ockleford C. Ricci A. et al. EFSA Scientific Committee, Benford. The principles and methods behind EFSA's guidance on uncertainty analysis in scientific assessment. *EFSA Journal*, 16(1):5122, 2018.

[20] B. Everitt. *Chance Rules: An Informal Guide to Probability, Risk and Statistics.* Springer, 2009.

[21] Failures of AI. Website: https://www.analyticsinsight.net/top-10-massive-failures-of-artificial-intelligence-till-date/.

[22] N. Fenton and M. Neil. *Risk Assessment and Decision Analysis with Bayesian Networks.* Springer, 2nd Edition, 2018.

[23] C.R. Fox and U. Gulden. Distinguishing two dimensions of uncertainty. In *Essays in Judgment and Decision Making*, pages –. Brun, W., Kirkebøen, G. and Montgomery, H., eds. Oslo: Universitetsforlaget., 2011.

[24] Maule J. French, S. and N. Papamichail. *Decision Behaviour Analysis and Support.* Cambridge University Press, 2009.

[25] G. Gigerenzer. *Reckoning with risk. Learning to live with uncertainty.* The Penguin Press, 2002.

[26] G. Gigerenzer. *Risk savy. How to make good decisions.* The Penguin Press, 2014.

[27] S.D. Guikema. Artificial intelligence for natural hazards risk analysis: Potential, challenges, and research needs. *Risk Analysis*, 40(6):1117–1123, 2020.

[28] S.D. Guikema. Natural disaster risk analysis for critical infrastructure systems: An approach based on statistical learning theory. *Reliability Engineering and System Safety*, 94(4):855–860, 2020.

[29] C. Hanck, M. Arnold, A. Gerber, and Schmelzer M. *Introduction to Econometrics with R.* Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License https://www.econometrics-with-r.org/index.html, 2021.

[30] S.O. Hansson. From the casino to the jungle. Dealing with uncertainty in technological risk management. *Synthese*, 168(-):423–432, 2009.

[31] M.A. Hernán and J.M. Robins. *Causal Inference: What If.* Boca Raton: Chapman Hall/CRC, available at https://www.hsph.harvard.edu/miguelhernan/causal-inference-book/, 2023.

[32] C.C. Holt. Forecasting seasonals and trends by exponentially weighted averages (onr memorandum no. 52), carnegie institute of technology, pittsburgh usa. reprinted in the international journal of forecasting, 2004, 20(1), pages 5-10.

[33] R.A. Howard and A.E. Abbas. *Foundations of decision analysis.* Pearson, 2016.

[34] R.J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice. 3rd ed.* OTexts: Melbourne, Australia. OTexts.com/fpp3, 2021.

[35] R. Kaas, Goovaerts, M., J. Dhaene, and M. Denuit. *Modern actuarial risk theory using R. Second edition.* Springer, 2009.

[36] D. Kahneman. *Thinking fast and slow.* Penguin; 1st edition, 2011.

[37] S. Kaplan and B.J. Garrick. One the quantitative definition of risk. *Risk Analysis*, 1(-):11–27, 1981.

[38] Accessed 24 August 2023 List of cognitive biases Wikipedia page. https://en.wikipedia.org/wiki/List$_o f_c ognitive_b iases$.

[39] C.L. Mallows. Some comments on Cp. *Technometrics*, 15(-):661–675, 1973.

[40] A.D.R. McQuarrie and C.L. Tsai. -. In *Regression and time series model selection*, pages –. Singapore: World Scientific, 1998.

[41] R. Nateghi and T. Aven. Risk analysis in the age of big data: the promises and pitfalls. *Risk Analysis*, -(-):doi/10.1111/risa.13682, 2021.

[42] Office for National Statistics. Dataset overseas travel and tourism time series, 2020. Data retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/datasets/internationalpassengersurveytimeseriesspreadsheet.

[43] T. O'Hagan. Dicing with the Unknown. *Significance*, 1(3):132–133, 2004.

[44] Collaborative Group on Hormonal Factors in Breast Cancer. Bootstrap prediction intervals for linear, nonlinear and nonparametric autoregressions. *Journal of Statistical Planning and Inference*, 177(341):1–27, 2016.

[45] Inoue L.Y.T. Parmigiani, G. and H.F. Lopes. *Decision Theory: Principles and Approaches.* Wiley, 2009.

[46] J. Pearl and D. MacKenzie. *The book of why. The new science of cause and effect.* Penguin Books LTD, 2018.

[47] P. Ponsko and B. Rybaczyk. Fan chart – a tool for nbp's monetary policy making. Paper obtained from https://www.nbp.pl/publikacje/materialy_i_studia241_en.pdf.

[48] Gerritsen A. Rautenberg, T. and M. Downes. Health Economic Decision Tree Models of Diagnostics for Dummies: A Pictorial Primer. *Diagnostics (Basel)*, 3(10):158, 2020.

[49] M.A. Rotondi. To ski or not to ski: Estimating transition matrices to predict tomorrow's snowfall using real data. *Journal of Statistics Education*, 18(3), 2010.

[50] Orlando R. Sanghera, S. and T. Roberts. Economic evaluations and diagnostic testing: An illustrative case study approach. *International Journal of Technology Assessment in Health Care*, 1(29):53–60, 2013.

[51] Benneyan J.C. Kiss I.G. Briggs-Gowan M.J. Copeland W. Sheldrick, R.C. and A.S. Carter. Thresholds and accuracy in screening tools for early detection of psychopathology. *J Child Psychol Psychiatry*, 56(9):936–48, 2015.

[52] R.H. Shumway and D.S. Stoffer. *Time Series Analysis and Its Applications: With R Examples. Fourth edition.* Springer Publisher, 2017.

[53] Andrienko N. Andrienko G. Shurkhovetskyy, G. and G. Fuchs. Data Abstraction for Visualizing Large Time Series. *Computer Graphics Forum*, 2017.

[54] G.A. Spedicato, T.S. Kang, S.B. Yalamanchi, D. Deepak Yadav, and I. Cordón. The markovchain Package: A Package for Easily Handling Discrete Markov Chains in R Project Website: https://cran.r-project.org/web/packages/markovchain/vignettes/an-introduction-to-markovchain-package.pdf.

[55] ISO Standards. ISO 31000 The international standard for risk assessment Website: https://www.iso.org/iso-31000-risk-management.html.

[56] R.A. Stine. Bootstrap Prediction Intervals for Regression. *Journal of the American Statistical Association*, 80(392):1026–1031, 1985.

[57] Tzioutzios D. Cruz A.M. Suarez-Paba, M.C. and E. Krausmann. Toward natech resilient industries. In *Essays in Judgment and Decision Making*, pages –. In: Yokomatsu, M., Hochrainer-Stigler, S. (eds) Disaster Risk Reduction and Resilience. Disaster and Risk Research: GADRI Book Series. Springer, Singapore., 2020.

[58] Tatar U. Santos J.R. Thekdi, S.A. and S. Chatterjee. Disaster risk and artificial intelligence: A framework to characterize conceptual synergies and future opportunities. *Risk analysis : an official publication of the Society for Risk Analysis*, 6, 2022.

[59] C.R. Thomas and S.C. Maurice. *Managerial economics, Foundations of business analysis and strategy.* McGraw-Hill Education, New York, 2016.

[60] A. Tversky and D. Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, -(-):1124–1131, 1974.

[61] A.M. van der Bles, S. van der Linden, A.L.J. Freeman, J. Mitchell, A.B. Galvao, and D.J. Spiegelhalter. Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, 6:181870(-):1–42, 2019.

[62] K. G. Volz and G. Gigerenzer. Cognitive processes in decisions under risk are not the same as in decisions under uncertainty. *Frontiers in Neuroscience*, 6(105):–, 2012.

[63] P.R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3)(-):324–342, 1960.

[64] M. Yar and C. Chatfield. Prediction intervals for the Holt-Winters forecasting procedure. *International Journal of Forecasting*, 6(1):127–137, 1990.

Index

AI-related risk, 7, 247, 256
amalgamation paradox, 4
artificial intelligence, 22, 42
axioms of probability, 26

Bayes's law, 36
Bayes's rule, 36
Bayes's theorem, 36
best fitting model, 60
Bootstrap, 110

communication of risk, 20
complete probability uncertainty, 186
conditional probability, 31
consequence, 13

data driven risk analysis, 18
data-driven risk analysis , 25
decision node, 216
decision theories, 17
decision tree, 215
decisions under risk, 19
digital attacks, 9

Ecological fallacy, 4
event, 26
events, exhaustive, 34
events, independent, 33
events, mutually exclusive , 34
exhaustive outcomes, 26
exploratory data analysis, EDA, 69

goodness-of-fit checks, 60

imprecise risk, 186
interpretation of probability, 29

joint probability, 29

knowledge, 13

marginal probability, 31
memoryless property, 147
model-based risk assessment, 6
multistage decisions, 215
mutually exclusive outcomes, 26

negative predictive value, 42

outcome, 13, 26

phased decisions, 215
positive predictive value, 41
precise risk, 185
preferences of stakeholders, 15
probabilistic risk analysis, 12
probability, classical, 27
probability, conditional, 31
probability, law of total, 34
probability, product rule, 33
probability, relative-frequency, 27
probability, subjective, 28

quantitative risk analysis, 12

random node, 216
risk, 5
risk analysis, 10
risk assessment, 10
risk communication, 41, 243
risk communication, three-point estimate, 104
risk descriptors, 13
risk evaluation, 10

sample space, 26
sensitivity, 41

sequential decisions, 215
Simpson's paradox, 4
Simpson's reversal, 4
social amplification or risk framework, 250
social cognitive theory, 250
specificity, 41
stakeholders, 10
strength of knowledge, 13

time series dataset, 58
time series, autocorrelation, 75, 78
time series, control, 54
time series, decomposition, 72
time series, detrended, 73
time series, EDA, 60
time series, forecast horizon, 104
time series, forecasting, 54, 58
time series, forecasting steps, 59
time series, Holt model, 93
time series, Holt-Winter's model, 94
time series, moving average, 73
time series, point forecast, 89, 104
time series, seasonal factors, 73
time series, smoothing, 58
tuxedo fallacy, 186

uncertainty, 7, 62
uncertainty measurement, 13
uncertainty quantification, 13
utilities, 15
utility function, 16

Yulo-Simpson effect, 4

ISBN 978-80-227-5341-8
SPEKTRUM STU Publishing