The Mathematics of Banking and Finance
For other titles in the Wiley Finance Series
please see www.wiley.com/finance
The Mathematics of Banking and Finance

Dennis Cox and Michael Cox

John Wiley & Sons, Ltd
Contents

Introduction xiii

1 Introduction to How to Display Data and the Scatter Plot 1
  1.1 Introduction 1
  1.2 Scatter Plots 2
  1.3 Data Identification 2
    1.3.1 An example of salary against age 2
  1.4 Why Draw a Scatter Plot? 3
  1.5 Matrix Plots 4
    1.5.1 An example of salary against age: Revisited 5

2 Bar Charts 7
  2.1 Introduction 7
  2.2 Discrete Data 7
  2.3 Relative Frequencies 8
  2.4 Pie Charts 12

3 Histograms 13
  3.1 Continuous Variables 13
  3.2 Cumulative Frequency Polygon 14
  3.3 Sturges’ Formula 20

4 Probability Theory 21
  4.1 Introduction 21
  4.2 Basic Probability Concepts 21
  4.3 Estimation of Probabilities 22
  4.4 Exclusive Events 22
  4.5 Independent Events 22
  4.6 Comparison of Exclusivity and Independence 23
  4.7 Venn Diagrams 23
  4.8 The Addition Rule for Probabilities 24
    4.8.1 A simple probability example using a Venn diagram 25
  4.9 Conditional Probability 25
    4.9.1 An example of conditional probability 26
## Contents

4.10 The Multiplication Rule for Probabilities 26
4.10.1 A classical example of conditional probability 27
4.11 Bayes’ Theorem 27
4.11.1 An example of Bayes’ theorem 28
4.11.2 Bayes’ theorem in action for more groups 29
4.11.3 Bayes’ theorem applied to insurance 29
4.12 Tree Diagram 30
4.12.1 An example of prediction of success 30
4.12.2 An example from an American game show: The Monty Hall Problem 34
4.13 Conclusion 35

5 Standard Terms in Statistics 37
5.1 Introduction 37
5.2 Maximum and Minimum 37
5.2.1 Mean 37
5.2.2 Median 38
5.2.3 Mode 39
5.3 Upper and Lower Quartile 39
5.4 MQNMQM Plot 40
5.5 Skewness 41
5.6 Variance and Standard Deviation 41
5.7 Measures for Continuous Data 44

6 Sampling 47
6.1 Introduction 47
6.2 Planning Data Collection 47
6.3 Methods for Survey Analysis 48
6.3.1 Random samples 49
6.3.2 Systematic sampling 49
6.3.3 Stratified sampling 49
6.3.4 Multistage sampling 50
6.3.5 Quota sampling 50
6.3.6 Cluster sampling 50
6.4 How It Can Go Wrong 50
6.5 What Might Be In a Survey? 51
6.6 Cautionary Notes 51

7 Probability Distribution Functions 53
7.1 Introduction 53
7.2 Discrete Uniform Distribution 53
7.2.1 Counting techniques 54
7.2.2 Combination 54
7.2.3 Permutation 55
7.3 Binomial Distribution 55
7.3.1 Example of a binomial distribution 56
8 Normal Distribution

8.1 Introduction
8.2 Normal Distribution
  8.2.1 A simple example of normal probabilities
  8.2.2 A second example of normal probabilities
8.3 Addition of Normal Variables
8.4 Central Limit Theorem
  8.4.1 An example of the Central Limit Theorem
8.5 Confidence Intervals for the Population Mean
  8.5.1 An example of confidence intervals for the population mean
8.6 Normal Approximation to the Binomial Distribution
  8.6.1 An example of the normal approximation to the binomial distribution
8.7 Normal Approximation to the Poisson Distribution
  8.7.1 An example of fitting a normal curve to the Poisson distribution

9 Comparison of the Means, Sample Sizes and Hypothesis Testing

9.1 Introduction
9.2 Estimation of the Mean
  9.2.1 An example of estimating a confidence interval for an experimental mean
9.3 Choice of the Sample Size
  9.3.1 An example of selecting sample size
9.4 Hypothesis Testing
  9.4.1 An example of hypothesis testing
9.5 Comparison of Two Sample Means
  9.5.1 An example of a two-sample t test
9.6 Type I and Type II Errors
  9.6.1 An example of type I and type II errors

10 Comparison of Variances

10.1 Introduction
10.2 Chi-Squared Test
  10.2.1 An example of the chi-squared test
10.3 F Test
  10.3.1 An example of the F test
  10.3.2 An example considering the normal distribution
## Contents

### 11 Chi-squared Goodness of Fit Test
- 11.1 Introduction 91
- 11.2 Contingency Tables 92
- 11.3 Multiway Tables 94
  - 11.3.1 An example of a four by four table 94

### 12 Analysis of Paired Data
- 12.1 Introduction 97
- 12.2 $t$ Test 97
- 12.3 Sign Test 98
- 12.4 The $U$ Test 99
  - 12.4.1 An example of the use of the $U$ test 101

### 13 Linear Regression
- 13.1 Introduction 103
- 13.2 Linear Regression 103
- 13.3 Correlation Coefficient 104
  - 13.3.1 An example of examining correlation 105
- 13.4 Estimation of the Uncertainties 109
- 13.5 Statistical Analysis and Interpretation of Linear Regression 110
- 13.6 ANOVA for Linear Regression 110
- 13.7 Equations for the Variance of $a$ and $b$ 112
- 13.8 Significance Test for the Slope 112
  - 13.8.1 An example of slope analysis 113
  - 13.8.2 A further example of correlation and linear regression 115

### 14 Analysis of Variance
- 14.1 Introduction 121
- 14.2 Formal Background to the ANOVA Table 121
- 14.3 Analysis of the ANOVA Table 122
- 14.4 Comparison of Two Causal Means 123
  - 14.4.1 An example of extinguisher discharge times 123
  - 14.4.2 An example of the lifetime of lamps 125

### 15 Design and Approach to the Analysis of Data
- 15.1 Introduction 129
- 15.2 Randomised Block Design 129
  - 15.2.1 An example of outsourcing 130
- 15.3 Latin Squares 131
- 15.4 Analysis of a Randomised Block Design 132
- 15.5 Analysis of a Two-way Classification 135
  - 15.5.1 An example of two-way analysis 137
  - 15.5.2 An example of a randomised block 140
  - 15.5.3 An example of the use of the Latin square 143

### 16 Linear Programming: Graphical Method
- 16.1 Introduction 149
16.2 Practical Examples
16.2.1 An example of an optimum investment strategy
16.2.2 An example of the optimal allocation of advertising

17 Linear Programming: Simplex Method
17.1 Introduction
17.2 Most Profitable Loans
17.2.1 An example of finance selection
17.3 General Rules
17.3.1 Standardisation
17.3.2 Introduction of additional variables
17.3.3 Initial solution
17.3.4 An example to demonstrate the application of the general rules for linear programming
17.4 The Concerns with the Approach

18 Transport Problems
18.1 Introduction
18.2 Transport Problem

19 Dynamic Programming
19.1 Introduction
19.2 Principle of Optimality
19.3 Examples of Dynamic Programming
19.3.1 An example of forward and backward recursion
19.3.2 A practical example of recursion in use
19.3.3 A more complex example of dynamic programming
19.3.4 The ‘Travelling Salesman’ problem

20 Decision Theory
20.1 Introduction
20.2 Project Analysis Guidelines
20.3 Minimax Regret Rule

21 Inventory and Stock Control
21.1 Introduction
21.2 The Economic Order Quantity Model
21.2.1 An example of the use of the economic order quantity model
21.3 Non-zero Lead Time
21.3.1 An example of Poisson and continuous approximation

22 Simulation: Monte Carlo Methods
22.1 Introduction
22.2 What is Monte Carlo Simulation?
22.2.1 An example of the use of Monte Carlo simulation: Theory of the inventory problem
22.3 Monte Carlo Simulation of the Inventory Problem
22.4 Queuing Problem 208
22.5 The Bank Cashier Problem 209
22.6 Monte Carlo for the Monty Hall Problem 212
22.7 Conclusion 214

23 Reliability: Obsolescence 215
23.1 Introduction 215
23.2 Replacement at a Fixed Age 215
23.3 Replacement at Fixed Times 217

24 Project Evaluation 219
24.1 Introduction 219
24.2 Net Present Value 219
24.2.1 An example of net present value 219
24.3 Internal Rate of Return 220
24.3.1 An example of the internal rate of return 220
24.4 Price/Earnings Ratio 222
24.5 Payback Period 222
24.5.1 Mathematical background to the payback period 222
24.5.2 Mathematical background to producing the tables 223

25 Risk and Uncertainty 227
25.1 Introduction 227
25.2 Risk 227
25.3 Uncertainty 227
25.4 Adjusting the Discount Rate 228
25.5 Adjusting the Cash Flows of a Project 228
25.5.1 An example of expected cash flows 228
25.6 Assessing the Margin of Error 229
25.6.1 An example of break-even analysis 229
25.7 The Expected Value of the Net Present Value 231
25.7.1 An example of the use of the distribution approach to the evaluation of net present value 231
25.8 Measuring Risk 232
25.8.1 An example of normal approximation 234

26 Time Series Analysis 235
26.1 Introduction 235
26.2 Trend Analysis 236
26.3 Seasonal Variations 236
26.4 Cyclical Variations 240
26.5 Mathematical Analysis 240
26.6 Identification of Trend 241
26.7 Moving Average 241
26.8 Trend and Seasonal Variations 242
26.9 Moving Averages of Even Numbers of Observations 244
26.10 Graphical Methods 247
# Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>Reliability</td>
<td>249</td>
</tr>
<tr>
<td>27.1</td>
<td>Introduction</td>
<td>249</td>
</tr>
<tr>
<td>27.2</td>
<td>Illustrating Reliability</td>
<td>249</td>
</tr>
<tr>
<td>27.3</td>
<td>The Bathtub Curve</td>
<td>249</td>
</tr>
<tr>
<td>27.4</td>
<td>The Continuous Case</td>
<td>251</td>
</tr>
<tr>
<td>27.5</td>
<td>Exponential Distribution</td>
<td>252</td>
</tr>
<tr>
<td>27.5.1</td>
<td>An example of exponential distribution</td>
<td>252</td>
</tr>
<tr>
<td>27.5.2</td>
<td>An example of maximum of an exponential distribution</td>
<td>254</td>
</tr>
<tr>
<td>27.6</td>
<td>Weibull Distribution</td>
<td>255</td>
</tr>
<tr>
<td>27.6.1</td>
<td>An example of a Weibull distribution</td>
<td>256</td>
</tr>
<tr>
<td>27.7</td>
<td>Log-Normal Distribution</td>
<td>257</td>
</tr>
<tr>
<td>27.8</td>
<td>Truncated Normal Distribution</td>
<td>260</td>
</tr>
<tr>
<td>28</td>
<td>Value at Risk</td>
<td>261</td>
</tr>
<tr>
<td>28.1</td>
<td>Introduction</td>
<td>261</td>
</tr>
<tr>
<td>28.2</td>
<td>Extreme Value Distributions</td>
<td>262</td>
</tr>
<tr>
<td>28.2.1</td>
<td>A worked example of value at risk</td>
<td>262</td>
</tr>
<tr>
<td>28.3</td>
<td>Calculating Value at Risk</td>
<td>264</td>
</tr>
<tr>
<td>29</td>
<td>Sensitivity Analysis</td>
<td>267</td>
</tr>
<tr>
<td>29.1</td>
<td>Introduction</td>
<td>267</td>
</tr>
<tr>
<td>29.2</td>
<td>The Application of Sensitivity Analysis</td>
<td>267</td>
</tr>
<tr>
<td>30</td>
<td>Scenario Analysis</td>
<td>271</td>
</tr>
<tr>
<td>30.1</td>
<td>Introduction to Scenario Analysis</td>
<td>271</td>
</tr>
<tr>
<td>30.2</td>
<td>Use of External Loss Data</td>
<td>271</td>
</tr>
<tr>
<td>30.3</td>
<td>Scaling of Loss Data</td>
<td>272</td>
</tr>
<tr>
<td>30.4</td>
<td>Consideration of Likelihood</td>
<td>272</td>
</tr>
<tr>
<td>30.5</td>
<td>Anonymised Loss Data</td>
<td>273</td>
</tr>
<tr>
<td>31</td>
<td>An Introduction to Neural Networks</td>
<td>275</td>
</tr>
<tr>
<td>31.1</td>
<td>Introduction</td>
<td>275</td>
</tr>
<tr>
<td>31.2</td>
<td>Neural Algorithms</td>
<td>275</td>
</tr>
<tr>
<td>Appendix</td>
<td>Mathematical Symbols and Notation</td>
<td>279</td>
</tr>
<tr>
<td>Index</td>
<td></td>
<td>285</td>
</tr>
</tbody>
</table>
Introduction

Within business in general and specifically within the banking industry, there are wide ranges of mathematical techniques that are in regular use. These are often embedded into computer systems, which means that the user of the system may be totally unaware of the mathematical calculations and assumptions that are being made. In other cases it would also appear that the banking industry uses mathematical techniques as a form of jargon to create its own mystique, effectively creating a barrier to entry to anyone seeking to join the industry. It also serves to effectively baffle clients with science.

But in practice things can be much worse than this. Business systems, including specifically those used by bankers or in treasury functions, make regular use of a variety of mathematical techniques without the users having a real appreciation of the objective of the technique, or of its limitations. The consequence of this is that a range of things can go wrong:

1. The user will not understand the output from the system and so will be unable to interpret the information that comes out.
2. The user will not appreciate the limitations in the modelling approach adopted, and will assume that the model works when it would not be valid in the circumstances under consideration.
3. The user may misinterpret the information arising and provide inaccurate information to management.
4. The user may not understand the uncertainties inherent in the model and may pass it to management without highlighting these uncertainties.
5. The user may use an invalid model to try to model something and come up with results that are not meaningful.
6. Management may not understand the information being provided to them by the analysts and may either ignore or misinterpret the information.

The consequence of this is that models and the mathematics that underpins them are one of the greatest risks that a business can encounter.

Within the banking industry the development of the rules for operational risk by the Bank for International Settlements have exacerbated the problem. In the past, operational areas would not be closely involved with mathematics, instead this would have been left to analysts, risk management and planning professionals. However, these new rules put a range of requirements on all levels of staff and have increased the incidence of the use of modelling in operational risk areas.
It is the challenge of this text to try to provide the reader with some understanding of the nature of the tools that they are using on a day-to-day basis. At present much of the mathematics are hidden – all the user sees is a menu of choices from which to select a particular approach. The system then produces a range of data, but without understanding, gives no information. Therefore we have attempted to provide these users with sufficient information to enable them to understand the basic nature of the concept and, in particular, any weaknesses or inherent problems.

In this work we attempt to remove the mystique of mathematical techniques and notation so that someone who has not done mathematics for many years will be able to gain some understanding of the issues involved. While we do use mathematical notation, this is either described in the chapter itself or in the Appendix on page 279. If you do not follow what we are trying to say with the mathematical notation, explanatory details are embedded within the chapters and the range of worked examples will provide the understanding you require.

Our objective is to try to reduce the number of times that we see the wrong model being used in the wrong place. Even at conferences and in presentations we often see invalid conclusions being drawn from incorrectly analysed material. This is an entry book to the subject. If you wish to know about any of the specific techniques included herein in detail, we suggest that you refer to more specialist works.
1.1 INTRODUCTION

The initial chapters of the book are related to data and how it should be portrayed. Often useful data is poorly served by poor data displays, which, while they might look attractive, are actually very difficult to interpret and mask trends in the data.

It has been said many times that ‘a picture is worth a thousand words’ and this ‘original’ thought has been attributed to at least two historical heavyweights (Mark Twain and Benjamin Disraeli). While tables of figures can be hard or difficult to interpret, some form of pictorial presentation of the data enables management to gain an immediate indication of the key issues highlighted within the data set. It enables senior management to identify some of the major trends within a complex data set without the requirement to undertake detailed mathematical work. It is important that the author of a pictorial presentation of data follows certain basic rules when plotting data to avoid introducing bias, either accidentally or deliberately, or producing inappropriate or misleading representations of the original data.

When asked to prepare a report for management which is either to analyse or present some data that has been accumulated, the first step is often to present it in a tabular format and then produce a simple presentation of the information, frequently referred to as a plot. It is claimed that a plot is interpreted with more ease than the actual data set out in some form of a table. Many businesses have standardised reporting packages, which enable data to be quickly transformed into a pictorial presentation, offering a variety of potential styles. While many of these software packages produce plots, they should be used with care. Just because a computer produces a graph does not mean it is an honest representation of the data. The key issue for the author of such a plot is to see if the key trends inherent in the data are better highlighted by the pictorial representation. If this is not the case then an alternative approach should be adopted.

Whenever you are seeking to portray data there are always a series of choices to be made:

1. What is the best way to show the data?
2. Can I amend the presentation so that key trends in the data are more easily seen?
3. Will the reader understand what the presentation means?

Often people just look at the options available on their systems and choose the version that looks the prettiest, without taking into consideration the best way in which the material should be portrayed.

Many people are put off by mathematics and statistics – perhaps rightly in many cases since the language and terminology are difficult to penetrate. The objective of good data presentation is not to master all the mathematical techniques, but rather to use those that are appropriate, given the nature of what you are trying to achieve.

In this chapter we consider some of the most commonly used graphical presentational approaches and try to assist you in establishing which is most appropriate for the particular
data set that is to be presented. We start with some of the simplest forms of data presentation, the scatter plot, the matrix plot and the histogram.

1.2 SCATTER PLOTS

Scatter plots are best used for data sets in which there is likely to be some form of relationship or association between two different elements included within the data. These different elements are generally referred to as variables. Scatter plots use horizontal and vertical axes to enable the author to input the information into the scatter plot, or, in mathematical jargon, to plot the various data points. This style of presentation effectively shows how one variable affects another. Such a relationship will reveal itself by highlighting any trend that will be apparent to the reader from a review of the chart.

1.3 DATA IDENTIFICATION

A scatter plot is a plot of the values of \( Y \) on the vertical axis, or ordinate, taken against the corresponding values of \( X \) on the horizontal axis, or abscissa. Here the letters \( X \) and \( Y \) are taken to replace the actual variables, which might be something like losses arising in a month (\( Y \)) against time (\( X \)).

- \( X \) is usually the independent variable.
- \( Y \) is usually the response or dependent variable that may be related to the independent variable.

We shall explain these terms further through consideration of a simple example.

1.3.1 An example of salary against age

Figure 1.1 presents the relationship between salary and age for 474 employees of a company. This type of data would be expected to show some form of trend since, as the staff gains experience, you would expect their value to the company to increase and therefore their salary to also increase.

The raw data were obtained from personnel records. The first individual sampled was 28.50 years old and had a salary of £16,080. To put this data onto a scatter plot we insert age onto the horizontal axis and salary onto the vertical axis. The different entries onto the plot are the 474 combinations of age and salary resulting from a selection of 474 employees, with each individual observation being a single point on the chart.

This figure shows that in fact for this company there is no obvious relation between salary and age. From the plot it can be seen that the age range of employees is from 23 to 65. It can also be seen that a lone individual earns a considerably higher salary than all the others and that starters and those nearing retirement are actually on similar salaries.

You will see that the length of the axis has been chosen to match the range of the available data. For instance, no employees were younger than 20 and none older than 70. It is not essential that the axis should terminate at the origin. The objective is to find the clearest way to show the data, so making best use of the full space available clearly makes sense. The process of starting from 20 for age and 6,000 for salaries is called truncation and enables the actual data to cover the whole of the area of the plot, rather than being stuck in one quarter.
1.4 WHY DRAW A SCATTER PLOT?

Having drawn the plot it is necessary to interpret it. The author should do this before it is passed to any user. The most obvious relationship between the variables $X$ and $Y$ would be a straight line or a linear one. If such a relationship can be clearly demonstrated then it will be of assistance to the reader if this is shown explicitly on the scatter plot. This procedure is known as linear regression and is discussed in Chapter 13.

An example of data where a straight line would be appropriate would be as follows. Consider a company that always charges out staff at £1,000 per day, regardless of the size of the contract and never allows discounts. That would mean that a one-day contract would cost £1,000 whereas a 7-day contract would cost £7,000 (seven times the amount per day). If you were to plot 500 contracts of differing lengths by taking the value of the contract against the number of days, then this would represent a straight line scatter plot.

In looking at data sets, various questions may be posed. Scatter plots can provide answers to the following questions:

- Do two variables $X$ and $Y$ appear to be related? Given what the scatter plot portrays, could this be used to give some form of prediction of the potential value for $Y$ that would correspond to a potential value of $X$?
- Are the two variables $X$ and $Y$ actually related in a straight line or linear relationship? Would a straight line fit through the data?
- Are the two variables $X$ and $Y$ instead related in some non-linear way? If the relationship is non-linear, will any other form of line be appropriate that might enable predictions of $Y$ to be made? Might this be some form of distribution? If we are able to use a distribution this will enable us to use the underlying mathematics to make predictions about the variables. This is discussed in Chapter 7.
- Does the amount by which \( Y \) changes depend on the amount by which \( X \) changes? Does the coverage or spread in the \( Y \) values depend on the choice of \( X \)? This type of analysis always helps to gain an additional insight into the data being portrayed.
- Are there data points that sit away from the majority of the items on the chart, referred to as outliers? Some of these may highlight errors in the data set itself that may need to be rechecked.

### 1.5 MATRIX PLOTS

Scatter plots can also be combined into multiple plots on a single page if you have more than two variables to consider. This type of analysis is often seen in investment analysis, for example, where there could be a number of different things all impacting upon the same data set. Multiple plots enable the reader to gain a better understanding of more complex trends hidden within data sets that include more than two variables. If you wish to show more than two variables on a scatter plot grid, or matrix, then you still need to generate a series of pairs of data to input into the plots. Figure 1.2 shows a typical example.

In this example four variables (\( a, b, c, d \)) have been examined by producing all possible scatter plots. Clearly while you could technically include even more variables, this would make the plot almost impossible to interpret as the individual scatter plots become increasingly small.

Returning to the analysis we set out earlier of salary and age (Figure 1.1), let us now differentiate between male salaries and female salaries, by age. This plot is shown as Figure 1.3.

![Figure 1.2 Example of a matrix plot.](image-url)
1.5.1 An example of salary against age: Revisited

It now becomes very clear that women have the majority of the lower paid jobs and that their salaries appear to be even less age dependent than those of men. This type of analysis would be of interest to the Human Resources function of the company to enable it to monitor compliance with legislation on sexual discrimination, for example. Of course there may be a range of other factors that need to be considered, including differentiating between full- and part-time employment by using either another colour or plotting symbol.

It is the role of the data presentation to facilitate the highlighting of trends that might be there. It is then up to the user to properly interpret the story that is being presented.

In summary the scatter plot attempts to uncover any relationship in the data. ‘Relationship’ means that there may be some structural association between two variables $X$ and $Y$. Scatter plots are a useful diagnostic tool for highlighting whether there is any form of potential association, but they cannot in themselves suggest an underlying cause-and-effect mechanism. A scatter plot can never prove cause and effect; this needs to be achieved through further detailed investigation, which should use the scatter plot to set out the areas where the investigation into the underlying data should commence.
2.1 INTRODUCTION

While a scatter plot is a useful way to show a lot of data on one chart, it does tend to need a reasonable amount of data and also quite a bit of analysis. By moving the data into discrete bands you are able to formulate the information into a bar chart or histogram. Bar charts (with vertical bars) or pie charts (where data is shown as segments of a pie) are probably the most commonly used of all data presentation formats in practice. Bar charts are suitable where there is discrete data, whereas histograms are more suitable when you have continuous data. Histograms are considered in Chapter 3.

2.2 DISCRETE DATA

Discrete data refers to a series of events, results, measurements or readings that may occur over a period of time. It may then be classified into categories or groups. Each individual event is normally referred to as an observation. In this context observations may be grouped into multiples of a single unit, for example:

- The number of transactions in a queue
- The number of orders received
- The number of calls taken in a call centre.

Since discrete data can only take integer values, this is the simplest type of data that a firm may want to present pictorially. Consider the following example:

A company has obtained the following data on the number of repairs required annually on the 550 personal computers (PCs) registered on their fixed asset ledger. In each case, when there is to be a repair to a PC, the registered holder of the PC is required to complete a repair record and submit this to the IT department for approval and action. There have been 341 individual repair records received by the IT department in a year and these have been summarised by the IT department in Table 2.1, where the data has been presented in columns rather than rows. This recognises that people are more accustomed to this form of presentation and therefore find it easier to discern trends in the data if it is presented in this way. Such a simple data set could also be represented by a bar chart. This type of presentation will assist the reader in undertaking an initial investigation of the data at a glance as the presentation will effectively highlight any salient features of the data. This first examination of the data may again reveal any extreme values (outliers), simple mistakes or missing values.

Using mathematical notation, this data is replaced by \((x_i, f_i: i = 1, \ldots, n)\). The notation adopted denotes the occurrence of variable \(x_i\) (the number of repairs) with frequency \(f_i\) (how often this happens). In the example, when \(i = 1\), \(x_1\) is 0 and \(f_1\) is 295, because 0 is the first observation, which is that there have been no repairs to these PCs. Similarly when \(i = 2\), \(x_2\) is 1 and \(f_2\) is 190 and so on until the end of the series, which is normally shown as the letter **table**.
Table 2.1 Frequency of repairs to PCs

<table>
<thead>
<tr>
<th>Number of repairs</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>295</td>
</tr>
<tr>
<td>1</td>
<td>190</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
</tr>
<tr>
<td>35</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>550</td>
</tr>
</tbody>
</table>

In this data set \( n = 6 \), \( x_6 \) has the value 5 and \( f_6 \) is 2. If the variable \( x \) is plotted on the horizontal axis and the frequency on the vertical axis, the vertical column of height \( f_i \) occurs at the position where there are \( x_i \) repairs. As explained below, a scaled form of the data is adopted since there needs to be some way to standardise the data to enable comparisons to be made between a number of plots.

Certain basic rules should be followed when plotting the data to ensure that the bar chart is an effective representation of the underlying data. These include the following:

- Every plot must be correctly labelled. This means a label on each axis and a heading for the graph as a whole.
- Every bar in the plot must be of an equal width. This is particularly important, since the eye is naturally drawn to wider bars and gives them greater significance than would actually be appropriate.
- There should be a space between adjacent bars, stressing the discrete nature of the categories.
- It is sensible to plot relative frequency vertically. While this is not essential it does facilitate the comparison of two plots.

2.3 RELATIVE FREQUENCIES

The IT department then calculates relative frequencies and intends to present them as another table. The relative frequency is basically the proportion of occurrences. This is a case where the superscript is used to denote successive frequencies. The relative frequency of \( f_i \) is shown as \( f'_i \). To obtain the relative frequencies \( (f'_i: i = 1, \ldots, 6) \), the observed frequency is divided by the total of all the observations, which in this case is 550.

This relationship may be expressed mathematically as follows: \( f'_i = f_i / F \), where \( F = f_1 + \ldots + f_6 \), in other words, the total of the number of possible observations. It is usual to write the expression \( f_1 + \ldots + f_6 \) as \( \sum_{i=1}^{6} f_i \) or, in words, ‘the sum from 1 to 6 of \( f_i \)’. This gives the property that the relative frequencies sum to 1. This data is best converted into a bar chart or histogram to enable senior management to quickly review the data set. This new representation of the data is shown in Table 2.2.

The total number of events is 550; therefore this is used to scale the total data set such that the total population occurs with a total relative frequency of 1. This table represents a subsidiary step in the generation of a bar chart. It is not something that would normally be presented to management since it is providing a greater level of information than they are likely to require and analysis is difficult without some form of pictorial presentation. The bar chart will represent a better representation of the data and will make it easier for the reader to analyse the data quickly. The resulting bar chart is shown in Figure 2.1.
Here the zero has been shifted on the horizontal axis away from the vertical axis to enable the first bar to be clearly reviewed in a form consistent with all of the other columns. While inclusion of an origin for the vertical axis is essential, an origin for a horizontal axis is only required if the observation ‘0’ was included in the original data. In general, we do not recommend the use of three-dimensional representations since the eye may be misled by the inclusion of perspective into exaggerating the importance of bars of similar height by subconsciously assigning them more weight. They may look attractive, but they do not assist the reader in discovering key trends within the data set itself. Similarly the author should always be careful in using a variety of colours since this could have the unfortunate consequence of reinforcing a specific part of the data set and should therefore be used with care.

From the plot it may be concluded that while the majority of PCs are actually trouble free, a significant proportion, 10%, exhibit two failures. While very few exhibit more than three or more failures, it is these that need investigating and any common causes of these faults identified and action taken by management. Obviously this is a simple data set and the

<table>
<thead>
<tr>
<th>Number of repairs</th>
<th>Frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>295</td>
<td>0.5364</td>
</tr>
<tr>
<td>1</td>
<td>190</td>
<td>0.3455</td>
</tr>
<tr>
<td>2</td>
<td>53</td>
<td>0.0964</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.0091</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>0.0091</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0.0036</td>
</tr>
<tr>
<td>Total</td>
<td>550</td>
<td>1</td>
</tr>
</tbody>
</table>

**Figure 2.1**  Bar chart of repairs to PCs.
information should have been clear from Table 2.2, but management will be able to save time by quickly reviewing the data as shown in the chart.

It is always best to include a narrative explanation to guide the reader to identify the key trends in the data set presented. It is also important for the author to ensure that anything that is to be compared is presented on equal scales, otherwise the relationships between the variables could be distorted. For extensive data sets the plot provides a concise summary of the raw data.

Here is an example of the use of comparative data.

An insurance company introduces a new homeowner’s policy. It covers the same range of risks as the traditional policy with the added benefit of an additional ‘new for old’ replacement clause. The analyst has been asked to assess whether the frequency of claim type varies between the two options.

Both policies cover

1. Hail damage – to roofs, air-conditioning units, windows and fences
2. Wind damage – to roofs, fences and windows
3. Water damage – any damage caused by leaking pipes, toilets, bathtubs, shower units, sinks, fridge freezers, dishwashers and washing machines
4. Fire damage
5. Vandalism
6. Smoke damage.

The analyst was able to obtain the information shown in Table 2.3 from the records of the insurance company.

<table>
<thead>
<tr>
<th>Claim type</th>
<th>Traditional policy</th>
<th>New for old policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hail damage</td>
<td>1,029</td>
<td>98</td>
</tr>
<tr>
<td>Wind damage</td>
<td>449</td>
<td>47</td>
</tr>
<tr>
<td>Water damage</td>
<td>2,730</td>
<td>254</td>
</tr>
<tr>
<td>Fire damage</td>
<td>4,355</td>
<td>453</td>
</tr>
<tr>
<td>Vandalism</td>
<td>70</td>
<td>7</td>
</tr>
<tr>
<td>Smoke damage</td>
<td>1,458</td>
<td>159</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>10,091</strong></td>
<td><strong>1,018</strong></td>
</tr>
</tbody>
</table>

The difference in the number of claims for each policy makes any comparisons difficult and trends within the data are unclear. It is easy to see that most claims are for fire damage and the least for vandalism, but relative performance is hard to identify. The analyst then converted the data into a series of relative frequencies, which are set out in Table 2.4, and then used them to produce the bar chart shown as Figure 2.2.

The similarity between the two policies is now clear. However, some key questions need to be addressed.

**How might the chart be improved?**

- It might be useful to prioritise the type of claim, showing those that occur least frequently to the right of the plot.
Table 2.4  Relative frequency of claim type

<table>
<thead>
<tr>
<th>Claim type</th>
<th>Frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Traditional policy</td>
<td>New for old policy</td>
</tr>
<tr>
<td>Hail damage</td>
<td>1,029</td>
<td>98</td>
</tr>
<tr>
<td>Wind damage</td>
<td>449</td>
<td>47</td>
</tr>
<tr>
<td>Water damage</td>
<td>2,730</td>
<td>254</td>
</tr>
<tr>
<td>Fire damage</td>
<td>4,355</td>
<td>453</td>
</tr>
<tr>
<td>Vandalism</td>
<td>70</td>
<td>7</td>
</tr>
<tr>
<td>Smoke damage</td>
<td>1,458</td>
<td>159</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>10,091</strong></td>
<td><strong>1,018</strong></td>
</tr>
</tbody>
</table>

What information has been lost?

- There is no information about the number of claims for each policy, only the relative frequency.
- There is no information about the cost, since all claims have been treated equally.
- There is no calendar information. The new policy would be expected to exhibit a growing number of claims as more customers adopt it. Older policies will have been in force for longer and therefore are more likely to exhibit claims.

This is a simple but useful form of data presentation, since it enables us to see simple trends in the data. There are more complex methods of showing data, which we consider in later chapters.

Figure 2.2  Bar chart of claim type against frequency.
Pie charts are often used in business to show data where there is a contribution to the total population from a series of events. Contribution to profit by the divisions of a company can be shown as a pie chart, which operates by transforming the lines in a table into segments of a circle.

Taking the information from Table 2.2, this can easily be changed into percentages as shown in Table 2.5. This can also be produced as a pie chart, as shown in Figure 2.3.

<table>
<thead>
<tr>
<th>Number of repairs</th>
<th>Frequency</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>295</td>
<td>53.64</td>
</tr>
<tr>
<td>1</td>
<td>190</td>
<td>34.55</td>
</tr>
<tr>
<td>2</td>
<td>53</td>
<td>9.64</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>9.10</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>9.10</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3.60</td>
</tr>
<tr>
<td>Total</td>
<td>550</td>
<td>100</td>
</tr>
</tbody>
</table>

Figure 2.3 Example of a pie chart.

The one advantage of the pie chart is that you can quickly see which is the largest segment. On the other hand, that is also obvious from a quick look at the underlying table. The problem with pie charts is that very little information is actually shown – again all you have is the relative frequency. Further, it is difficult to compare different pie charts with each other. As a presentation to make it easy for the reader to understand the trends in data, it is generally rather poor. However, in practice it is a well-used and popular form of data presentation.
3.1 CONTINUOUS VARIABLES

The next issue is how to present observations of continuous variables successfully, for example:

- Height or weight of a company’s employees
- The time taken by a series of teams to process an invoice.

While we use a bar chart where there is discrete data, a histogram is employed where there is continuous data. Many of the basic rules employed for bar charts are also used in histograms. However, there is one additional requirement: there is a need to standardise class intervals. This has an echo from bar charts, where it was insisted that all bars were to be of equal width.

The actual form of presentation will be based on the specific data set selected. Displayed in Table 3.1 is some data collected on overtime payments made to the processing and IT functions within a financial institution. All such payments are made weekly and it is expected that staff will work some overtime to supplement their salaries.

The range is referred to as the class interval and the following notation is adopted:

- \( L_i \) = the left point of the \( i \)th class interval,
- \( R_i \) = the right point, and
- \( f_i \) = the observed frequency in the interval.

In the example, \( L_1 \) is £210, \( R_1 \) is £217 with a frequency \( f_1 \) of 1. So one employee earns a salary in the range, \( £210 \leq \text{salary} < £217 \). The final interval has \( L_{25} \) at £355 with \( R_{25} \) at £380 and a frequency \( f_{25} \) of 4.

There are two issues with plotting this type of data.

Firstly (which is not a problem here), there is the possibility that a right end point may not be identical to the following left end point, so that a gap exists. For example, if \( R_1 = 216.5 \) and \( L_2 = 217 \), then an intermediate value of 216.75 would be used to summarise the data, and this would be adopted for both end points.

Secondly, a problem is raised by unequal class intervals which occurs when the difference between the left end point and the right end point is not a constant throughout the data set. Using the notation where \([335, 355)\) means \(335 \leq x < 355\), there may for instance in another data set be 12 items in the range \([335, 355)\) and six in the range \([237, 241)\), and to compare these values it is best to think of the 12 items as being 12/5 of an item in each interval \([335, 336), \ldots, [354, 355)\). Using mathematical notation you should replace \(f_i\) by \(f_i'' = f_i/[(R_i - L_i)F]\), where \(F = f_1 + \ldots + f_{25}\). This has two important properties: (1) it correctly represents the proportional height for each range, and (2) it forces the total area under the graph to become 1.

The data from Table 3.1 needs to be prepared for plotting, as shown in Table 3.2, with the resulting histogram shown in Figure 3.1.

An alternative way to show the same information is the cumulative frequency polygon or ogive.
### Table 3.1  Overtime earnings for processing staff

<table>
<thead>
<tr>
<th>Overtime earnings (£ per week)</th>
<th>Number of staff</th>
</tr>
</thead>
<tbody>
<tr>
<td>210</td>
<td>217</td>
</tr>
<tr>
<td>217</td>
<td>221</td>
</tr>
<tr>
<td>221</td>
<td>225</td>
</tr>
<tr>
<td>225</td>
<td>229</td>
</tr>
<tr>
<td>229</td>
<td>233</td>
</tr>
<tr>
<td>233</td>
<td>237</td>
</tr>
<tr>
<td>237</td>
<td>241</td>
</tr>
<tr>
<td>241</td>
<td>245</td>
</tr>
<tr>
<td>245</td>
<td>249</td>
</tr>
<tr>
<td>249</td>
<td>253</td>
</tr>
<tr>
<td>253</td>
<td>257</td>
</tr>
<tr>
<td>257</td>
<td>261</td>
</tr>
<tr>
<td>261</td>
<td>265</td>
</tr>
<tr>
<td>265</td>
<td>269</td>
</tr>
<tr>
<td>269</td>
<td>273</td>
</tr>
<tr>
<td>273</td>
<td>277</td>
</tr>
<tr>
<td>277</td>
<td>281</td>
</tr>
<tr>
<td>281</td>
<td>285</td>
</tr>
<tr>
<td>285</td>
<td>295</td>
</tr>
<tr>
<td>295</td>
<td>305</td>
</tr>
<tr>
<td>305</td>
<td>315</td>
</tr>
<tr>
<td>315</td>
<td>325</td>
</tr>
<tr>
<td>325</td>
<td>335</td>
</tr>
<tr>
<td>335</td>
<td>355</td>
</tr>
<tr>
<td>355</td>
<td>380</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1,706</strong></td>
</tr>
</tbody>
</table>

### 3.2 CUMULATIVE FREQUENCY POLYGON

A cumulative frequency polygon is constructed to indicate what proportions of the observations have been achieved prior to a particular point on the horizontal axis. Employing the relative frequencies, using the calculations in section 2.3, there have been 0 observations before $L_1$ and $f_1'$ prior to reaching $R_1$. So the points $(0, L_1)$ and $(f_1', R_1)$ are joined. Similarly the proportion $f_1' + f_2'$ is observed by $R_2$ and the second line segment can then be drawn. In general, the cumulative distribution is defined by $F_i = f_1' + \ldots + f_i'$. The final right end point must correspond to the cumulative frequency of 1.

The calculations are summarised in Table 3.3 with the results presented in Figure 3.2.

The same employer has data on the weekly overtime earnings of contractors analysed by the sex of the contractor (Table 3.4). Here a histogram can be employed to represent the data and enable the reader to make comparisons.

The figures in Table 3.5 are then required to enable the histograms in Figures 3.3 and 3.4 to be prepared.

To enable these two charts to be compared, the data should be presented on axes that have identical scales. To further facilitate comparison it would be worth while to overlay the figures, as shown in Figure 3.5.