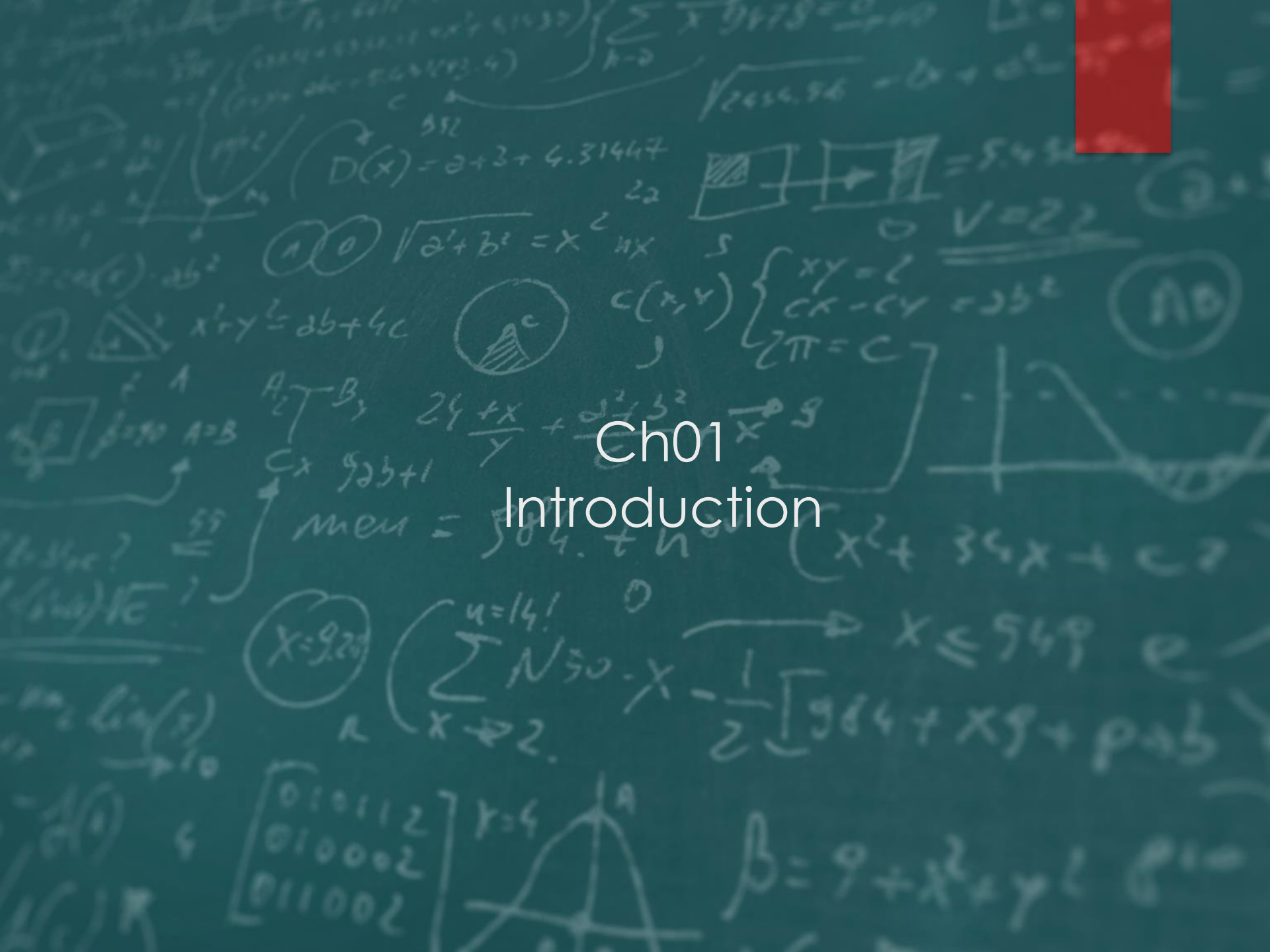
Statistics for Economics,



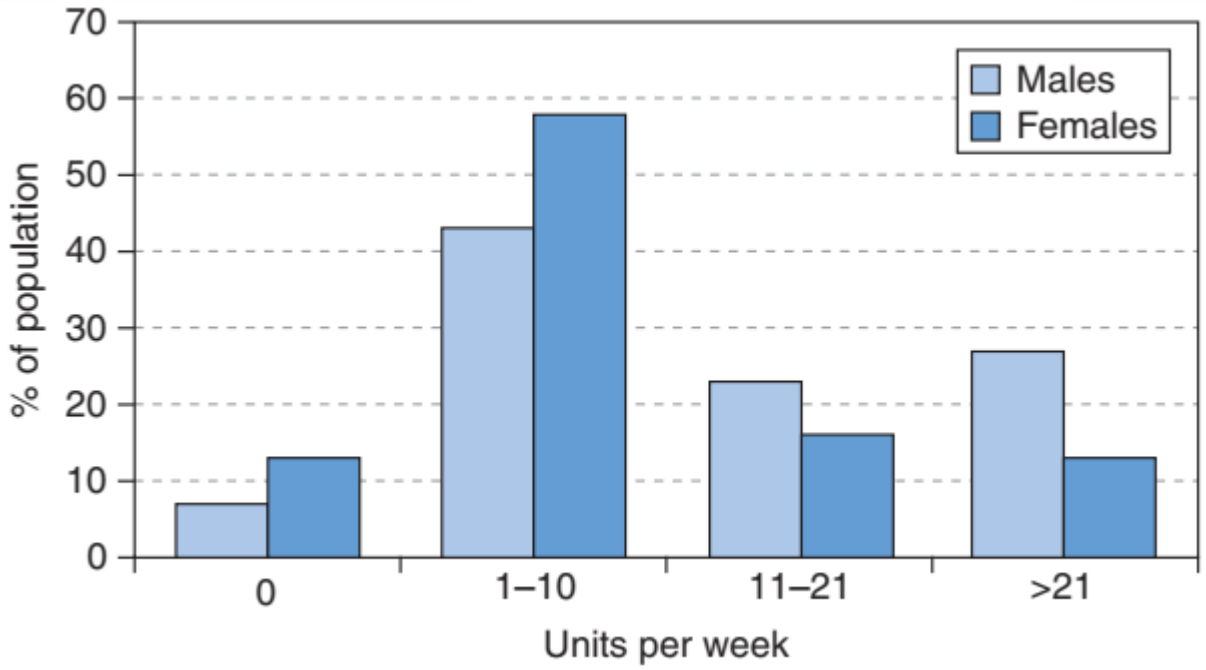
Ch01
Introduction

1.1 Introduction

- Statistics is a subject which can be (and is) applied to every aspect of our lives.
- In common usage people think of statistics as numerical data—the unemployment rate last month, total government expenditure last year, the number of impaired drivers charged during the recent holiday season, the crime rates of cities, and so forth.
- Although there is nothing wrong with viewing statistics in this way, we are going to take a deeper approach.
- We will view statistics the way professional statisticians view it—as a methodology for collecting, classifying, summarizing, organizing, presenting, analysing and interpreting numerical information.
- The subject of statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting data (numerical information). The power and utility of statistics derives from being able to draw valid conclusions (inferences), and make reasonable decisions, on the basis the available data.
- The term statistics is also used in a much narrower sense when referring to the data themselves or various other numbers derived from any given set of data

1.1 Introduction

- **Two types of statistics**
- Descriptive statistics are used to summarise information which would otherwise be too complex to take in, by means of techniques such as averages and graphs.
- The graph shown in Figure 1.1 is an example, summarising consuming habits in the United Kingdom.

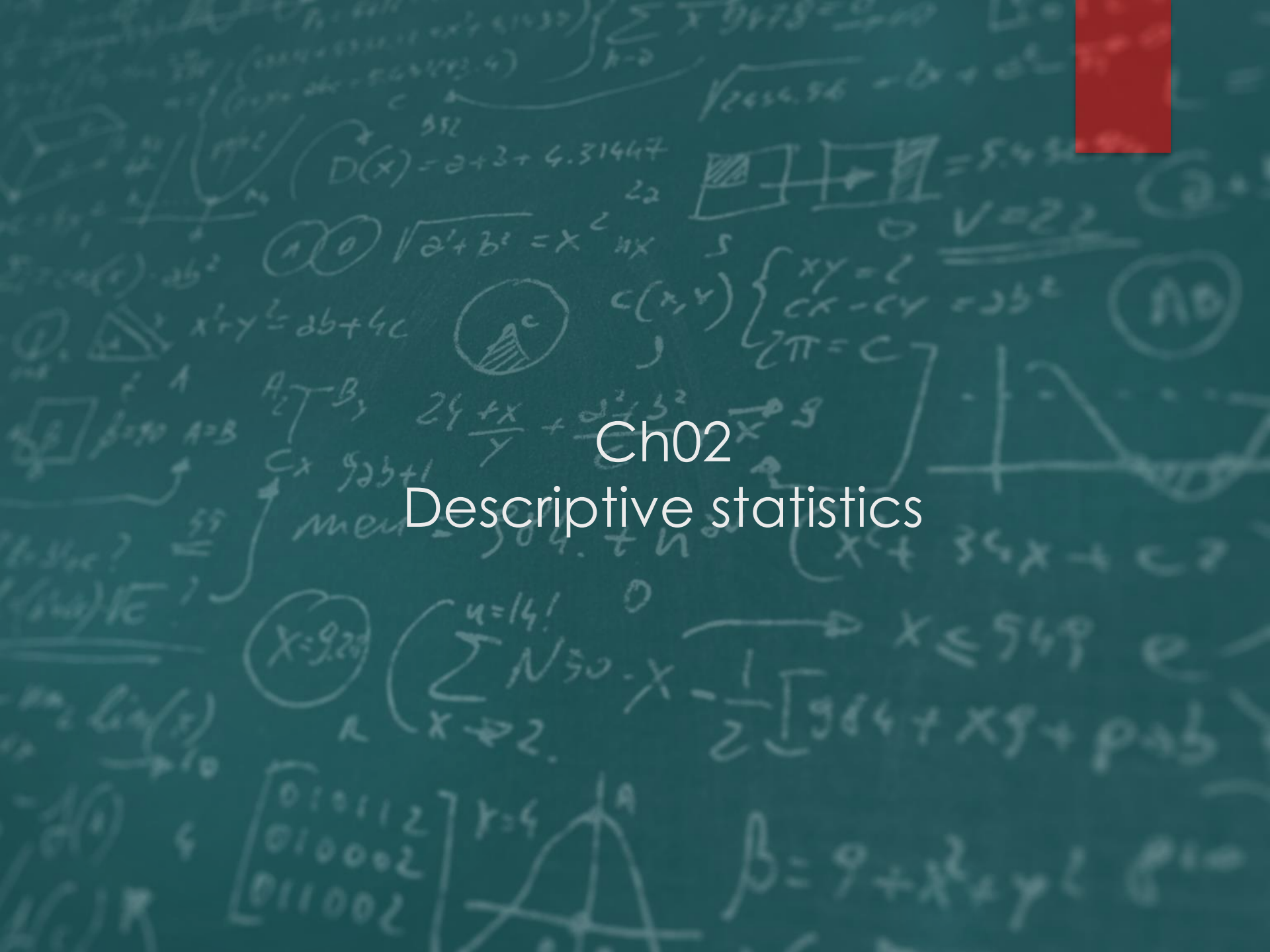


1.1 Introduction

- The graph reveals, for instance, that about 43% of men and 57% of women consume between 1 and 10 units of the good per week.
- The graph also shows that men tend to consume more than women, with higher proportions consuming 11 to 20 units and over 21 units per week.
- This simple graph has summarised a vast amount of information, the consumption levels of about 45 million adults.
- Even so, it is not perfect and much information is hidden.
- It is not obvious from the graph that the average consumption of men is 16 units per week, of women only 6 units.
- From the graph, you would probably have expected the averages to be closer together. This shows that graphical and numerical summary measures can complement each other.

1.1 Introduction

- Graphs can give a very useful visual summary of the information but are not very precise. For example, it is difficult to convey in words the content of a graph; you have to see it.
- Numerical measures such as the average are more precise and are easier to convey to others.
- Imagine you had data for student consumption; how do you think this would compare to the graph? It would be easy to tell someone whether the average is higher or lower, but comparing the graphs is difficult without actually viewing them.
- Conversely, the average might not tell you important information. The problem of ‘binge’ consuming is related not to the average (though it does influence the average) but to extremely high consumption by some individuals.
- Other numerical measures (or an appropriate graph) are needed to address the issue.
- Statistical inference, the second type of statistics covered, concerns the relationship between a sample of data and the population (in the statistical sense, not necessarily human) from which it is drawn. In particular, it asks what inferences can be validly drawn about the population from the sample. Sometimes the sample is not representative of the population (either due to bad sampling procedures or simply due to bad luck) and does not give us a true picture of reality.



Ch02
Descriptive statistics

2.1 Introduction

- The aim of descriptive statistical methods is simple: to present information in a clear, concise and accurate manner.
- The difficulty in analysing many phenomena, be they economic, social or otherwise, is that there is simply too much information for the mind to assimilate.
- The task of descriptive methods is therefore to summarise all this information and draw out the main features, without distorting the picture.
- The appropriate method of analysing the data will depend on a number of factors: the type of data under consideration, the sophistication of the audience and the ‘message’ which it is intended to convey.
- One would use different methods to persuade academics of the validity of one’s theory about inflation than one would use to persuade consumers that Brand X powder washes whiter than Brand Y.

2.1 Introduction

- To illustrate the use of the various methods, three different topics are covered in this chapter.
- **First,** we look at the relationship between educational attainment and employment prospects. Do higher qualifications improve your employment chances? The data come from people surveyed in 2009, so we have a sample of **cross-section data** giving an illustration of the situation at one point in time.
- We will look at the distribution of educational attainments amongst those surveyed, as well as the relationship to employment outcomes. In this example, we simply count the numbers of people in different categories (e.g. the number of people with a degree qualification who are employed).

2.1 Introduction

- **Second,** we examine the distribution of wealth in the United Kingdom in 2005. The data are again cross-section, but this time we can use more sophisticated methods since wealth is measured on **a ratio scale.**
- Someone with £200 000 of wealth is twice as wealthy as someone with £100 000, for example, and there is a meaning to this ratio. In the case of education, one cannot say with any precision that one person is twice as educated as another.
- The educational categories may be ordered (so one person can be more educated than another, although even that may be ambiguous) but we cannot measure the 'distance' between them. We therefore refer to educational attainment being measured on an **ordinal scale**.
- In contrast, there is not an obvious natural ordering to the three employment categories (employed, unemployed, inactive), so this is measured on a **nominal scale.**

2.1 Introduction

- **Third,** we look at national spending on investment over the period 1977–2009. This is time-series data since we have a number of observations on the variable measured at different points in time. Here it is important to take account of the time dimension of the data: things would look different if the observations were in the order 1977, 1989, 1982, . . . rather than in correct time order.
- We also look at the relationship between two variables, investment and output, over that period of time and find appropriate methods of presenting it.
- In all three cases, we make use of both graphical and numerical methods of summarising the data. Although there are some differences between the methods used in the three cases, these are not watertight compartments: the methods used in one case might also be suitable in another, perhaps with slight modification.
- Part of the skill of the statistician is to know which methods of analysis and presentation are best suited to each particular problem.

2.2 Summarising data using graphical techniques

- Education improves one's life chances in various ways, one of the possible benefits being that it reduces the chances of being out of work.
- But by how much does it reduce those chances? We shall use a variety of graphical techniques to explore the question.
- The raw data for this investigation come from the Education and Training Statistics for the UK 2009.
- Some of these data are presented in Table 1.1 and show the numbers of people by employment status (either in work, unemployed or inactive, i.e. not seeking work) and by educational qualification (higher education, A levels, other qualification or no qualification).
- The table gives a cross-tabulation of employment status by educational qualification and is simply a count (the frequency) of the number of people falling into each of the 12 cells of the table.
- For example, there were 9,713,000 people in work who had experience of higher education.
- This is part of a total of nearly 38 million people of working age. Note that the numbers in the table are in thousands, for the sake of clarity.

2.2 Summarising data using graphical techniques

Table 1.1 Economic status and educational qualifications, 2009 (numbers in 000s)

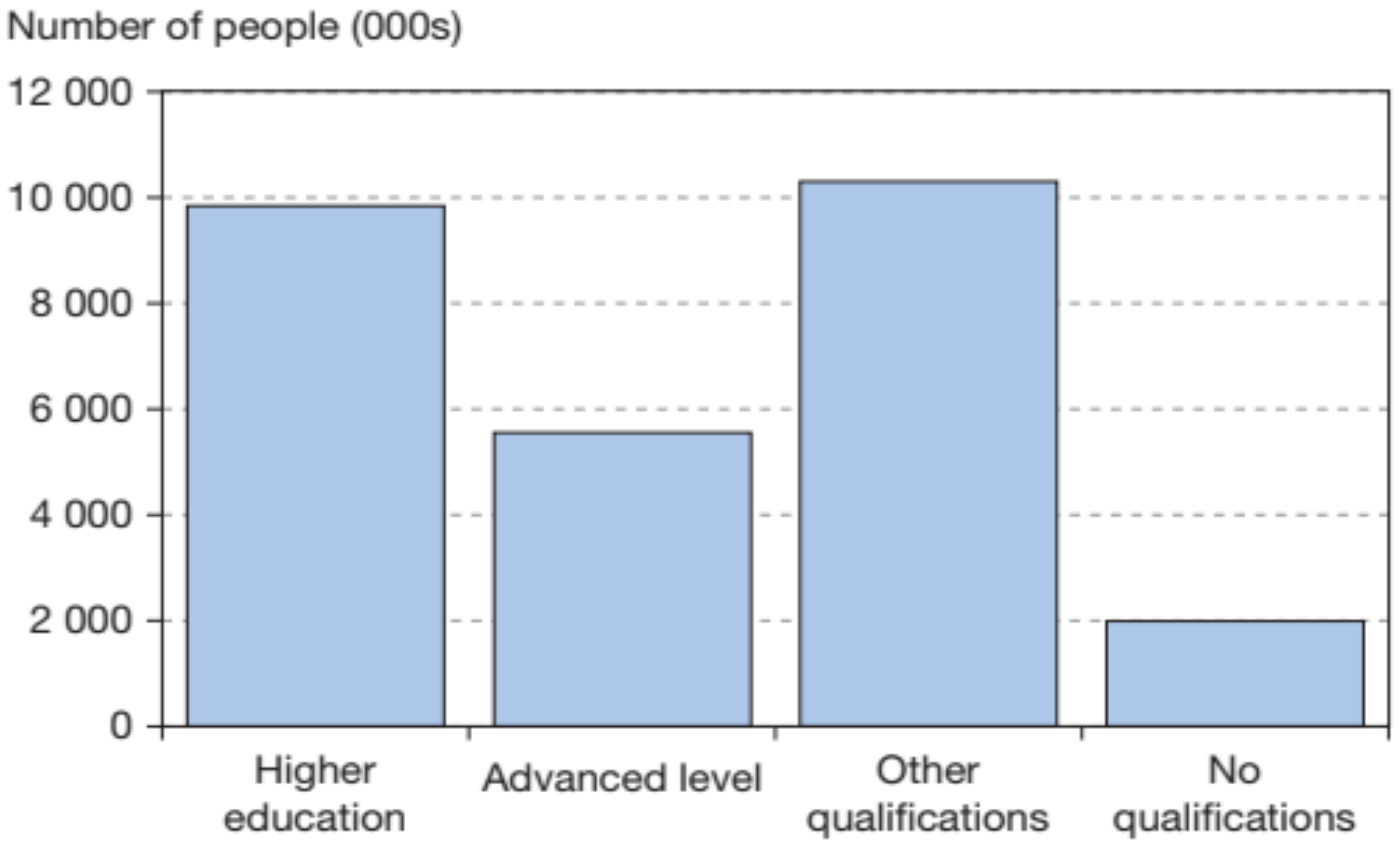
	Higher education	A levels	Other qualification	No qualification	Total
In work	9 713	5 479	10 173	1 965	27 330
Unemployed	394	432	1 166	382	2 374
Inactive	1 256	1 440	3 277	2 112	8 085
Total	11 363	7 351	14 616	4 459	37 789

Source: Adapted from Department for Children, Schools and Families, Education and Training Statistics for the UK 2009, http://dera.ioe.ac.uk/15353/, contains public sector information licensed under the Open Government Licence (OGL) v3.0. http://www.nationalarchives.gov.uk/doc/open-government-licence/open-government

2.2 Summarising data using graphical techniques

- From the table, we can see some messages from the data; for example, being unemployed or inactive seems to be more prevalent amongst those with lower qualifications: $56\%\ (= (382 + 2112)/4458)$ of those with no qualifications are unemployed or inactive compared to only about 15% of those with higher education.
- However, it is difficult to go through the table by eye and pick out these messages. It is easier to draw some graphs of the data and use them to form conclusions.
- ***The bar chart***
- The bar chart summarises the educational qualifications of those in work, (the data in the first row of the Table.
- The four educational categories are arranged along the horizontal (x) axis, while the frequencies are measured on the vertical (y) axis. The height of each bar represents the numbers in work for that category.
- The biggest groups are those with higher education and those with ‘other qualifications’ which are of approximately equal size.
- The graph also shows that there are relatively few people working who have no qualifications.

2.2 Summarising data using graphical techniques

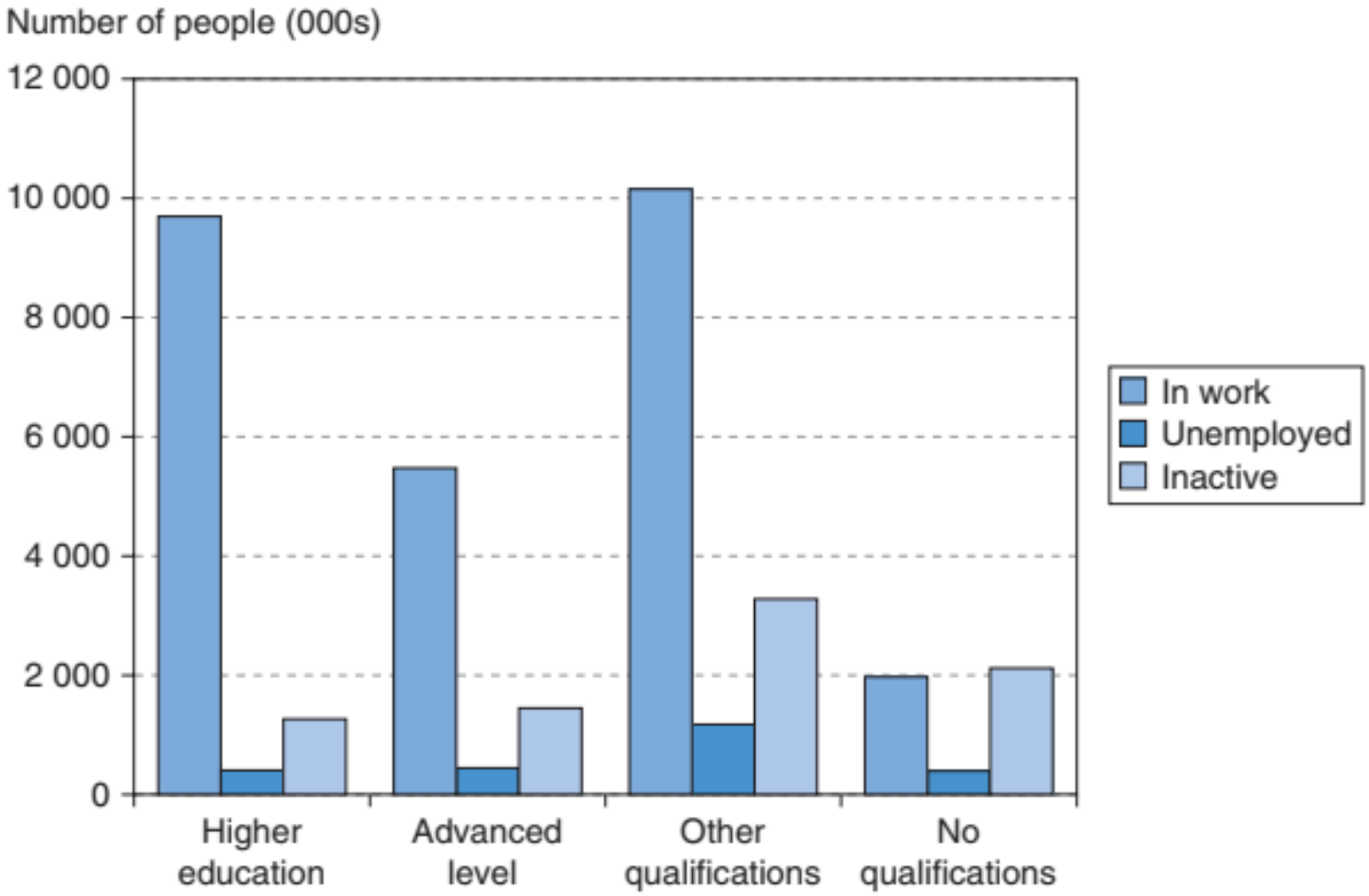


Note: The height of each bar is determined by the associated frequency. The first bar is 9713 units high, the second is 5479 units high and so on. The ordering of the bars could be reversed ('no qualifications' becoming the first category) without altering the message.

2.2 Summarising data using graphical techniques

- It is important to realise what the graph does not show: it does not say anything about your likelihood of being in work, given your educational qualifications.
- For that, we would need to compare the proportions of each education category in work; for the moment, we are only looking at the absolute numbers.
- It would be interesting to compare the distribution in Figure with those for the unemployed and inactive categories.
- This is done in the next Figure, which adds bars for these other two categories.
- This multiple bar chart shows that, as for the 'in work' category, amongst the inactive and unemployed, the largest group consists of those with 'other' qualifications (which are typically vocational qualifications).
- These findings simply reflect the fact that 'other qualifications' is the largest category.

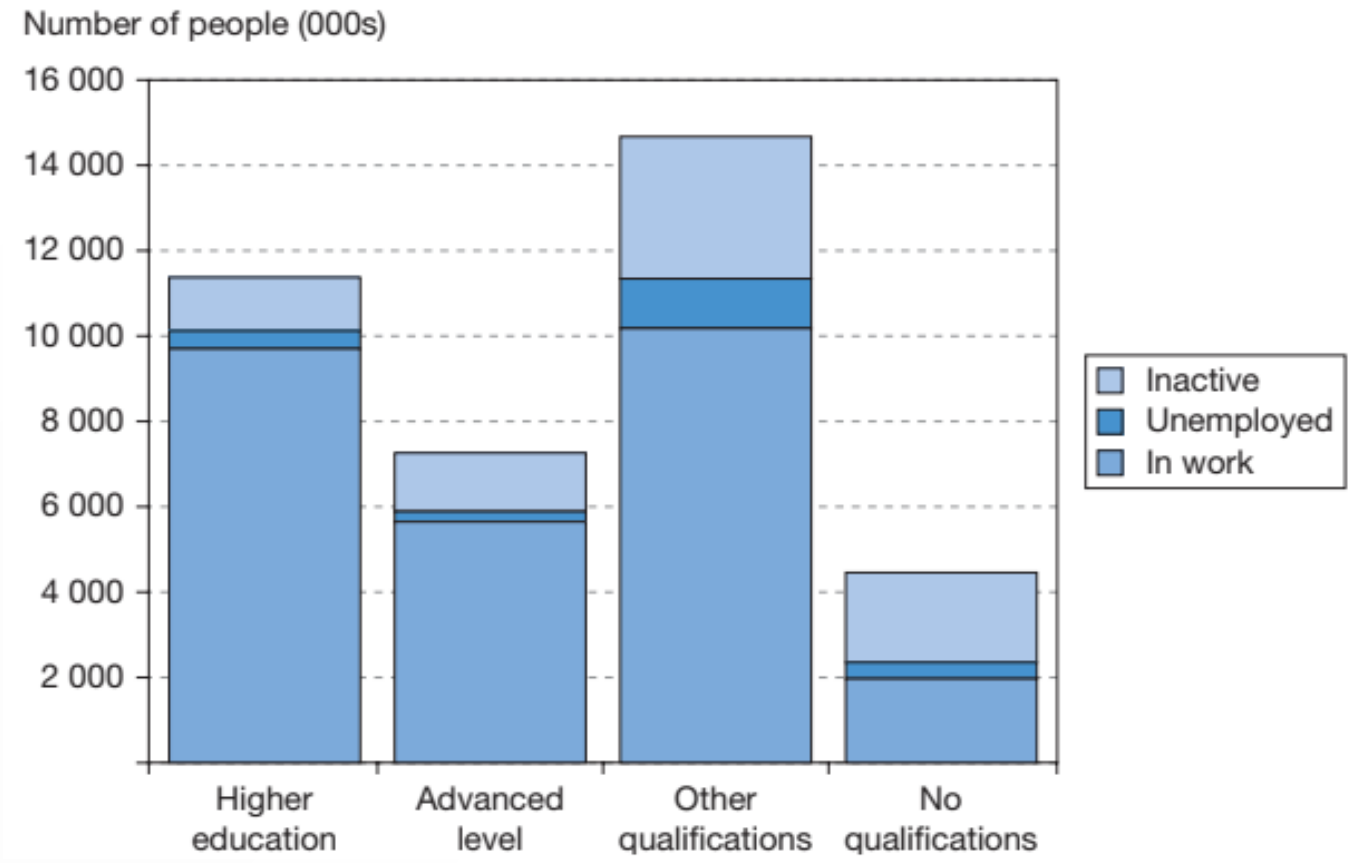
2.2 Summarising data using graphical techniques



2.2 Summarising data using graphical techniques

- We can also now begin to see whether more education increases your chance of having a job.
- For example, compare the height of the ‘in work’ bar to the ‘inactive’ bar.
- we have to make these judgements about the relative heights of different bars simply by eye, so it is easy to make a mistake. It would be better if we could draw charts that clearly highlight the differences.
- Figure below shows an alternative method of presentation: the stacked bar chart.
- In this case, the bars (for each education category) are stacked one on top of another instead of being placed side by side.
- This is perhaps slightly better, and the different overall size of each category is clearly brought out. However, we still have to make tricky visual judgements about proportions.
- As you may be starting to realise, we can present the same data in different ways depending upon our purpose. Here, we are going through different types of graph in turn and seeing what each can tell us. In practice, one would more likely identify the purpose first and then choose the type of graph most suited to it.

2.2 Summarising data using graphical techniques



2.2 Summarising data using graphical techniques

- A clearer picture emerges if the data are transformed into (column) percentages, i.e. the columns are expressed as percentages of the column totals (e.g. the proportion of graduates in work, rather than the number).
- This makes it easier to directly compare the different educational categories and to see whether graduates are more or less likely to be employed than others.

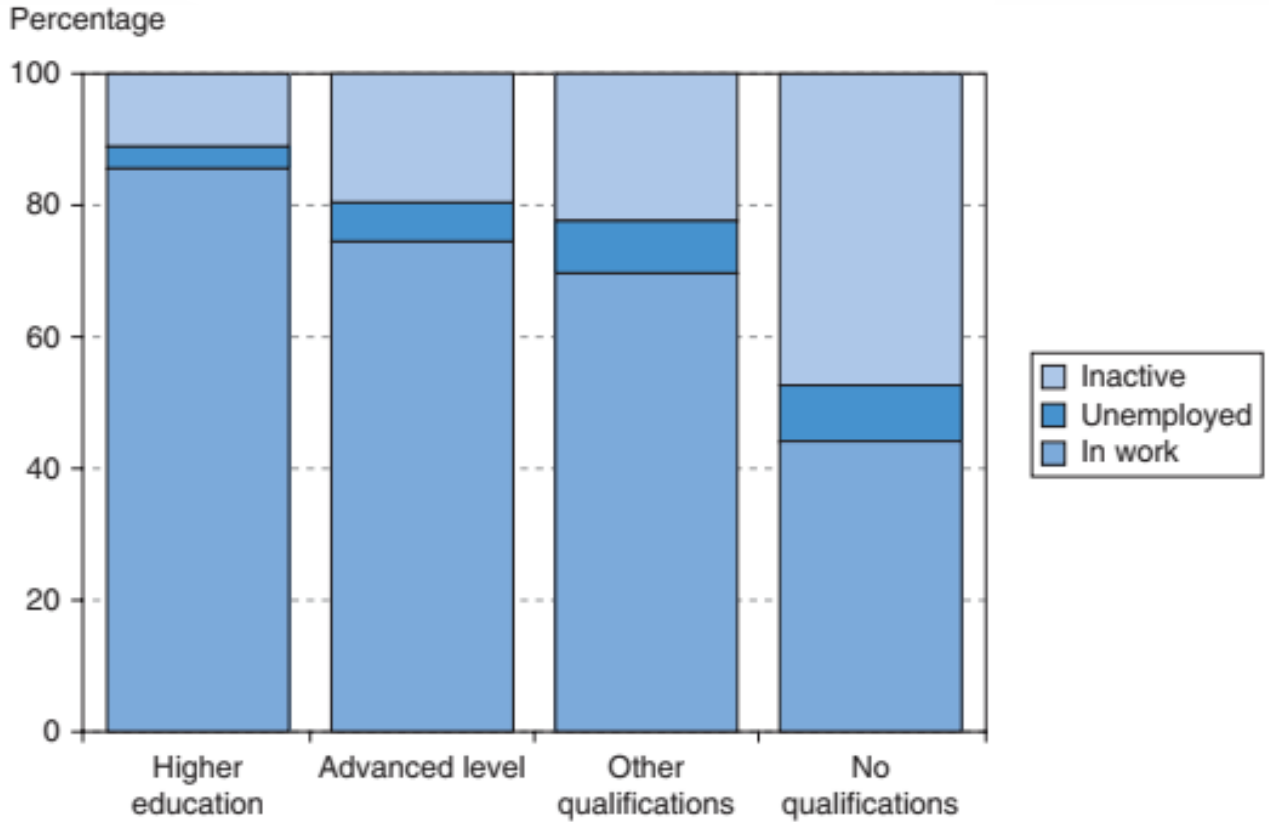
Table 1.2 Economic status and educational qualifications (column percentages)

	Higher education	A levels	Other qualification	No qualification	All
In work	85	75	70	44	72
Unemployed	3	6	8	9	6
Inactive	11	20	22	47	21
Totals	99	101	100	100	99

Note: The column percentages are obtained by dividing each frequency by the column total. For example, 85% is 9713 divided by 11 362; 75% is 5479 divided by 7352, etc. Some columns do not sum to 100% due to rounding.

2.2 Summarising data using graphical techniques

Having done this, it is easier to make a direct comparison of the different education categories (columns). This is shown in the next Figure 1.4 (based on the data in Table 1.2), where all the bars are of the same height (representing 100%) and the components of each bar now show the proportions of people in each educational category either in work, unemployed or inactive.



2.2 Summarising data using graphical techniques

- It is now clear how economic status differs according to education and the result is quite dramatic. In particular:
  - The proportion of people unemployed or inactive increases rapidly with lower educational attainment.
  - The biggest difference is between the no qualifications category and the other three, which have relatively smaller differences between them. In particular, A levels and other qualifications show a similar pattern.
- Thus we have looked at the data in different ways, drawing different charts and seeing what they can tell us.
- You need to consider which type of chart is most suitable for the data you have and the questions you want to ask. *There is no one graph which is ideal for all circumstances.*
- Can we safely conclude therefore that the probability of your being unemployed is significantly reduced by education? Could we go further and argue that the route to lower unemployment generally is via investment in education? The answer may be ‘yes’ to both questions, but we have not proved it.

2.2 Summarising data using graphical techniques

- Two important considerations are as follows:
  - Innate ability has been ignored. Those with higher ability are more likely to be employed and are more likely to receive more education. Ideally we would like to compare individuals of similar ability but with different amounts of education.
  - Even if additional education does reduce a person's probability of becoming unemployed, this may be at the expense of someone else, who loses their job to the more educated individual. In other words, additional education does not reduce total unemployment but only shifts it around amongst the labour force. Of course, it is still rational for individuals to invest in education if they do not take account of this externality.

2.2 Summarising data using graphical techniques

- **Try Something Yourself :** Producing charts using Microsoft Excel

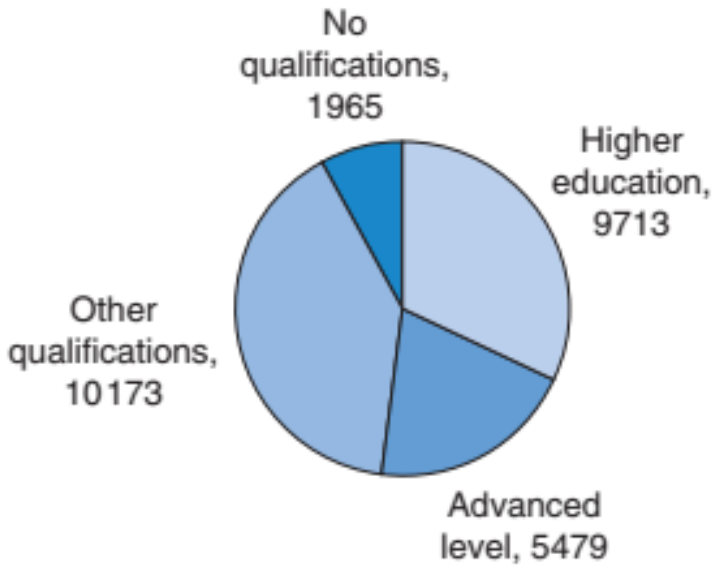
You can draw charts by hand on graph paper, and this is still a very useful way of really learning about graphs. Nowadays, however, most charts are produced by computer software, notably Excel. Most of the charts in this text were produced using Excel's charting facility. You should aim for a similar, uncluttered look. Some tips you might find useful are:

- Make the grid lines dashed in a light grey colour (they are not actually part of the chart, and hence should be discrete) or eliminate them altogether.
- Get rid of any background fill (grey by default; alter to 'No fill'). It will look much better when printed.
- On the x-axis, make the labels horizontal or vertical, not slanted – it is difficult to see which point they refer to.
- On the y-axis, make the axis title horizontal and place it at the top of the axis. It is much easier for the reader to see.
- Colour charts look great on-screen but unclear if printed in black and white. Change the style of the lines or markers (e.g. make some of them dashed) to distinguish them on paper.
- Both axes start at zero by default. If all your observations are large numbers, then this may result in the data points being crowded into one corner of the graph. Alter the scale on the axes to fix this – set the minimum value on the axis to be slightly less than the minimum observation. Note, however, that this distorts the relative heights of the bars and could mislead. Use with caution.

2.2 Summarising data using graphical techniques

- ***The pie chart***
- Another common way of presenting information graphically is the pie chart, which is a good way to describe how a variable is distributed between different categories.
- For example, from Table 1.1 we have the distribution of educational qualifications for those in work (the first row of the table).
- This can alternatively be shown as a pie chart, as in next Figure.
- The area (and angle) of each slice is proportional to the respective frequency, and the pie chart is an alternative means of presentation to the bar chart shown in Figure.
- The numbers falling into each education category have been added around the chart, but this is not essential. For presentational purposes, it is best not to have too many slices in the chart: beyond about six the chart tends to look crowded. It might be worth amalgamating less important categories to make such a chart look clearer.

2.2 Summarising data using graphical techniques



Note: If you have to draw a pie chart by hand, the angle of each slice can be calculated as follows:

$$angle = \frac{frequency}{total\,frequency} \times 360.$$

The angle of the first slice, for example, is

$$\frac{9713}{27330} \times 360 = 127.9°.$$

2.2 Summarising data using graphical techniques

Exercise 1

The following table shows the total numbers (in millions) of tourists visiting each country and the numbers of English tourists visiting each country:

	France	Germany	Italy	Spain
All tourists	12.4	3.2	7.5	9.8
English tourists	2.7	0.2	1.0	3.6

Adapted from data from the Office for National Statistics licensed under the Open Government Licence v.3.0.
Source: Office for National Statistics.

a) Draw a bar chart showing the total numbers visiting each country.

(b) Draw a stacked bar chart which shows English and non-English tourists making up the total visitors to each country.

(c) Draw a pie chart showing the distribution of all tourists between the four destination countries. Do the same for English tourists and compare results.

2.3 Looking at cross-section data:

- ***Frequency tables and charts***
- We now move on to examine data in a different form. The data on employment and education consisted simply of frequencies, where a characteristic (such as higher education) was either present or absent for a particular individual.
- We now look at the distribution of wealth, a variable which can be measured on a ratio scale so that a different value is associated with each individual.
- For example, one person might have £1000 of wealth, and another might have £1 million. Different presentational techniques will be used to analyse this type of data.
- We use these techniques to investigate questions such as *how much wealth does the average person have and whether wealth is evenly distributed or not.*
- The data are given in Table 1.3 shows the distribution of wealth in the United Kingdom for the year 2005 (the latest available at the time of writing),

2.3 Looking at cross-section data:

Table 1.3 The distribution of wealth, United Kingdom, 2005

Class interval (£)	Numbers (thousands)
0–9999	1 668
10 000–24 999	1 318
25 000–39 999	1 174
40 000–49 999	662
50 000–59 999	627
60 000–79 999	1 095
80 000–99 999	1 195
100 000–149 999	3 267
150 000–199 999	2 392
200 000–299 999	2 885
300 000–499 999	1 480
500 000–999 999	628
1 000 000–1 999 999	198
2 000 000 or more	88
Total	18 667

Note: It would be impossible to show the wealth of all 18 million individuals, so it has been summarised in this **frequency table**.

2.3 Looking at cross-section data:

- This is an example of a frequency table. Wealth is difficult to define and to measure; the data shown here refer to marketable wealth (i.e. items such as the right to a pension, which cannot be sold, are excluded) and are estimates for the population (of adults) as a whole based on taxation data.
- Wealth is divided into 14 *class intervals*: £0 up to (but not including) £10 000; £10 000 up to £24 999, etc., and the number (or *frequency*) of individuals within each class interval is shown.
- Note that the widths of the intervals (the *class widths*) vary up the wealth scale: the first is £10 000, the second
- £15000 (= 25 000 − 10 000), the third £15000 also and so on. This will prove an important factor when it comes to graphical presentation of the data.

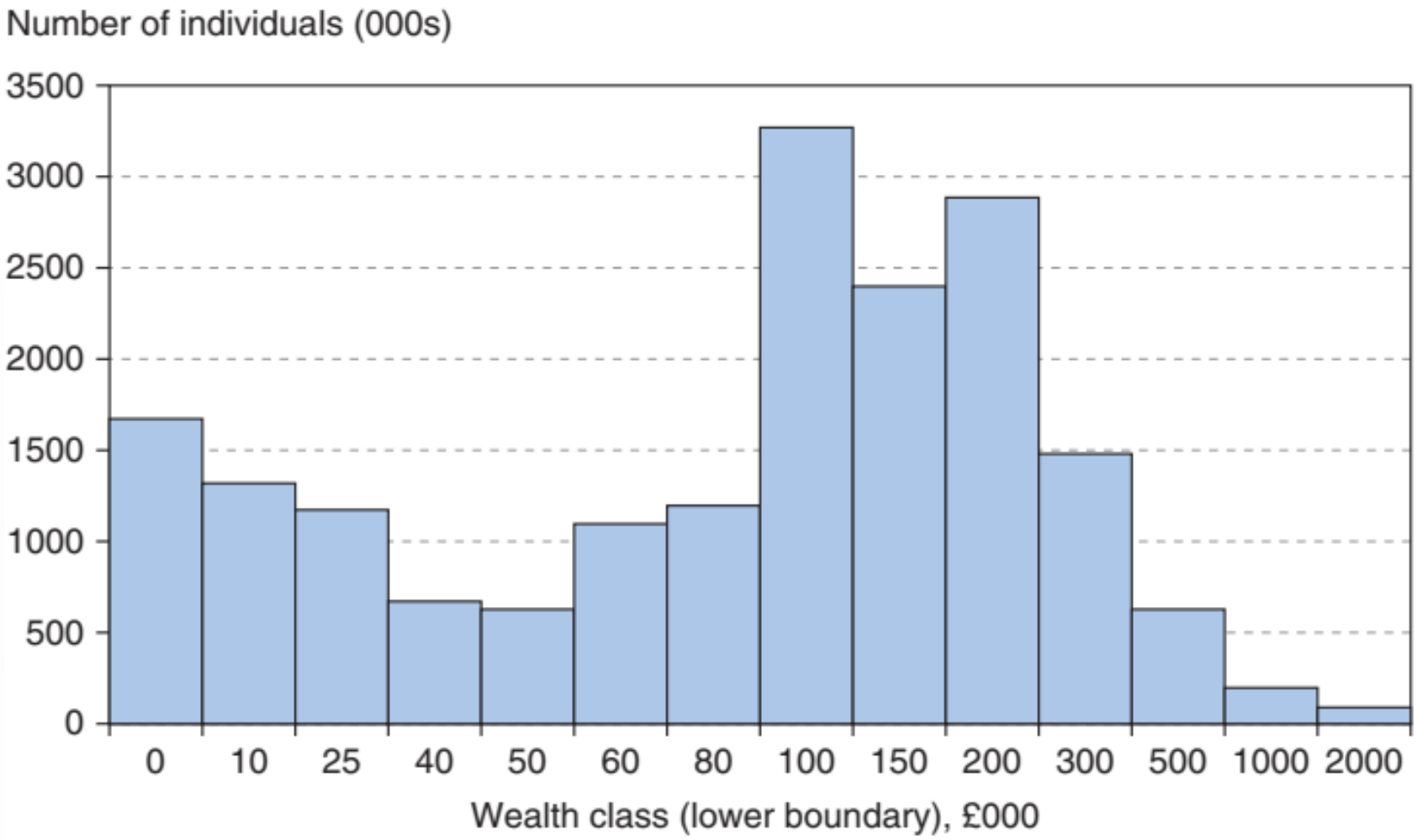
2.3 Looking at cross-section data:

- This table has been constructed from the original 18 667 000 observations on individuals' wealth, so it is already a summary of the original data (note that all the frequencies have been expressed in thousands in the table) and much of the original information is unavailable.
- The first decision to make if one had to draw up such a frequency table from the raw data is how many class intervals to have and how wide they should be.
- It simplifies matters if they are all of the same width, but in this case it is not feasible: if 10000 were chosen as the standard width for each class, there would be many intervals between 500000 and 1000000 (50 of them in fact), most of which would have a zero or very low frequency.
- If 100000 were the standard width, there would be only a few intervals and the first of them (0 - 100000) would contain 7739 observations (41% of all observations), so almost all the interesting detail would be lost.
- A compromise between these extremes has to be found.
- A useful rule of thumb is that the number of class intervals should equal the square root of the total frequency, subject to a maximum of about 12 intervals.

2.3 Looking at cross-section data:

- Thus, for example, a total of 25 observations should be allocated to 5 intervals; 100 observations should be grouped into 10 intervals and 18 667 should be grouped into about 12 (14 are used here). The class widths should be equal insofar as this is feasible but should increase when the frequencies become very small.
- To present these data graphically one could draw a bar chart, as in the case of education above, and this is presented in Figure below.
- Note that although the original data are on a ratio scale, we have transformed them so that we are now counting individuals in each category.
- Hence we can make use of the bar chart again, although note that the x-axis has categories differentiated by the value of wealth rather than some characteristic such as education.

2.3 Looking at cross-section data:



2.3 Looking at cross-section data:

- ***The histogram***
- A better method would make the shape of the distribution independent of how the class intervals are arranged. This can be done by drawing a histogram.
- A histogram is similar to a bar chart except that it corrects for differences in class widths. If all the class widths are identical, then there is no difference between a bar chart and a histogram. The calculations required to produce the histogram are shown in Table 1.4.
- The new column in the table shows the frequency density, which measures the frequency per unit of class width. Hence it allows a direct comparison of different class intervals, i.e. accounting for the difference in class widths. The frequency density is defined as follows:

$$frequency\ density\ = \frac{frequency}{classwidth}$$

2.3 Looking at cross-section data:

Table 1.4 Calculation of frequency densities

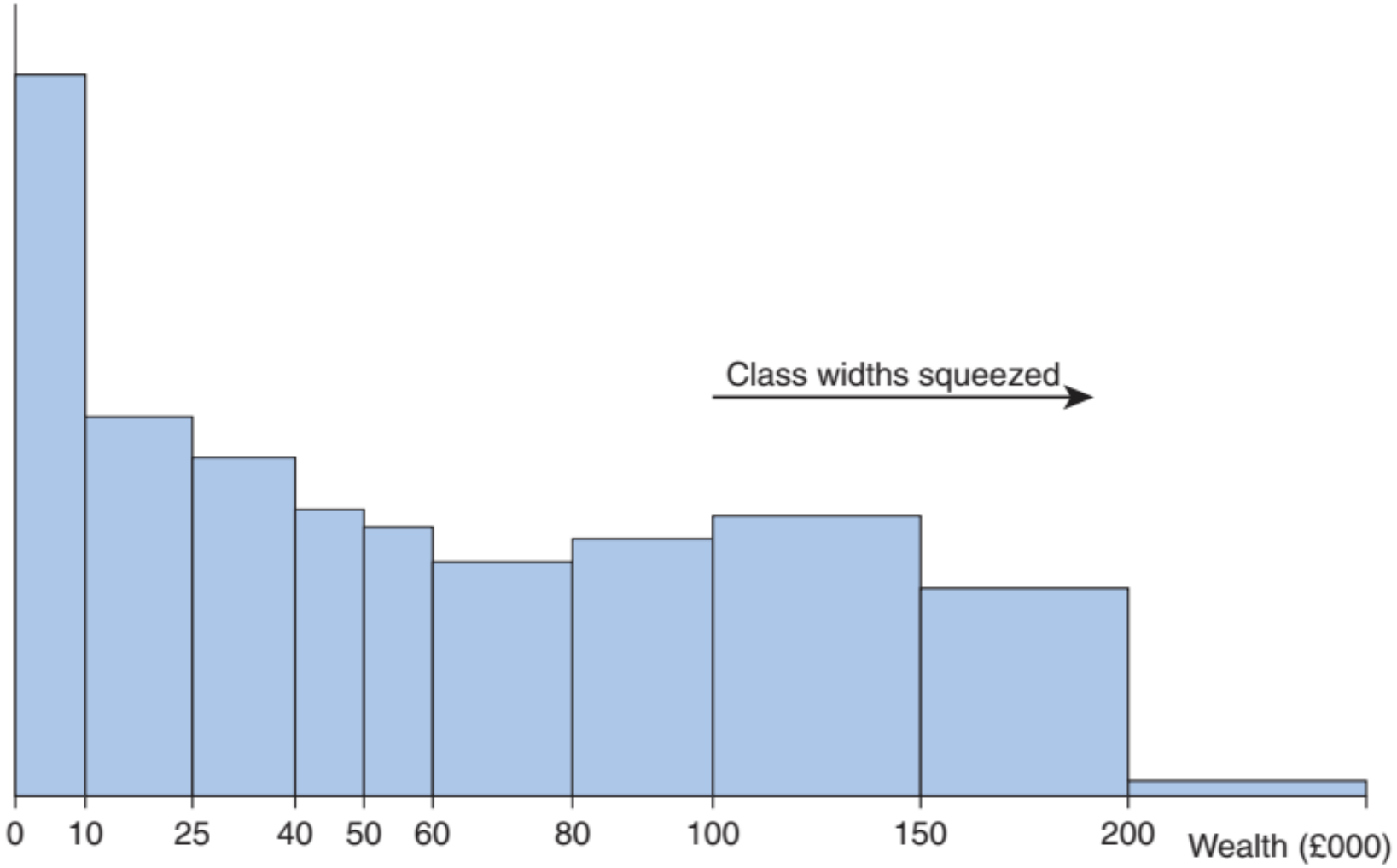
Range	Frequency	Class width	Frequency density
0–	1668	10 000	0.1668
10 000–	1318	15 000	0.0879
25 000–	1174	15 000	0.0783
40 000–	662	10 000	0.0662
50 000–	627	10 000	0.0627
60 000–	1095	20 000	0.0548
80 000–	1195	20 000	0.0598
100 000–	3267	50 000	0.0653
150 000–	2392	50 000	0.0478
200 000–	5279	3 800 000	0.0014

Note: As an alternative to the frequency density, one could calculate the frequency per 'standard' class width, with the standard width chosen to be 10 000 (the narrowest class). The values in column 4 would then be 1668; 879(= 1318 ÷ 1.5); 783, etc. This would lead to the same shape of histogram as using the frequency density.

2.3 Looking at cross-section data:

- Above £200 000, the class widths are very large and the frequencies small (too small to be visible on a histogram), so these classes have been combined.
- The width of the final interval is unknown, so it has to be estimated in order to calculate the frequency density. It is likely to be extremely wide since the wealthiest person may well have assets valued at several £m (or even £bn); the value we assume will affect the calculation of the frequency density and therefore of the shape of the histogram. Fortunately, it is in the tail of the distribution and only affects a small number of observations. Here we assume (arbitrarily) a width of £3.8m to be a 'reasonable' figure, giving an upper class boundary of £4m.
- The frequency density, not the frequency, is then plotted on the vertical axis against wealth on the horizontal axis to give the histogram. One further point needs to be made: for clarity, the scale on the horizontal wealth axis should be linear as far as possible, e.g. £50 000 should be twice as far from the origin as £25 000. However, it is difficult to fit all the values onto the horizontal axis without squeezing the graph excessively at lower levels of wealth, where most observations are located. Therefore, the classes above £100 000 have been squeezed, and the reader's attention is drawn to this. The result is shown in Figure below.

2.3 Looking at cross-section data:



Note: A **frequency polygon** would be the result if, instead of drawing blocks for the histogram, one drew lines connecting the centres of the top of each block. The diagram is better drawn with blocks, in general.

2.3 Looking at cross-section data:

The effect of taking frequency densities is to make the area of each block in the histogram represent the frequency, rather than the height, which now shows the density. This has the effect of giving an accurate picture of the shape of the distribution. Note that it is very different from the preceding graph.

Now that all this has been done, what does the histogram show?

● The histogram is heavily skewed to the right (i.e. the long tail is to the right).

● The modal class interval is £0 to £10 000 (i.e. has the greatest density: no other £10 000 interval has more individuals in it).

Looking at the graph, it appears that more than half of all people have wealth of less than £100 000. However, this is misleading as the graph is squeezed beyond £100 000. In fact, about 41% have wealth below this figure.

The figure shows quite a high degree of inequality in the wealth distribution. Whether this is acceptable or even desirable is a value judgement. It should be noted that part of the inequality is due to differences in age: younger people have not yet had enough time to acquire much wealth and therefore appear worse off, although

in lifetime terms this may not be the case. To get a better picture of the distribution of wealth would require some analysis of the acquisition of wealth over the life-cycle (or comparison of individuals of a similar age). In fact, correcting for age differences does not make a big difference to the pattern of wealth distribution.

2.3 Looking at cross-section data:

- ***Relative frequency and cumulative frequency distributions***
- An alternative way of illustrating the wealth distribution uses the relative and cumulative frequencies of the data. The relative frequencies show the proportion of observations that fall into each class interval, so, for example, 3.5% of individuals have wealth holdings between £40 000 and £50 000 (662 000 out of 18 677 000 individuals).
- Relative frequencies are shown in the third column of Table 1.5, calculated using the following formula:

$$Relative\ Frequency = \frac{frequency}{sume\ of\ frequencies}$$

2.3 Looking at cross-section data:

Table 1.5 Calculation of relative and cumulative frequencies

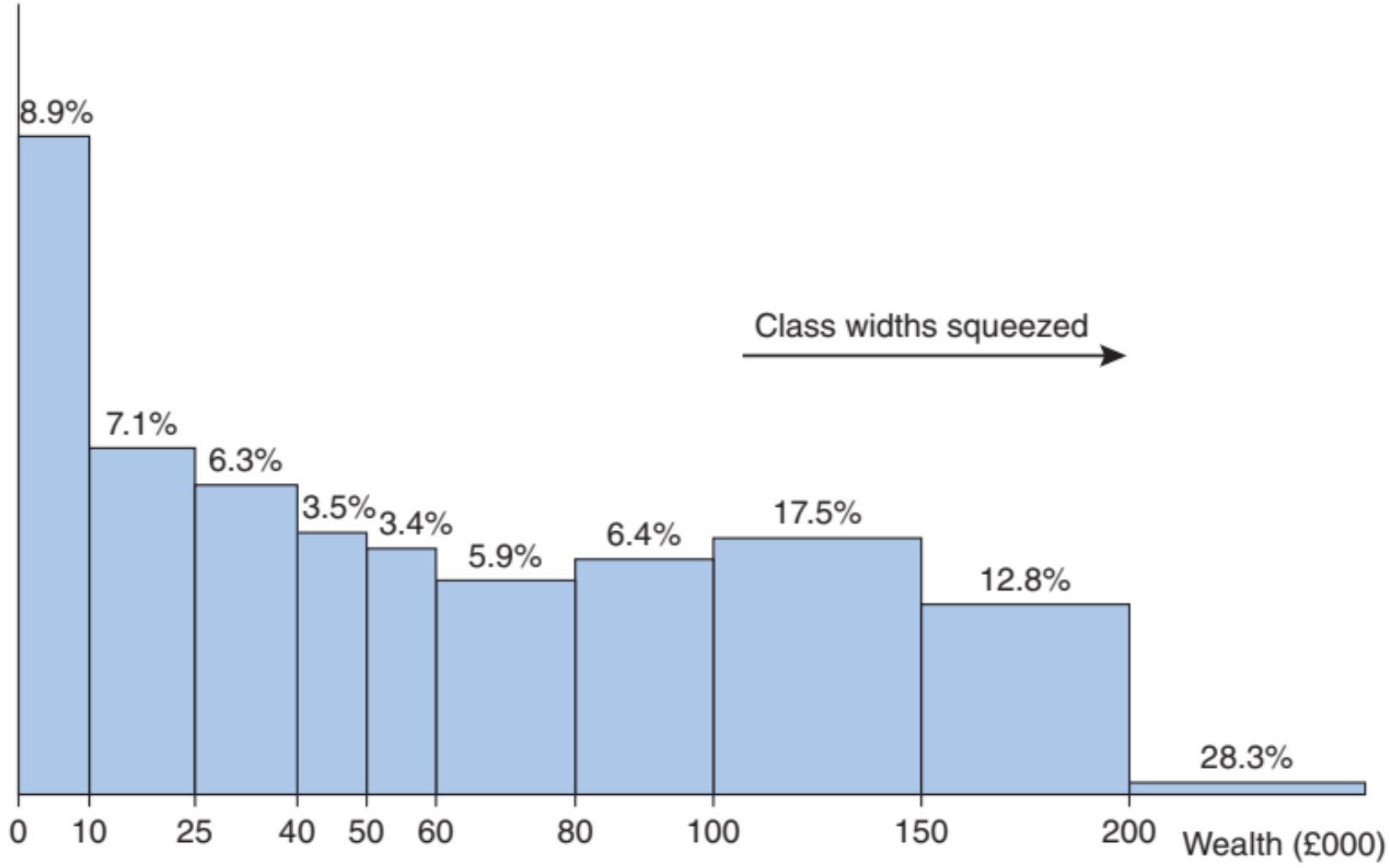
Range	Frequency, *f*	Relative frequency (%)	Cumulative frequency, *F*
0–	1 668	8.9	1 668
10 000–	1 318	7.1	2 986
25 000–	1 174	6.3	4 160
40 000–	662	3.5	4 822
50 000–	627	3.4	5 449
60 000–	1 095	5.9	6 544
80 000–	1 195	6.4	7 739
100 000–	3 267	17.5	11 006
150 000–	2 392	12.8	13 398
200 000–	2 885	15.4	16 283
300 000–	1 480	7.9	17 763
500 000–	628	3.4	18 391
1 000 000–	198	1.1	18 589
2 000 000–	88	0.5	18 677
Total	18 677	100.0	

Note: Relative frequencies are calculated in the same way as the column percentages in Table 1.2. Thus for example, 8.9% is 1668 divided by 18 667. Cumulative frequencies are obtained by cumulating, or successively adding, the frequencies. For example, 2986 is 1668 + 1318, 4160 is 2986 + 1174, etc.

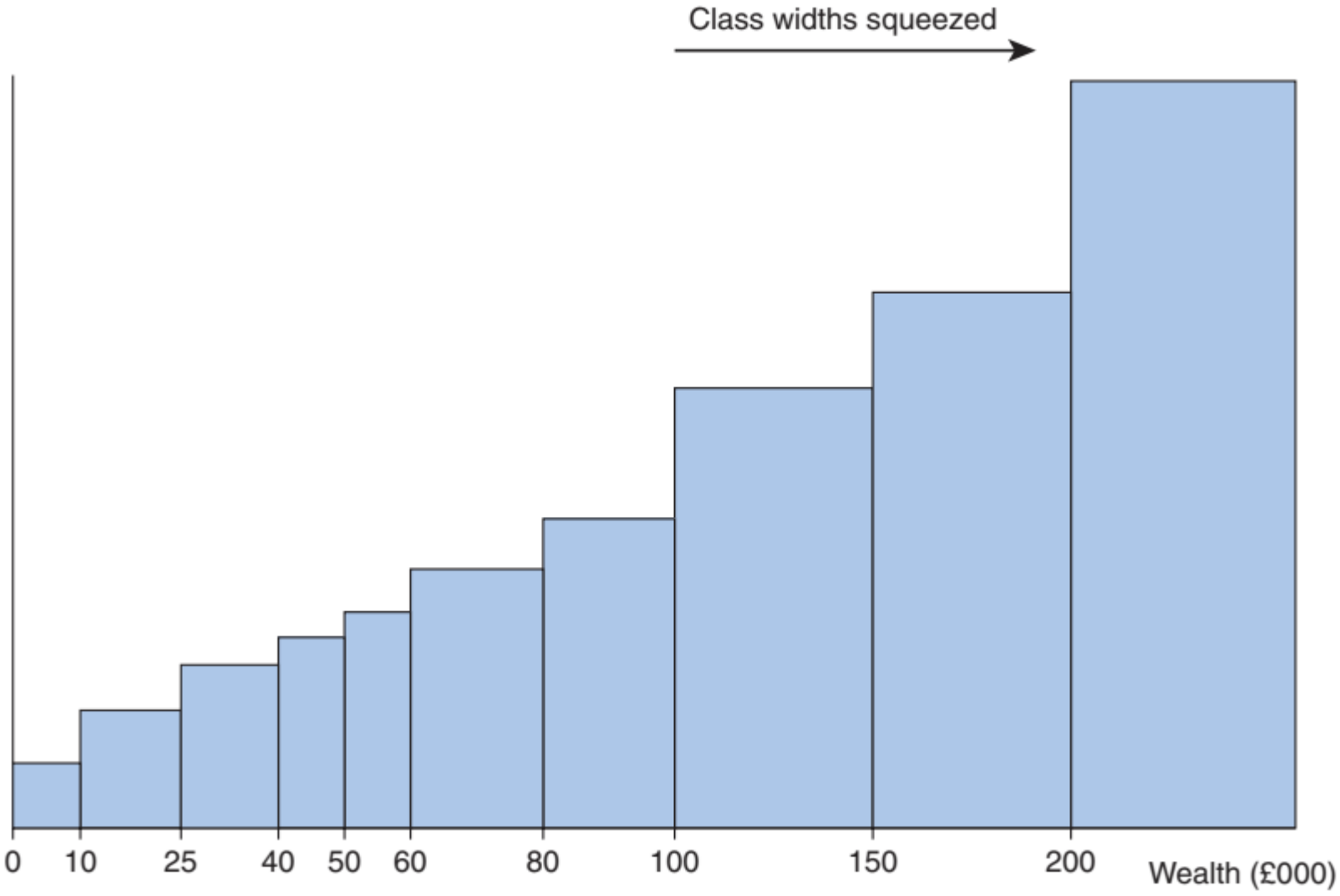
2.3 Looking at cross-section data:

- The sum of the relative frequencies has to be 100%, and this acts as a check on the calculations.
- The cumulative frequencies, shown in the fourth column, are obtained by cumulating (successively adding) the frequencies. The cumulative frequencies show the total number of individuals with wealth up to a given amount; for example, about 7.7 million people have less than £100 000 of wealth.
- Both relative and cumulative frequency distributions can be drawn, in a similar way to the histogram. In fact, the relative frequency distribution has exactly the same shape as the frequency distribution. This is shown in Figure below. This time we have written the relative frequencies above the appropriate column, although this is not essential.
- The cumulative frequency distribution is shown in the next Figure, where the blocks increase in height as wealth increases. The simplest way to draw this is to cumulate the frequency densities (shown in the final column of Table 1.4) and to use these values as the y-axis coordinates.

2.3 Looking at cross-section data:



2.3 Looking at cross-section data:



2.3 Looking at cross-section data:

- **Exercise 2** Given the following data:

Range	Frequency
0 – 10	20
11 – 30	40
31 – 60	30
61 – 100	20

- (a) Draw both a bar chart and a histogram of the data and compare them.
- (b) Calculate cumulative frequencies and draw a cumulative frequency diagram.

2.3 Looking at cross-section data:

Example: The data on the variable X and its frequencies f are shown in the following table, with the calculations required:

x	Frequency	Relative Frequency	Cumulative Frequency
10	6	$\frac{6}{35} = 0.17$	
11	8		$8 + 6 = 14$
12	15		
13	5		
14	1		
Total	**35**		

The resulting bar chart …. and cumulative frequency distribution are….

2.4 Summarising data using numerical techniques:

- Graphical methods are an excellent means of obtaining a quick overview of the data, but they are not particularly precise, nor do they lend themselves to further analysis.
- For this, we must turn to numerical measures such as the average. There are a number of different ways in which we may describe a distribution such as that for wealth. If we think of trying to describe the histogram, it is useful to have:
- A measure of location giving an idea of whether people own a lot of wealth or a little. An example is the average, which gives some idea of where the distribution is located along the x-axis. In fact, we will encounter three different measures of the 'average':

  (1) The mean (2) The median (3) The mode

- A measure of dispersion showing how wealth is dispersed around the average, whether it is concentrated close to the average or is generally far away from it. *An example here is the standard deviation*.
- A measure of skewness showing how symmetric the distribution is, i.e. whether the left half of the distribution is a mirror image of the right half. This is obviously not the case for the wealth distribution.

2.4 Summarising data using numerical techniques:

- ***Measures of location: the mean***
- The *arithmetic mean*, commonly called the average, is the most familiar measure of location and is obtained simply by adding all the wealth observations and dividing by the number of observations.
- If we denote the wealth of the i^{th} household by x_i (so that the index i runs from 1 to N, where N is the number of observations; as an example, x_3 would be the wealth of the third household), then the mean is given by the following formula:

$$\mu = \frac{\sum_{i=1}^{i=N} x_i}{N} \qquad (1)$$

- ***Example:*** for the values 17, 25, 28, 20, 35 , find the mean.

2.4 Summarising data using numerical techniques:

- Formula 1 can only be used when all the individual x values are known.
- The frequency table for wealth does not show all 18 million observations, however, but only the range of values for each class interval and the associated frequency.
- In the case of such **grouped data** the following equivalent formula may be used:

$$\mu = \frac{\sum_{i=1}^{i=N} f_i x_i}{\sum_{i=1}^{i=N} f_i} \qquad (2)$$

- x denotes the mid-point of each class interval, since the individual x values are unknown. The mid-point is used as the representative x value for each class. In the first class interval, for example, we do not know precisely where each of the 1668 observations lies. Hence we assume they all lie at the mid-point, £5000. This will cause a slight inaccuracy – because the distribution is so skewed, there are likely more households below the mid-point than above it in every class interval except, perhaps, the first. We ignore this problem here, and it is less of a problem for most distributions which are less skewed than this one.
- The summation runs from 1 to C, the number of class intervals, or mid-point x values. f times x gives the total wealth in each class interval. If we sum over the 14 class intervals, we get the total wealth of all individuals.
- $\Sigma f_i = N$ gives the total number of observations, the sum of the individual frequencies.

2.4 Summarising data using numerical techniques:

Table 1.6 The calculation of average wealth

Range	x	f	fx
0–	5.0	1668	8340
10000–	17.5	1318	23065
25000–	32.5	1174	38155
40000–	45.0	662	29790
50000–	55.0	627	34485
60000–	70.0	1095	76650
80000–	90.0	1195	107550
100000–	125.0	3267	408375
150000–	175.0	2392	418600
200000–	250.0	2885	721250
300000–	400.0	1480	592000
500000–	750.0	628	471000
1000000–	1500.0	198	297000
2000000–	3000.0	88	264000
Total		18677	3490260

Note: The fx column gives the product of the values in the f and x columns (so, for example, $5.0 \times 1668 = 8340$, which is the total wealth held by those in the first class interval). The sum of the fx values gives total wealth.

2.4 Summarising data using numerical techniques:

- ***Example:*** Suppose we have 10 families with a single television in their homes, 12 families with two televisions each and three families with three. Setting this out formally and calculate the average number of televisions per family

x	f	$f \times x$
1	10	$1 * 10 = 10$
2		
3		
Totals		

- $\mu = -$

2.4 Summarising data using numerical techniques:

- ***The mean as the expected value***
- We also refer to the mean as the expected value of x and write:

$$E(x) = \mu$$

- The mean is the expected value in the sense that if we selected a household at random from the population, we would ‘expect’ its wealth to be £186 875.
- It is important to note that this is a statistical expectation, rather than the everyday use of the term. Most of the random individuals we encounter have wealth substantially below this value. Most people might therefore ‘expect’ a lower value because that is their everyday experience; but statisticians are different; they refer to the mean as the expected value.

2.4 Summarising data using numerical techniques:

- ***The sample mean and the population mean***
- Very often we have only a sample of data (as in worked example), and it is important to distinguish this case from the one where we have all the possible observations. For this reason, the sample mean is given by:

$$\bar{x} = \frac{\Sigma x}{n} \quad or \quad \bar{x} = \frac{\Sigma f x}{\Sigma f} \text{ for grouped data} \qquad (3)$$

- Note the distinctions between μ (the population mean) and x (the sample mean), and between N (the size of the population) and n (the sample size).
- Otherwise, the calculations are identical. It is a convention to use Greek letters, such as μ, to refer to the population and Roman letters, such as x, to refer to a sample.

2.4 Summarising data using numerical techniques:

- ***The weighted average***
- Sometimes observations have to be given different weightings in calculating the average, as in the following example.
- Consider the problem of calculating the average spending per pupil by an education authority. Some figures for spending on primary (ages 5–11), secondary (11–16) and post-16 pupils are given in Table 1.7.
- Clearly, significantly more is spent on secondary and post-16 pupils (a general pattern throughout England and most other countries) and the overall average should lie somewhere between 1750 and 3820. However, taking a simple average of these three values would give the wrong answer, because there are different numbers of children in the three age ranges.

Table 1.7 Cost per pupil in different types of school (£ p.a.)

	Primary	Secondary	Post-16
Unit cost	1750	3100	3820

2.4 Summarising data using numerical techniques:

- The numbers and proportions of children in each age group are given in Table 1.8.

Table 1.8 Numbers and proportions of pupils in each age range

	Primary	Secondary	Post-16	Total
Numbers	8000	7000	3000	18 000
Proportion	44.4%	38.9%	16.7%	

- Since there are relatively more primary schoolchildren than secondary, and relatively fewer post-16 pupils, the primary unit cost should be given greatest weight in the averaging process and the post-16 unit cost the least. The **weighted average** is obtained by multiplying each unit cost figure by the proportion of children in each category and summing. The weighted average is therefore

$$0.444 \times 1750 + 0.389 \times 3100 + 0.167 \times 3820 = 2620.8$$

- The weighted average gives an answer closer to the primary unit cost than does the simple average of the three figures (2890 in this case), which would be misleading. The formula for the weighted average is

$$\bar{x}_w = \sum w_i x_i$$

2.4 Summarising data using numerical techniques:

- ***The median***
- Returning to the study of wealth, the unrepresentative result for the mean suggests that we may prefer a measure of location which is not so strongly affected by outliers (extreme observations) and skewness.
- The **median** is a measure of location which is *more robust to such extreme values*; it may be defined by the following procedure. Imagine everyone in a line from poorest to wealthiest. Go to the individual located halfway along the line.
- Ask her what her wealth is. Her answer is the median. The median is clearly unaffected by extreme values, unlike the mean: if the wealth of the richest person were doubled (with no reduction in anyone else's wealth), there would be no effect upon the median. The calculation of the median is not so straightforward as for the mean, especially for grouped data.
- The following worked example first shows how to calculate the median for ungrouped data.
- ***Example:*** Calculate the median of the following values: 45, 12, 33, 80, 77.
- ***Example:*** Find the median of 12, 33, 45, 63, 77, 80.

2.4 Summarising data using numerical techniques:

- **For grouped data** there are two stages to the calculation: **first** we must first identify the class interval which contains the median person, **then** we must calculate where in the interval that person lies.
- (1) To find the appropriate class interval: since there are 17 636 000 observations, we need the wealth of the person who is 8 818 000 in rank order. The table of cumulative frequencies (see Table 1.5 above) is the most suitable for this. There are 8 106 000 individuals with wealth of less than £80 000 and 9 746 000 with wealth of less than £100 000. The middle person therefore falls into the £80 000–100 000 class. Furthermore, given that 8 818 000 falls roughly half way between 8 106 000 and 9 746 000 it follows that the median is close to the middle of the class interval. We now go on to make this statement more precise.
- (2) To find the position in the class interval, we can now use formula (4) below

2.4 Summarising data using numerical techniques:

$$median = x_l + (x_u - x_l)\left\{\frac{\frac{N+1}{2} - F}{f}\right\}$$

- x_L = the lower limit of the class interval containing the median. x_U = the upper limit of this class interval
- N = the number of observations (using $N + 1$ rather than N in the formula is only important when N is relatively small). F = the cumulative frequency of the class intervals up to (but not including) the one containing the median
- f = the frequency for the class interval containing the median.
- For the wealth distribution we have

$$median = 8000 + (100000 - 80000)\left\{\frac{\frac{17636000+1}{2} - 8106000}{1640000}\right\} = 90823$$

- This alternative measure of location gives a very different impression: it is less than two-thirds of the mean. Nevertheless, it is equally valid despite having a different meaning. It demonstrates that the person ‘in the middle’ has wealth of £90 829 and in this sense is typical of the UK population.

2.4 Summarising data using numerical techniques:

- ***The mode***
- The mode is defined as that level of wealth which occurs with the greatest frequency, in other words the value that occurs most often. It is most useful and easiest to calculate when one has all the data and there are relatively few distinct observations.
- This is the case in the simple example below.
- ***Example:*** Suppose we have the following data on sales of dresses by a shop, according to size

Size	Sales
8	7
10	25
12	36
14	11
16	11
18	3

2.4 Summarising data using numerical techniques:

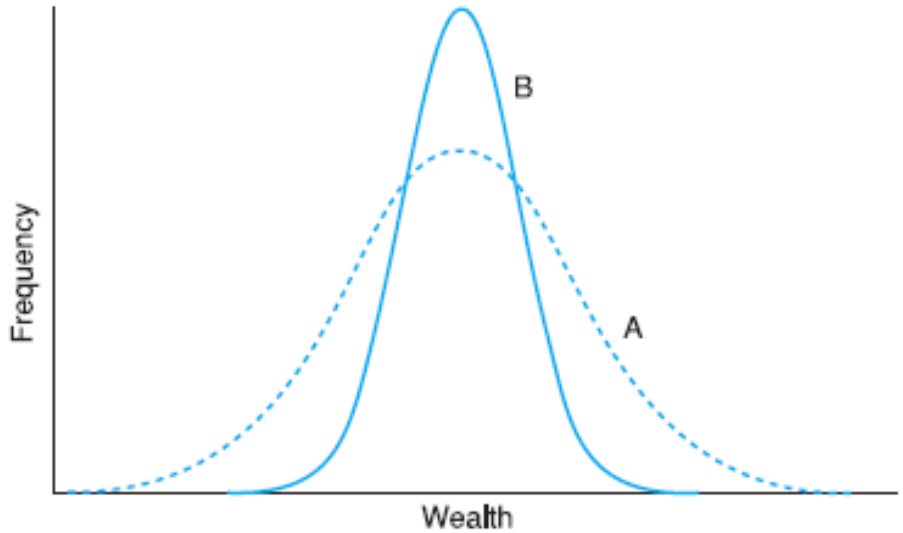
- The modal size is 12. There are more women buying dresses of this size than any other. This may be the most useful form of average as far as the shop is concerned. Although it needs to stock a range of sizes, it knows it needs to order more dresses in size 12 than in any other size. The mean would not be so helpful in this case (it is X = 11.7) as it is not an actual dress size.
- **In the case of grouped data** matters are more complicated. It is the modal class interval which is required, once the intervals have been corrected for width (otherwise a wider class interval is unfairly compared with a narrower one).
- For this, we can again make use of the frequency densities.
- From Table 1.4 it can be seen that it is the first interval, from £0 to £10 000, which has the highest frequency density.
- It is 'typical' of the distribution because it is the one which occurs most often (using the frequency densities, not frequencies).
- The wealth distribution is most concentrated at this level and more people are like this in terms of wealth than anything else. Once again it is notable how different it is from both the median and the mean.

2.4 Summarising data using numerical techniques:

- **Exercise 3** Given the data in exercise 2:
- a) calculate the mean, median and mode of the data.
- (b) Mark these values on the histogram you drew for Exercise 1.2.

2.4 Summarising data using numerical techniques:

- ***Measures of dispersion***
- Two different distributions (e.g. wealth in two different countries) might have the same mean yet look very different, as shown in the next Figure (the distributions have been drawn using smooth curves rather than bars to improve clarity).
- In one country everyone might have a similar level of wealth (curve B). In another, although the average is the same there might be extremes of great wealth and poverty (curve A).



Note: Distribution A has a greater degree of dispersion than B, where everyone has a similar level of wealth.

2.4 Summarising data using numerical techniques:

- A measure of dispersion is a number which allows us to distinguish between these two situations.
- The simplest measure of dispersion is **the range**, which is the difference between the smallest and largest observations. It is impossible to calculate accurately from the table of wealth holdings since the largest observation is not available. In any case, it is not a very useful figure since it relies on two extreme values and ignores the rest of the distribution. In simpler cases it might be more informative.
- For example, in an exam the marks may range from a low of 28% to a high of 74%. In this case the range is 74 − 28 = 46 and this tells us something useful.
- An improvement is the **inter-quartile range (IQR),** which is the difference between the first and third quartiles. It therefore defines the limits of wealth of the middle half of the distribution and ignores the very extremes of the distribution.
- To calculate the first quartile (which we label Q1) we have to go one-quarter of the way along the line of wealth holders (ranked from poorest to wealthiest) and ask the person in that position what their wealth is.

2.4 Summarising data using numerical techniques:

- Their answer is the first quartile. The calculation is as follows:
  - one-quarter of 17 636 is 4409; the person ranked 4409 is in the £25 000–40 000 class;
  - adapting formula (4)

$$Q_1 = 25000 + (40000 - 25000)\left\{\frac{4409 - 4271}{1375}\right\} = 26505.5$$

- The third quartile is calculated in similar fashion:
  - three-quarters of 17 636 is 13 227; the person ranked 13 227 is in the £150 000–200 000 class;
  - again using formula (4)
- $Q_3 = 150000 + (200000 - 150000)\left\{\frac{13227-11897}{2215}\right\} = 180022.6$
- and therefore the inter-quartile range is Q3 − Q1 = 180 022 − 26 505 = 153 517.
- This might be reasonably rounded to £150 000 given the approximations in our calculation, and is a much more memorable figure.

2.4 Summarising data using numerical techniques:

- This gives one summary measure of the dispersion of the distribution: the higher the value the more spread-out is the distribution. Two different wealth distributions might be compared according to their inter-quartile ranges therefore, with the country having the larger figure exhibiting greater inequality.
- Note that the figures would have to be expressed in a common unit of currency for this comparison to be valid.
- ***Example:*** Suppose 110 children take a test, with the following results:

Mark, X	Frequency, f	Cumulative frequency, F
13	5	5
14	13	18
15	29	47
16	33	80
17	17	97
18	8	105
19	4	109
20	1	110
Total	110	

2.4 Summarising data using numerical techniques:

- The range is simply 20 - 13 = 7. The inter-quartile range requires calculation of the quartiles. Q 1 is given by the value of the 27.5th observation (= 110>4), which is 15. Q 3 is the value of the 82.5th observation (= 110 * 0.75) which is 17. The IQR is therefore 17 - 15 = 2 marks.
- Notice that a slight change in the data (three more students getting 16 rather than 17 marks) would alter the IQR to 1 mark (16 - 15).
- The result should therefore be treated with some caution. This is a common problem when there are few distinct values of the variable (eight in this example). It is often worth considering whether a few small changes to the data could alter a calculation considerably. In such a case, the original result might not be very robust.

2.4 Summarising data using numerical techniques:

- ***The variance***

- A more useful measure of dispersion is the variance, which makes use of all of the information available, rather than just trimming the extremes of the distribution.

- The variance is denoted by the symbol σ^2. σ is the Greek lower-case letter sigma, so itis read 'sigma squared'. It has a completely different meaning from Σ (capital sigma) used before. Its formula is:

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

- In this formula, x - m measures the distance from each observation to the mean.

- Squaring these makes all the deviations positive, whether above or below the mean. We then take the average of all the squared deviations from the mean.

- A more dispersed distribution will tend to have larger deviations from the mean and hence a larger variance.

- In comparing two distributions with similar means, therefore, we could examine their variances to see which of the two has the greater degree of dispersion.

- With **grouped data** the formula becomes: $\sigma^2 = \frac{\sum f(x-\mu)^2}{\sum f}$

2.4 Summarising data using numerical techniques:

- ***Example:*** calculate the variance of wealth is shown in Table 1.9,

Table 1.9 The calculation of the variance of wealth

Range	Mid-point x (£000)	Frequency, f	Deviation $(x - \mu)$	$(x - \mu)^2$	$f(x - \mu)^2$
0–	5.0	1668	−181.9	33078.4	55174821.9
10000–	17.5	1318	−169.4	28687.8	37810535.3
25000–	32.5	1174	−154.4	23831.6	27978261.2
40000–	45.0	662	−141.9	20128.4	13325033.3
50000–	55.0	627	−131.9	17391.0	10904128.1
60000–	70.0	1095	−116.9	13659.7	14957383.4
80000–	90.0	1195	−96.9	9384.7	11214740.7
100000–	125.0	3267	−61.9	3828.5	12507665.8
150000–	175.0	2392	−11.9	141.0	337296.1
200000–	250.0	2885	63.1	3984.8	11496134.1
300000–	400.0	1480	213.1	45422.4	67225100.9
500000–	750.0	628	563.1	317110.0	199145098.5
1000000–	1500.0	198	1313.1	1724297.9	341410980.4
2000000–	3000.0	88	2813.1	7913673.6	696403275.4
Totals		18677			1499890455.1

2.4 Summarising data using numerical techniques:

- Note that the variance is $\sigma^2 = 80306.8$
- This calculated value is before translating back into the original units of measurement, as was done for the mean by multiplying by 1000.
- In the case of the variance, however, we must multiply by 1000000, which is the square of 1000.
- The variance of the original data is therefore 80306800 000.
- Multiplying by the square of 1000 is a consequence of using squared deviations in the variance formula.
- One thus needs to be a little careful about the units of measurement. If the mean is reported at 186.875, then it is appropriate to report the variance as 80 306.8. If the mean is reported as 186 875, then the variance should be reported as 80 306 800 000. Note that it is only the presentation which changes; the underlying facts are the same.

2.4 Summarising data using numerical techniques:

- ***The standard deviation***

- In what units is the variance measured? Since we have used a squaring procedure in the calculation we end up with something like 'squared' £s, which is not very convenient, nor does it make much sense. Because of this, it is useful to define the standard deviation as the square root of the variance, which is therefore back in £s. The standard deviation is therefore given by:

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

- for grouped data: $\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{\sum f}}$

- The standard deviation of wealth is therefore $\sqrt{80306.8} = 283.385$. This is in £000, so the standard deviation is actually £283385 (note that this is the square root of 80 306 800 000, as it should be).

- On its own the standard deviation (and the variance) is not easy to interpret since it is not something we have an intuitive feel for, unlike the mean. It is more useful when used in a comparative setting. This will be illustrated later on.

2.4 Summarising data using numerical techniques:

- ***The variance and standard deviation of a sample***

- In what units is the variance measured? Since we have used a squaring procedure in the calculation we end up with something like 'squared' £s, which is not very convenient, nor does it make much sense. Because of this, it is useful to define the standard deviation as the square root of the variance, which is therefore back in £s. The standard deviation is therefore given by:

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

- for grouped data: $\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{\sum f}}$

- The standard deviation of wealth is therefore $\sqrt{80306.8} = 283.385$. This is in £000, so the standard deviation is actually £283385 (note that this is the square root of 80 306 800 000, as it should be).

- On its own the standard deviation (and the variance) is not easy to interpret since it is not something we have an intuitive feel for, unlike the mean. It is more useful when used in a comparative setting. This will be illustrated later on.

2.4 Summarising data using numerical techniques:

Example: calculate the variance and standard deviation for the students marks shown in Table 1.10,

x	f	fx	$x-\mu$	$(x-\mu)^2$	$f(x-\mu)^2$
13	5	$13*5=65$	$13-15.81=-2.81$	$(-2.81)^2=7.89$	$5*7.89=39.45$
14	13				
15	29				
16	33				
17	17				
18	8				
19	4				
20	1				
Totals	110	1739			222.99

2.4 Summarising data using numerical techniques:

- ***Alternative formulae for calculating the variance and standard deviation***
- The following formulae give the same answers as equations before but are simpler to calculate, either by hand or using a spreadsheet. For the population variance one can use

$$\sigma^2 = \frac{\sum x^2}{N} - \mu^2$$

- for grouped data: $\sigma^2 = \frac{\sum f x^2}{\sum f} - \mu^2$
- ***The variance and standard deviation of a sample***

2.4 Summarising data using numerical techniques:

- ***The variance and standard deviation of a sample***
- As with the mean, a different symbol is used to distinguish a variance calculated from the population and one calculated from a sample.
- In addition, the sample variance is calculated using a slightly different formula from the one for the population variance. The sample variance is denoted by s2 and its formula is given by equations

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

- and, for grouped data: $s^2 = \frac{\sum f(x-\bar{x})^2}{n-1}$
- where n is the sample size. The reason $n - 1$ is used in the denominator rather than n (as one might expect) is the following. Our real interest is in the population variance, and the sample variance is an estimate of it. The former is measured by the dispersion around μ and the sample variance should ideally be measured around μ also. However, μ is unknown, so x is used in the formula instead. But the variation of the sample observations around x tends to be smaller than that around μ. Using $n - 1$ rather than n in the formula compensates for this and the result is an unbiased3 (i.e. correct on average) estimate of the population variance.

2.4 Summarising data using numerical techniques:

- ***The coefficient of variation***
- The measures of dispersion examined so far are all measures of absolute dispersion and, in particular, their values depend upon the units in which the variable is measured.
- It is therefore difficult to compare the degrees of dispersion of two variables which are measured in different units.
- For example, one could not compare wealth in the United Kingdom with that in Germany if the former uses £s and the latter euros for measurement. Nor could one compare the wealth distribution in one country between two points in time because inflation alters the value of the currency over time.
- The solution is to use a measure of relative dispersion, which is independent of the units of measurement. One such measure is the coefficient of variation, defined as:

$$coefficient\ of\ variation = \frac{\sigma}{\mu}$$

2.4 Summarising data using numerical techniques:

- ***Independence of units of measurement***
- It is worth devoting a little attention to this idea, that some summary measures are independent of the units of measurement and some are not, as it occurs quite often in statistics and is not often appreciated at first.
- A statistic which is independent of the units of measurement is one which is unchanged, even when the units of measurement are changed. It is therefore more useful in general than a statistic which is not independent, since one can use it to make comparisons, or judgements, without worrying too much about how it was measured.
- The mean is not independent of the units of measurement. If we are told the average income in the United Kingdom is 30 000, for example, we need to know whether it is measured in pounds sterling, euros or even dollars. The underlying level of income is the same, of course, but it is measured differently.

2.4 Summarising data using numerical techniques:

- ***Measuring deviations from the mean: z scores***
- Imagine the following problem.
- *A man and a woman are arguing over their career records. The man says he earns more than she does, so he is more successful. The woman replies that women are discriminated against and that, relative to other women, she is doing better than the man is, relative to other men. Can the argument be resolved?*
- Suppose the data are as follows: the average male salary is £19 500 and the average female salary £16 800. The standard deviation of male salaries is £4750 and for women it is £3800. The man's salary is £31 375, while the woman's is £26 800. The man is therefore £11 875 above the mean, and the woman is £10 000 above. However, women's salaries are less dispersed than men's, so the woman has done well to get to £26 800.
- One way to resolve the problem is to calculate the z score, which gives the salary in terms of the number of standard deviations from the mean. Thus for the man, the z score is

$$z = \frac{X - \mu}{\sigma}$$

2.4 Summarising data using numerical techniques:

- **Exercise 4** Given the data in exercise 2:
- (a) calculate the inter-quartile range, the variance and the standard deviation.
- (b) Calculate the coefficient of variation.
- (c) Approximately how much of the distribution lies within one standard deviation either side of the mean?

2.4 Summarising data using numerical techniques:

- ***Measuring skewness***
- he skewness of a distribution is the third characteristic that was mentioned earlier, in addition to location and dispersion.
- The wealth distribution is heavily skewed to the right, or positively skewed; it has its long tail in the right-hand end of the distribution.
- A measure of skewness gives a numerical indication of how asymmetric is the distribution.
- One measure of skewness, known as the coefficient of skewness, is

$$Coefficient\ of\ skewness = \frac{\sum f(x-\mu)^3}{N\sigma^3}$$

- The result of applying the above formula is positive for a right-skewed distribution (such as wealth), zero for a symmetric one, and negative for a left-skewed one. Table 1.11 shows the calculation for the wealth data (some rows are omitted for brevity).

2.4 Summarising data using numerical techniques:

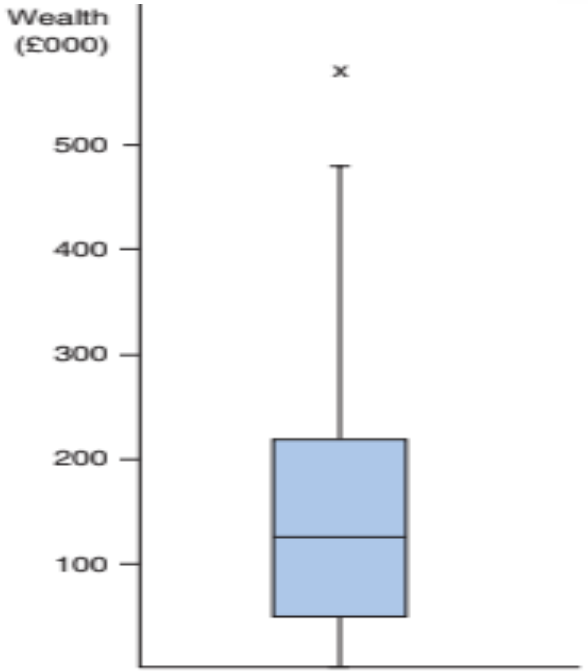
Table 1.11 Calculation of the skewness of the wealth data

Range	Mid-point x	Frequency f	$x - \mu$	$(x - \mu)^3$	$f(x - \mu)^3$
0–	5.0	1 668	−181.9	−6 016 132	−10 034 907 815
10 000–	17.5	1 318	−169.4	−4 858 991	−6 404 150 553
:	:	:	:	:	:
1 000 000–	1 500.0	198	1 313.1	2 264 219 059	448 315 373 613
2 000 000–	3 000.0	88	2 813.1	22 262 154 853	1 959 069 627 104
Total		18 677	3 898.8	24 692 431 323	2 506 882 551 023

- The measure of skewness is much less useful in practical work than measures of location and dispersion, and even knowing the value of the coefficient does not always give much idea of the shape of the distribution: two quite different distributions can share the same coefficient.
- In descriptive work, it is probably better to draw the histogram itself.

2.5 The box and whiskers diagram:

- Having calculated these various summary statistics, we can now return to a useful graphical method of presentation.
- This is the **box and whiskers diagram** (sometimes called a box plot) which shows the median, quartiles and other aspects of a distribution on a single diagram.
- A simple diagram thus reveals a lot of information about the distribution. Figure 1.15 shows the box plot for the wealth data. Here, Wealth is measured on the vertical axis.



2.5 The box and whiskers diagram:

- The rectangular box stretches (vertically) from the first to third quartile and therefore encompasses the middle half of the distribution.
- The horizontal line through it is at the median and lies slightly less than halfway up the box. This tells us that there is a degree of skewness even within the central half of the distribution, though it does not appear very severe.
- The two ‘whiskers’ extend above and below the box as far as the highest and lowest observations, excluding outliers.
- An outlier is defined to be any observation which is more than 1.5 times the inter-quartile range (which is the same as the height of the box) above or below the box.
- Earlier we found the IQR to be 173 443 and the upper quartile to be 221 135, so an (upper) outlier lies beyond 221 135 + 1.5 * 173 443 = 481 300. There are no outliers below the box as wealth cannot fall below zero.
- The top whisker is thus substantially longer than the bottom one, and indicates the extent of dispersion towards the tails of the distribution. The crosses indicate the outliers and in reality extend far beyond those shown in the diagram.

2.6 Time-series data:

- The data on the wealth distribution give a snapshot of the situation at particular points in time, and comparisons can be made between the 1979 and 2003 snapshots. Often, however, we wish to focus on the time-path of a variable and therefore we use time-series data.
- The techniques of presentation and summarizing are slightly different than for cross-section data.
- As an example, we use data on investment in the UK for the period 1973–2005. These data were taken from Statbase (http://www.statistics.gov.uk/statbase/) although you can find the data in Economic Trends Annual Supplement. Investment expenditure is important to the economy because it is one of the primary determinants of growth. Until recent years, the UK economy's growth record had been poor by international
- standards and lack of investment may have been a cause. The variable studied here is total gross (i.e. before depreciation is deducted) domestic fixed capital formation, measured in £m. The data are shown in Table 1.12.

2.6 Time-series data:

Table 1.12 UK investment, 1973–2005

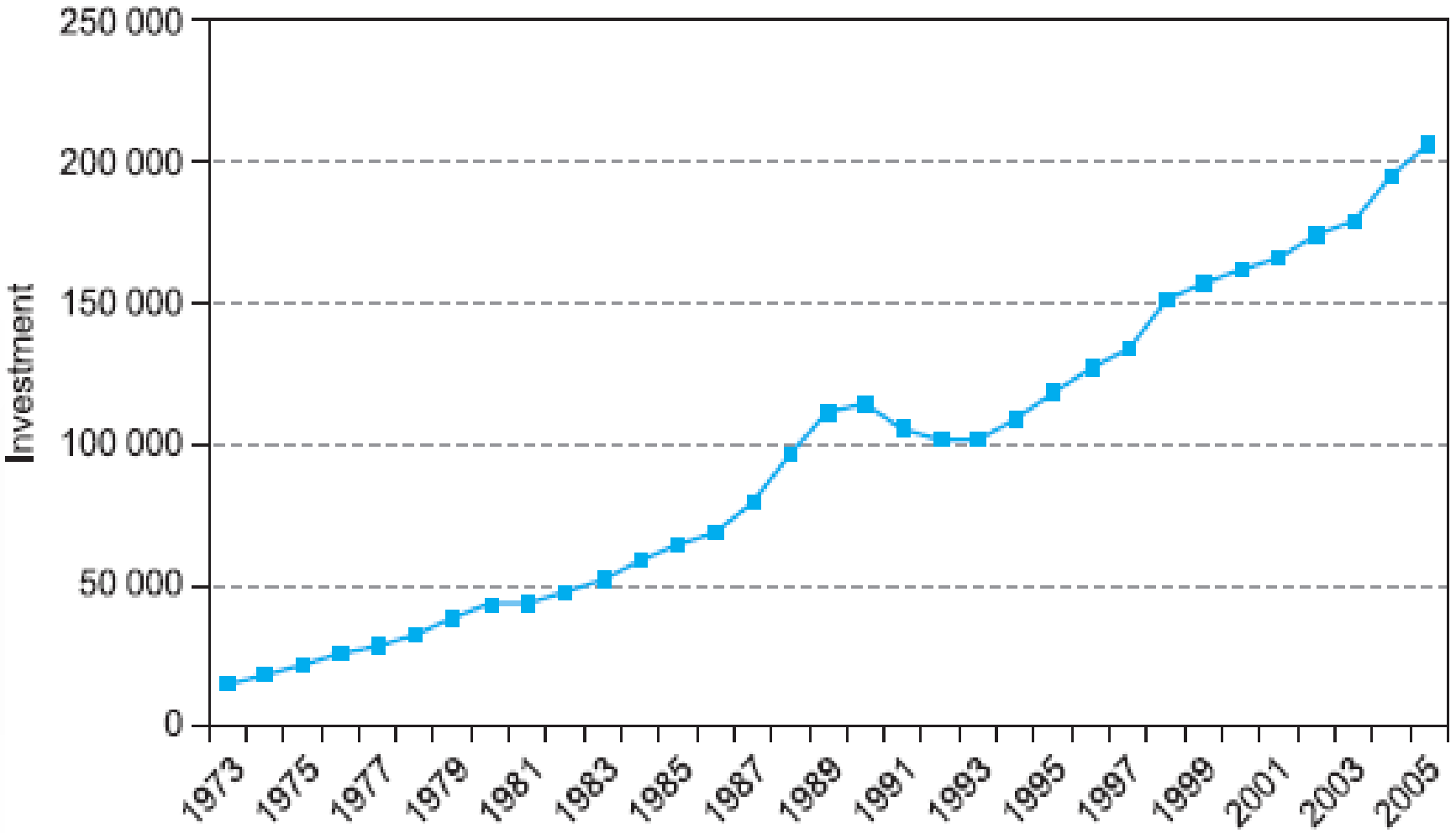
Year	Investment	Year	Investment	Year	Investment
1973	15 227	1984	58 589	1995	118 031
1974	18 134	1985	64 400	1996	126 593
1975	21 856	1986	68 546	1997	133 620
1976	25 516	1987	78 996	1998	151 083
1977	28 201	1988	96 243	1999	156 344
1978	32 208	1989	111 324	2000	161 468
1979	38 211	1990	114 300	2001	165 472
1980	43 238	1991	105 179	2002	173 525
1981	43 331	1992	101 111	2003	178 751
1982	47 394	1993	101 153	2004	194 491
1983	51 490	1994	108 534	2005	205 843

Note: Time-series data consist of observations on one or more variables over several time periods. The observations can be daily, weekly, monthly, quarterly or, as here, annually.

2.6 Time-series data:

- It should be remembered that the data are in current prices so that the figures reflect price increases as well as changes in the volume of physical investment. The series in Table 1.12 thus shows the actual amount of cash that was spent each year on investment.
- The techniques used below for summarising the investment data could equally well be applied to a series showing the volume of investment.
- **First of all** we can use graphical techniques to gain an insight into the characteristics of investment. Figure below shows a time-series graph of investment.
- The graph plots the time periods on the horizontal axis and the investment variable on the vertical.

2.6 Time-series data:



Note: The *X*, *Y* coordinates are the values {year, investment}; the first data point has the coordinates {1973, 15 227}, for example.

2.6 Time-series data:

- Plotting the data in this way brings out clearly some key features of the series:
- The trend in investment is upwards, with only a few years in which there was either no increase or a decrease.
- There is a 'hump' in the data in the late 1980s/early 1990s, before the series returns to its trend. Something unusual must have happened around that time. If we want to know what factors determine investment (or the effect of investment upon other economic magnitudes) we should get some useful insights from this period of the data.
- The trend is slightly non-linear – it follows an increasingly steep curve over time. This is essentially because investment grows by a percentage or proportionate amount each year. As we shall see shortly, it grows by about 8.5% each year. Therefore, as the level of investment increases each year, so does the increase in the level, giving a non-linear graph.

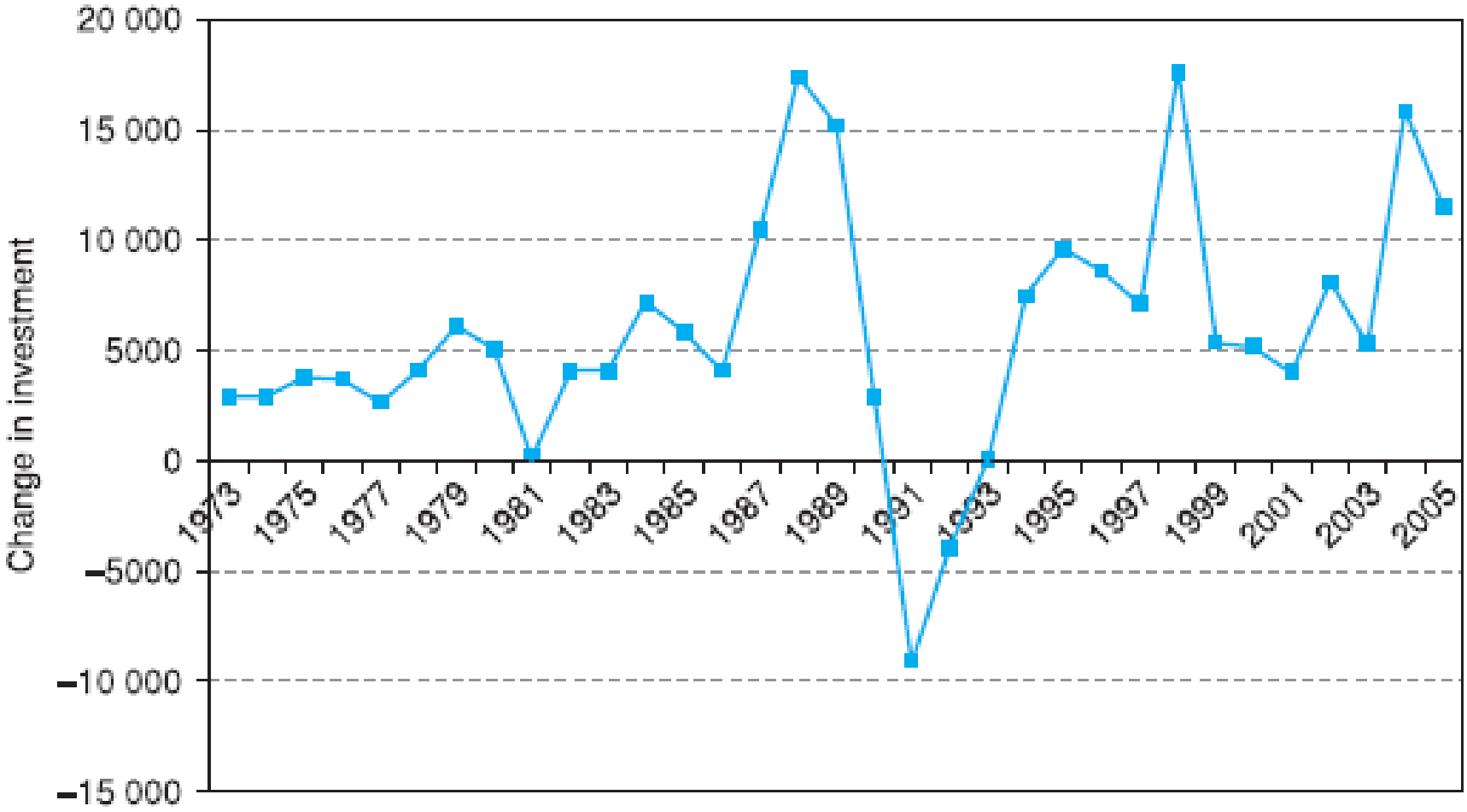
2.6 Time-series data:

- Successive values of the investment variable are similar in magnitude, i.e. the value in year t is similar to that in $t - 1$. In fact, the value in one year appears to be based on the value in the previous year, plus (in general) 8.5% or so. We refer to this phenomenon as serial correlation and it is one of the aspects of the data that we might wish to investigate.
- The ordering of the data matters, unlike the case with cross-section data where the ordering is usually irrelevant. In deciding how to model investment behaviour, we might focus on changes in investment from year to year.
- The series seems '**smoother**' in the earlier years (up to perhaps 1986) and exhibits greater volatility later on. In other words, there are greater fluctuations around the trend in the later years.
- We could express this more formally by saying that the variance of investment around its trend appears to change (increase) over time. This is known as **heteroscedasticity**; a constant variance is termed **homoscedasticity**.

2.6 Time-series data:

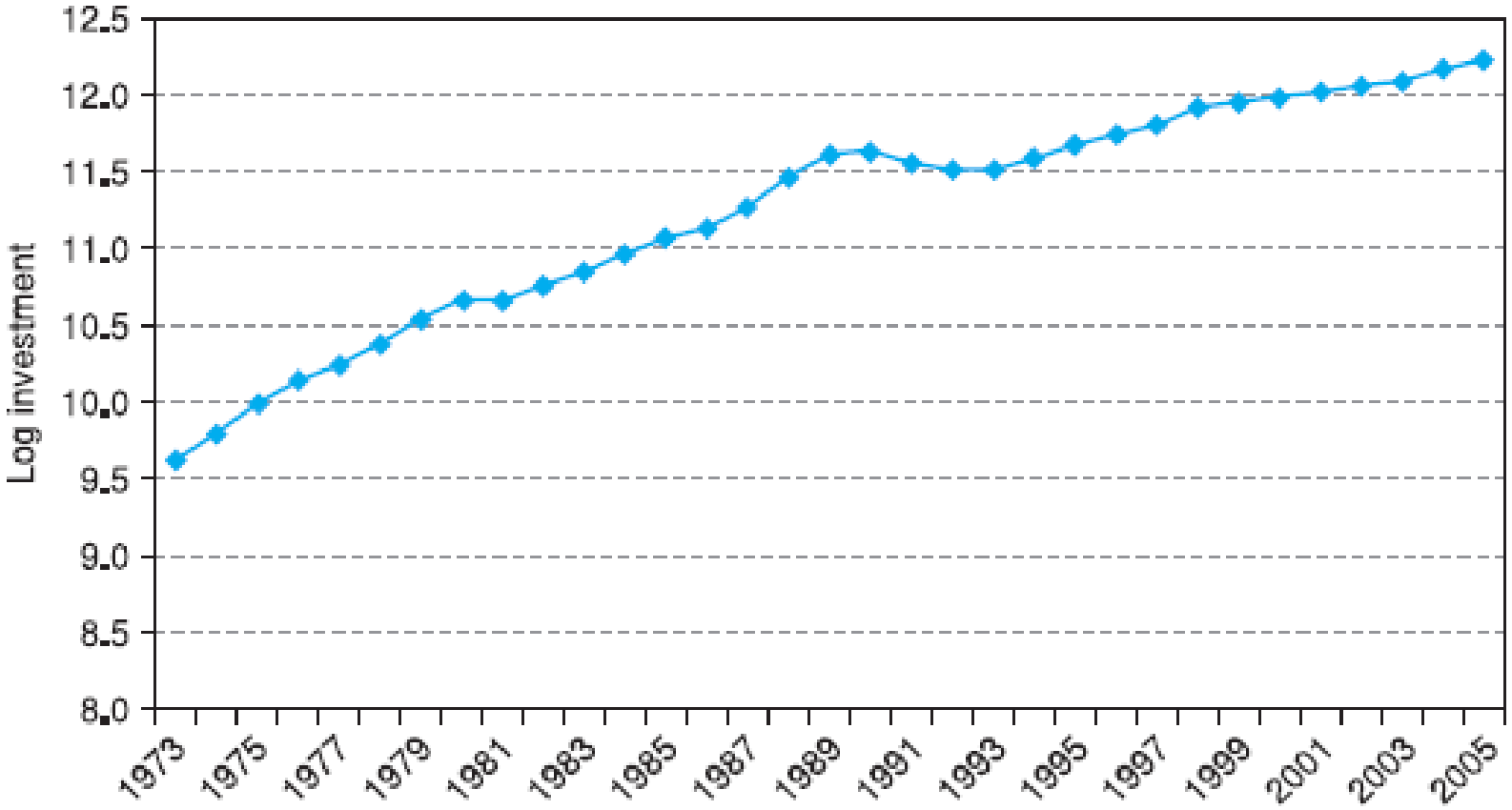
- We may gain further insight into how investment evolves over time by focusing on the change in investment from year to year.
- If we denote investment in year t by It then the change in investment, ΔI_t, is given by $I_t - I_{t-1}$.
- Next figure shows the changes in investment each year and Figure 1.17 provides a timeseries graph.
- The series is made up of mainly positive values, indicating that investment increases over time. It also shows that the increase grows each year, with perhaps some greater volatility (of the increase) towards the end of the period. The graph also shows dramatically the change that occurred around 1990.

2.6 Time-series data:



2.6 Time-series data:

Another useful way of examining the data is to look at the logarithm of investment. This transformation has the effect of straightening out the nonlinear investment series. Figure below shows the transformed values graphs the series. In this case we use the natural (base e) logarithm.



2.6 Time-series data:

- This new series is much smoother than the original one (as is usually the case when taking logs) and is helpful in showing the long-run trend, though it tends to mask some of the volatility of investment.
- The slope of the graph gives a close approximation to the average rate of growth of investment over the period (expressed as a decimal).
- This is calculated as follows

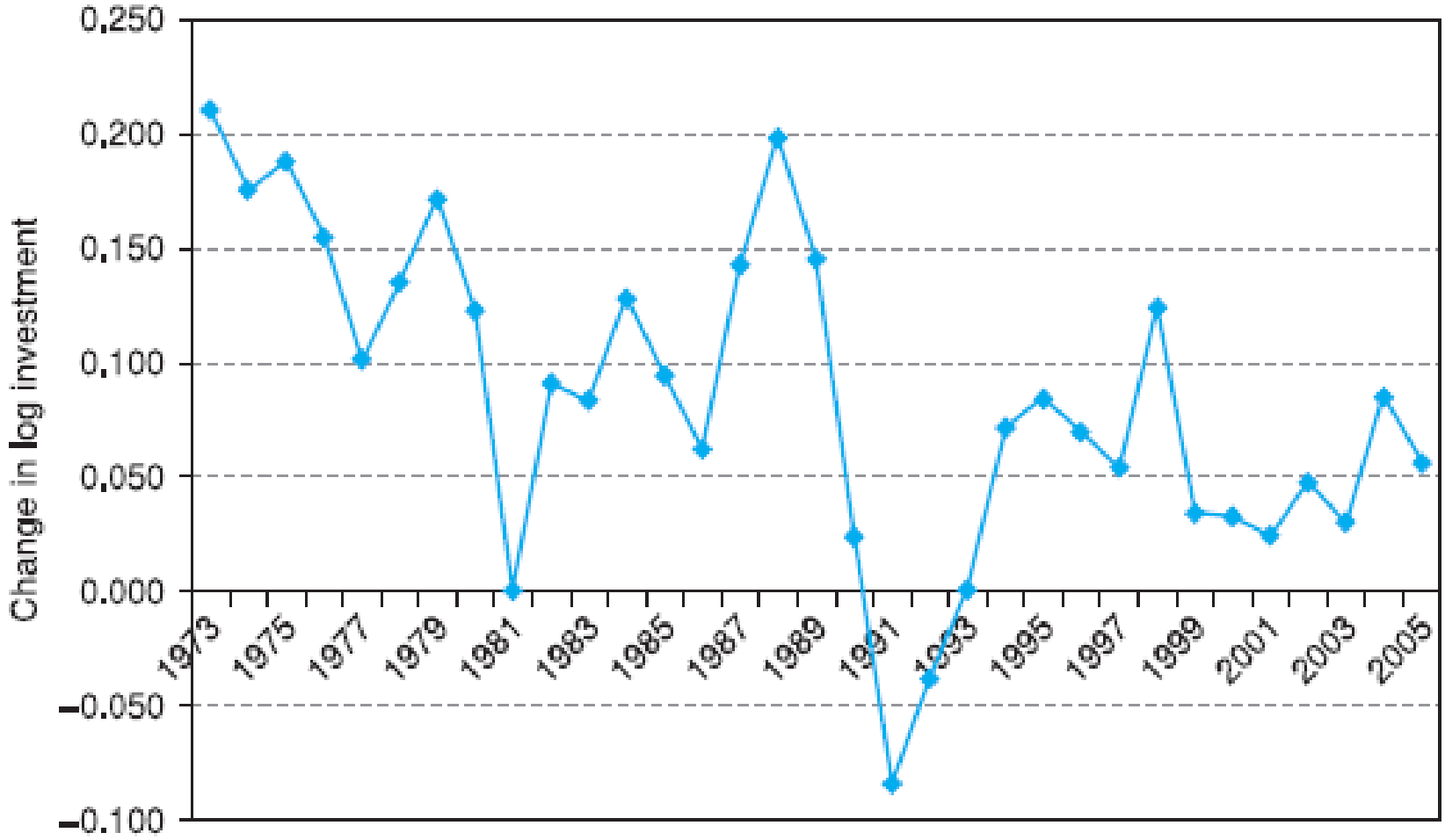
$$slope = \frac{\Delta \ln(I_t)}{Number\ of\ years}$$

- In our case, $slope = \frac{12.35-9.631}{32} = 0.081$ (i.e. 8.1% per annum)
- Note that although there are 33 observations, there are only 32 years of growth. A word of warning: you must use natural (base e) logarithms, not logarithms to the base 10, for this calculation to work. Remember also that the growth of the volume of investment will be less than 8.1% per annum, because part of it is due to price increases.
- The logarithmic presentation is useful when comparing two different data series: when graphed in logs it is easy to see which is growing faster – just see which series has the steeper slope.

2.6 Time-series data:

- A corollary of slope equation is that change in the natural logarithm of investment from one year to the next represents the percentage change in the data over that year.
- For example, the natural logarithm of investment in 1973 is 9.631, while in 1974 it is 9.806. The difference is 0.175, so the rate of growth is 17.5%.
- Remember that this is an approximation and the result of a quick and easy calculation. It is reasonably accurate up to a figure of about 20%.
- Finally we can graph the difference of the logarithm, as we graphed the difference of the level. This is shown in Figure below.
- This is quite revealing. It shows the series fluctuating about the value of approximately 0.08 (the average calculated in equation above), with a slight downwards trend.
- Furthermore, the series does not seem to show increasing volatility over time, as the others did.
- The graph therefore demonstrates that in proportionate terms there is no increasing volatility; the variance of the series around 0.08 does not change much over time (although 1991 still seems to be an ‘unusual’ observation).

2.6 Time-series data:



2.6 Time-series data:

- ***Measuring deviations from the mean: z scores***
- The graphs have revealed quite a lot about the data already, but we can also calculate numerical descriptive statistics as we did for the cross-section data. First we consider the mean, then the variance and standard deviation.
- ***The mean of a time series***
- We could calculate the mean of investment itself, but would this be helpful?
- Because the series is trended, it passes through the mean at some point between 1973 and 2005, but never returns to it.
- The mean of the series is actually £95.103bn, which is not very informative since it tells nothing about its value today, for instance. The problem is that the variable is trended, so that the mean is not typical of the series.
- The annual increase in investment is also trended, so is subject to the same criticism.

2.6 Time-series data:

- It is better in this case to calculate the average growth rate, as this is more likely to be representative of the whole time period.
- It seems more reasonable to say that a series is growing at (for example) 8% per annum than that it is growing at 5000 per annum.
- The average growth rate was calculated in slope equation as 8.1% per annum, by measuring the slope of the graph of the log investment series.
- That was stated to be an approximate answer. We can obtain an accurate value in the following way:
- 1. Calculate the overall growth factor of the series, i.e. $\frac{x_T}{x_1}$ where x_T is the final observation and x_1 is the initial observation. This is $\frac{x_T}{x_1} = 13.518$, i.e. investment expenditure is 13.5 times larger in 2005 than in 1973.
- (2) Take the $T-1$ root of the growth factor. Since $T = 33$ we calculate $\sqrt[32]{13.518} = 1.085$.
- (3) Subtract 1 from the result in the previous step, giving the growth rate as a decimal. In this case we have $1.085 - 1 = 0.085$.
- Thus the average growth rate of investment is 8.5% per annum, rather than the 8.1% calculated earlier.

2.6 Time-series data:

- ***The geometric mean***
- In calculating the average growth rate of investment we have implicitly calculated the geometric mean of a series.
- If we have a series of n values, then their geometric mean is calculated as the n^{th} root of the product of the values, i.e.

$$geometric\,mean = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- The x values in this case are the growth factors in each year, as in Table 1.15 (the values in intermediate years are omitted). The 'Π' symbol is similar to the use of Σ, but means 'multiply together' rather than 'add up'.
- The product of the 32 growth factors is 13.518 (the same as is obtained by dividing the final observation by the initial one – why?) and the 32nd root of this is 1.085. This latter figure, 1.085, is the geometric mean of the growth factors and from it we can derive the growth rate of 8.5% p.a. by subtracting 1.

2.6 Time-series data:

Whenever one is dealing with growth data (or any series that is based on a multiplicative process) one should use the geometric mean rather than the arithmetic mean to get the answer. However, using the arithmetic mean in this case generally gives only a small error, as is indicated below.

Table 1.15 Calculation of the geometric mean – annual growth factors

	Investment	Growth factors	
1973	15 227		
1974	18 134	1.191	(= 18 134/15 227)
1975	21 856	1.205	(= 21 856/18 134)
1976	25 516	1.167	Etc.
⋮	⋮	⋮	
2002	173 525	1.049	
2003	178 751	1.030	
2004	194 491	1.088	
2005	205 843	1.058	

Note: Each growth factor simply shows the ratio of that year's investment to the previous year's.

2.6 Time-series data:

- ***The geometric mean***
- We have seen that when calculating rates of growth one should use the geometric mean, but if the growth rate is reasonably small then taking the arithmetic mean of the growth factors will give approximately the right answer.
- Use of the arithmetic mean is justified in this context if one needs only an approximation to the right answer and annual growth rates are reasonably small.
- It is usually quicker and easier to calculate the arithmetic rather than geometric mean, especially if one does not have a computer to hand.
- By now you might be feeling a little overwhelmed by the various methods we have used, all to get an idea of the average – methods which give similar but not always identical answers.

2.6 Time-series data:

- Let us summarise the findings:
- (a) measuring the slope of the log graph: gives approximately the right answer;
- (b) transforming the slope using the formula $e^b - 1$: gives the precise answer (b is the measured slope);
- (c) calculating $\sqrt[T-1]{\frac{x_T}{x_1}} - 1$: gives the precise answer (as in (b));
- (d) calculating the geometric mean of the growth factors: gives the precise answer;
- (e) calculating the arithmetic mean of the growth factors: gives approximately the right answer (although not the same approximation as (a) above).
- Remember also that the 'precise' answer could be slightly misleading if either initial or final value is an outlier.

2.6 Time-series data:

- ***The variance of a time series***
- The variance of the investment data can be calculated, but it would be uninformative in the same way as the mean. As the series is trended, and this is likely to continue in the longer run, the variance is in principle equal to infinity.
- The calculated variance would be closely tied to the sample size: the larger it is, the larger the variance.
- Again it makes more sense to calculate the variance of the growth rate, which has little trend in the long run.
- This variance can be calculated from the formula

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} = \frac{\sum x^2 - n\bar{x}^2}{n - 1}$$

- where $\bar{x}$ is the average rate of growth.
- Note three things about this calculation: **first,** we have used the arithmetic mean (using the geometric mean makes very little difference); **second,** we have used the formula for the sample variance since the period 1974–2005 constitutes a sample of all the possible data we could collect; and **third,** we could have equally used the growth factors for the calculation of the variance

2.6 Time-series data:

- ***Example:*** Given the following data, find the variance

Year	1999	2000	2001	2002	2003	
Price of a PC	1100	900	800	750	700	

- *Solution:*
- *1. The overall growth factor is ….*
- *2. The annual rate (the fourth root of step 1) …*
- *3. The variance is given by …*

2.4 Summarising data using numerical techniques:

- **Exercise 5** Given the following data in

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Profit	50	60	25	–10	10	45	60	50	20	40
Sales	300	290	280	255	260	285	300	310	300	330

- (a) Draw a multiple time series graph of the two variables. Label both axes appropriately and provide a title for the graph.
- (b) Adjust the graph by using the right-hand axis to measure profits, the left-hand axis sales. What difference does this make?
- (c) calculate the average level of profit over the time period and the average growth rate of profit over the period. Which appears more useful?
- (d) Calculate the variance of profit and compare it to the variance of sales.

2.7 Graphing bivariate data, the scatter diagram:

- The analysis of investment is an example of the use of univariate methods: only a single variable is involved. However, we often wish to examine the relationship between two (or sometimes more) variables and we have to use bivariate (or multivariate) methods.
- To illustrate the methods involved we shall examine the relationship between investment expenditures and gross domestic product (GDP).
- Economics tells us to expect a positive relationship between these variables, higher GDP is usually associated with higher investment.
- A scatter diagram (also called an XY chart) plots one variable (in this case investment) on the y axis, the other (GDP) on the x axis, and therefore shows the relationship between them. For example, one can see whether high values of one variable tend to be associated with high values of the other.

2.7 Graphing bivariate data, the scatter diagram:

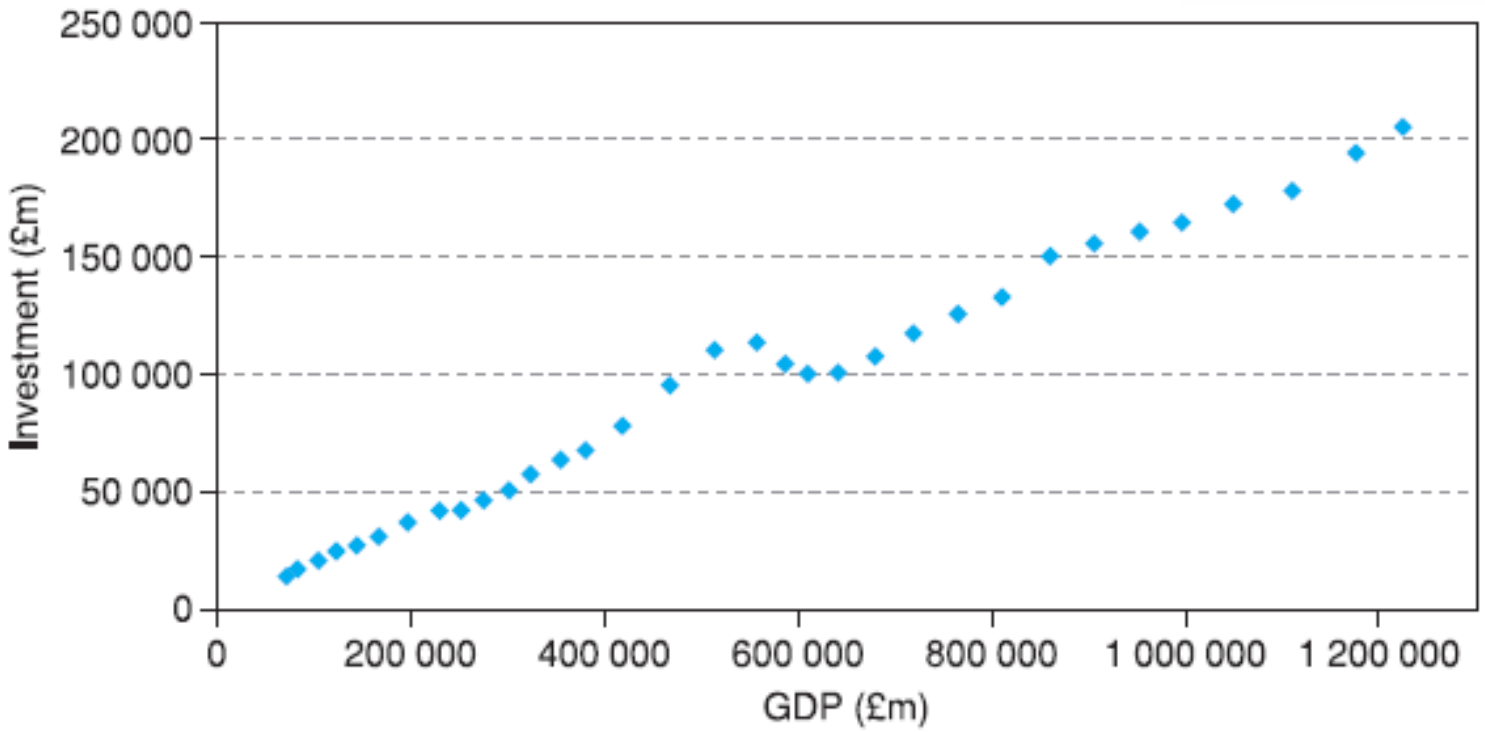
- Table 1.17 provides data on GDP for the UK.

Table 1.17 GDP data

Year	GDP	Year	GDP	Year	GDP
1973	74 020	1984	324 633	1995	719 747
1974	83 793	1985	355 269	1996	765 152
1975	105 864	1986	381 782	1997	811 194
1976	125 203	1987	420 211	1998	860 796
1977	145 663	1988	469 035	1999	906 567
1978	167 905	1989	514 921	2000	953 227
1979	197 438	1990	558 160	2001	996 987
1980	230 800	1991	587 080	2002	1 048 767
1981	253 154	1992	611 974	2003	1 110 296
1982	277 198	1993	642 656	2004	1 176 527
1983	302 973	1994	680 978	2005	1 224 715

2.7 Graphing bivariate data, the scatter diagram:

- Figure below shows the relationship for investment and GDP.



2.7 Graphing bivariate data, the scatter diagram:

- The chart shows a strong linear relationship between the two variables, apart from a curious dip in the middle.
- This reflects the sharp fall in investment after 1990, which is not matched by a fall in GDP (if it were, the XY chart would show a linear relationship without the dip).
- It is important to recognise the difference between the time-series plot and the XY chart.
- Because of inflation later observations tend to be towards the top right of the XY chart (both investment and GDP are increasing over time) but this does not have to happen; if both variables fluctuated up and down, later observations could be at the bottom left (or centre, or anywhere).
- By contrast, in a time series plot, later observations are always further to the right.

2.8 Data transformations: “Do it yourself”

In analysing employment and investment data in the examples above we have often changed the variables in some way in order to bring out the important characteristics. In statistics one usually works with data that have been transformed in some way rather than using the original numbers. It is therefore worth summarising the main data transformations available, providing justifications for their use and exploring the implications of such adjustments to the original data. You should be able to briefly deal with the following transformations:

- rounding;
- grouping;
- dividing or multiplying by a constant;
- differencing;
- taking logarithms;
- taking the reciprocal;
- deflating.