Section 1
First principles

1

What's in a number?

Chapter 1 provides a brief review of the development of quantitative analysis in Geography, Earth and Environmental Science and related disciplines. It also discusses the relative merits of using numerical data and how numbers can be used to represent qualitative characteristics. A brief introduction to mathematical notation and calculation is provided to a level that will help readers to understand subsequent chapters. Overall this introductory chapter is intended to define terms and to provide a structure for the remainder of the book.

Learning outcomes

This chapter will enable readers to:

- outline the difference between quantitative and qualitative approaches, and their relationship to statistical techniques;
- describe the characteristics of numerical data and scales of measurement;
- recognize forms of mathematical notation and calculation that underlie analytical procedures covered in subsequent chapters;
- plan their reading of this text in relation to undertaking an independent research investigation in Geography and related disciplines.

Practical Statistics for Geographers and Earth Scientists Nigel Walford

1.1 Introduction to quantitative analysis

Quantitative analysis comprises one of two main approaches to researching and understanding the world around us. In simple terms quantitative analysis can be viewed as the processing and interpretation of data about things, sometimes called phenomena, which are held in a numerical form. In other words, from the perspective of Geography and other Earth Sciences, it is about investigating the differences and similarities between people and places that can be expressed in terms of numerical quantities rather than words. In contrast, qualitative analysis recognizes the uniqueness of all phenomena and the important contribution towards understanding that is provided by unusual, idiosyncratic cases as much as by those conforming to some numerical pattern. Using the two approaches together should enable researchers to develop a more thorough understanding of how processes work that lead to variations in the distribution of phenomena over the Earth's surface than by employing either methodology on its own.

If you are reading this book as a student on a university or college course, there will be differences and similarities between you and the other students taking the same course in terms of such things as your age, height, school level qualifications, home town, genetic make-up, parental annual income and so on. You will also be different because each human being, and for that matter each place on the Earth, is unique. There is no one else exactly like you, even if you have an identical twin, nor is there any place exactly the same as where you are reading this book. You are different from other people because your own attitudes, values and feelings have been moulded by your upbringing, cultural background and physical characteristics. In some ways, it is the old argument of nature versus nurture, but in essence we are unique combinations of both sets of factors. You may be reading this book in your room in a university hall of residence, and there are many such places in the various countries of the world and those in the same institution often seem identical, but the one where you are now is unique. Just as the uniqueness of individuals does not prevent analysis of people as members of various different groups, so the individuality of places does not inhibit investigation of their distinctive and shared characteristics.

What quantitative analysis attempts to do is concentrate on those factors that seem to be important in producing differences and similarities between individual phenomena and to disregard those producing aberrant outcomes. Confusion between quantitative and qualitative analysis may arise because the qualitative characteristics of people and places are sometimes expressed in numerical terms. For example, areas in city may be assigned to a series of categories, such as downtown, suburbs, shopping mall, commercial centre, housing estate and so on, to which numerical codes can be attached, or a person's gender may be labelled as male or female with equivalent numerical values such as 1 or 2. Just as qualitative characteristics can be turned into numerical codes, so too can numerical measurements be converted, either singly or in combination, into descriptive labels. Many geodemographic descriptions of neighbourhoods, such as (e.g. *Old people, detached houses; Larger families, prosperous* *suburbs; Older rented terraces;* and *Council flats, single elderly*) are based on having taken a selection of different socioeconomic and demographic numerical counts for different locations and combined them in an analytical melting pot so as to produce a label describing what the places are like.

The major focus of this book is on quantitative analysis as applied to Geography and other Earth Sciences. However, this should not be taken to imply that quantitative analysis is in some definitive sense regarded as 'better', or even more scientific, than qualitative analysis. Nor are these approaches mutually exclusive, since researchers from many disciplines have come to appreciate the advantages of combining both forms of analysis in recent years. This book concentrates on quantitative analysis since for many students, and perhaps researchers, dealing with numbers and statistics is difficult, and trying to understand what these can usefully tell us about the 'real' Geography or Earth Science topics that interest us is perplexing. Why should we bother with numbers and statistics, when all we want to do is to understand the process of globalization in political economy, or to explain why we are experiencing a period of global temperature increase, or to identify areas vulnerable to natural hazards?

We can justify bothering with numbers and statistics to answer such research questions in a number of different ways. As with most research, when we actually try to explain things such as globalization, global warming and the occurrence of natural hazards, the answers often seem like commonsense and perhaps even rather obvious. Maybe this is a sign of 'good' research because it suggests the process of explaining such things is about building on what has become common knowledge and understanding. If this is true, then ongoing academic study both rests upon and questions the endeavours of previous generations of researchers, and puts across complex issues in ways that can be generally understood. Yet despite the apparent certainty with which the answers to research questions might be conveyed, these are often underlain by an analysis of numerical information that is anything but certain. It is likely that the results of the research are correct, but they may not be. Using statistical techniques gives us a way of expressing this uncertainty and of hedging our bets against the possibility that our particular set of results has only arisen by chance. At some time in the future another researcher might come along and contradict our findings.

But what do we really mean by the phrase 'the results of the research'. For the 'consumers' of research, whether the public at large or particular professional groups, the findings, results or outcomes of research are often some piece of crucial factual information. The role of such information is to either confirm facts that are already known or believed, or to fulfil the unquenchable need for new 'facts'. For the academic, these detailed factual results may be of less direct interest than the implications of the research findings, perhaps with regard to some overarching theory. The student undertaking a research investigation as part of the programme of study sits somewhere, perhaps uncomfortably between these two positions. Realistically, many undergraduate students recognize that their research endeavours are unlikely to contribute significantly to theoretical advance, although obviously there are exceptions.

Yet they also recognize that their tutors are unlikely to be impressed simply by the presentation of new factual information. Further, such research investigations are typically included in undergraduate degree programmes in order to provide students with a training that prepares them for a career where such skills as collecting and assimilating information will prove beneficial, whether this be in academia or more typically in other professional fields. Students face a dilemma to which there is no simple answer. They need to demonstrate that they have carried out their research in a rigorous scientific manner using appropriate quantitative and qualitative techniques, but they do not want to overburden the assessors with an excess of detail that obscures the implications of their results.

In the 1950s and 1960s a number of academic disciplines 'discovered' quantitative analysis and few geography students of the last five decades can fail to have heard of the so-called 'quantitative revolution' in their subject area, and some may not have forgiven the early pioneers of this approach. There was, and to some extent still is, a belief that the principles of rigour, replication and respectability enshrined in scientific endeavour sets it apart from, and possibly above, other forms of more discursive academic enquiry. The adoption of the quantitative approach was seen implicitly, and in some cases explicitly, as providing the passport to recognition as a scientific discipline. Geography and other Earth Sciences were not alone, although perhaps were more sluggish than some disciplines, in seeking to establish their scientific credentials. The classical approach to geographical enquiry followed on from the colonial and exploratory legacies of the 18th and 19th centuries. This permeated regional geography in the early 20th century and concentrated on the collection of factual information about places. Using this information to classify and categorize places seemed to correspond with the inductive scientific method that served the purpose of recognizing pattern and regularity in the occurrence of phenomena. However, the difficulty of establishing inductive laws about intrinsically unique places and regions led other geographers to search for ways of applying the deductive scientific method, which was also regarded as more rigorous. The deductive method involves devising hypotheses with reference to existing conditions and testing them using empirical evidence obtained through the measurement and observation of phenomena.

Geography and to a lesser extent the other Earth Sciences have emerged from a period of self-reflection on the scientific nature of their endeavour with an acceptance that various philosophies can coexist and further their collective enterprise. Thus, many university departments include physical geographers and Earth scientists, adhering to generally positivist scientific principles, working alongside human geographers following a range of traditions including humanism, Marxism and structuralism as well as more positivist social science. The philosophical basis of geographical and Earth scientific enquiry has received a further twist in recent decades on account of the growing importance of information and communications technology (ICT). Hence, students in academic departments need to be equipped with the skills not only to undertake research investigations in these areas, but also to handle geographical and spatial data in a digital environment.

Johnston (1979) commented that statistical techniques provide a way of testing hypotheses and the validity of empirical measurements and observations. However, the term statistics is used in a number of different ways. In general usage, it typically refers to the results of data collection by means of censuses and surveys that are published in books, over the Internet or on other media. The associated term 'official statistics' is usually reserved for information that has been collected, analysed and published by national, regional or local government and are therefore deemed to have a certain authority and a connotation of conveying the 'truth'. This belief may be founded upon a presumption of impartiality and rigour, although such neutrality of method or intent cannot realistically be justified or assumed in all instances. Statistics also refers to a branch of mathematics that may be used in scientific investigations to substantiate or refute the results of scientific research. In this sense statistics also has a double meaning, either comprising a series of **techniques** ranging from simple summarizing measures to complex models involving many variables, or the term may refer to the numerical **quantities** produced by these techniques. All these senses of the term statistics are relevant to this text, since published statistics may well contribute to research investigations, and statistical techniques and the measures associated with them are an essential part of quantitative analysis. Such techniques are applied to numerical data and serve two general purposes: to confirm or otherwise the significance of research results towards the accumulation of knowledge with respect to a particular area of study; and to establish whether empirical connections between different characteristics for a given set of phenomena are likely to be genuine or spurious.

Different areas of scientific study and research have over the years carved out their own particular niches. For example, in simplistic terms the Life Sciences are concerned with living organisms, the Chemical Sciences with organic and inorganic materials, Political Science with national and international government, Sociology with social groups and Psychology with individuals' mental condition. When these broad categories have become too general then often subdivision occurred with the emergence of fields in the Life Sciences such as cell biology, biochemistry, physiology, etc. Geography and the other Earth Sciences do not seem to fit easily into this seemingly straightforward partitioning of scientific subject matter, since their broad perspective leads to an interest in all the things covered by other academic disciplines, even to the extent of Earth Scientists transferring their allegiance to examine **terrestrial** processes on other planetary and celestial bodies. No doubt if, or when, ambient intelligent life is found on other planets, 'human' geographers will be there investigating its spatial arrangement and distribution. It is commonly argued that the unifying focus of Geography is its concern for the spatial and earthly context in which those phenomena of interest to other disciplines make their home. Thus, for example human geographers are concerned with the same social groups as the sociologist, but emphasize their spatial juxtaposition and interaction rather than the social ties that bind them, although geographers cannot disregard the latter. Similarly, geochemists focus on the chemical properties of minerals not only for their individual characteristics but also for how assemblages combine to form different rocks in distinctive locations. Other disciplines may wonder what Geography and Earth Science add to their own academic endeavour, but geographers and Earth scientists are equally certain that if their area of study did not exist, it would soon need to be invented.

The problem that all this raises for geographers and Earth scientists is how to define and specify the units of observation, the things that are of interest to them, and them alone. One possible solution that emerged during the quantitative revolution was that geography was pre-eminently the science of spatial analysis and therefore it should be concerned with discovering the laws that governed spatial processes. A classical example of this approach in human geography was the search for regions or countries where the collections of settlements conformed to the spatial hierarchy anticipated by central place theory. Physical geographers and Earth scientists also became interested in spatial patterns. Arguably they had more success in associating the occurrence of environmental phenomena with underlying explanatory processes, as evidenced by the development plate tectonics theory in connection with the spatial pattern of earthquake and volcanic zones around the Earth. According to this spatial analytic approach, once the geographer and Earth scientist have identified some spatially distributed phenomena, such as settlements, hospitals, earthquakes or volcanoes, then their investigation can proceed by measuring distance and determining pattern.

This foray into spatial analysis was soon undermined, when it became apparent that exception rather than conformity to proposed spatial laws was the norm. Other approaches, or possibly paradigms, emerged, particularly in human geography, that sought to escape from the straightjacket of positivist science. Advocates of Marxist, behavioural, politicoeconomic and cultural geography have held sway at various times during the last 40 years. However, it is probably fair to say that none of these have entirely rejected using numerical quantities as a way of expressing geographical difference. Certainly in physical geography and Earth Science where many would argue that positivism inevitably still forms the underlying methodology, quantitative analysis has never receded into the background. Despite the vagaries of all these different approaches most geographers and Earth scientists still hold on to the notion that what interests them and what they feel other people should be reminded of is that the Earth, its physical phenomena, its environment and its inhabitants are differentiated and are unevenly distributed over space.

From the practical perspective of this text, what needs to be decided is what constitutes legitimate data for the study of Geography and Earth Science. Let us simply state that a geographical or Earth scientific dataset needs to comprise a collection of data items, facts if you prefer, that relate to a series of spatially distributed phenomena. Such a collection of data needs to relate to at least one discrete and defined section of the Earth and/or its immediate atmosphere. This definition deliberately provides wide scope for various types and sources of data with which to investigate geographical and Earth scientific questions.

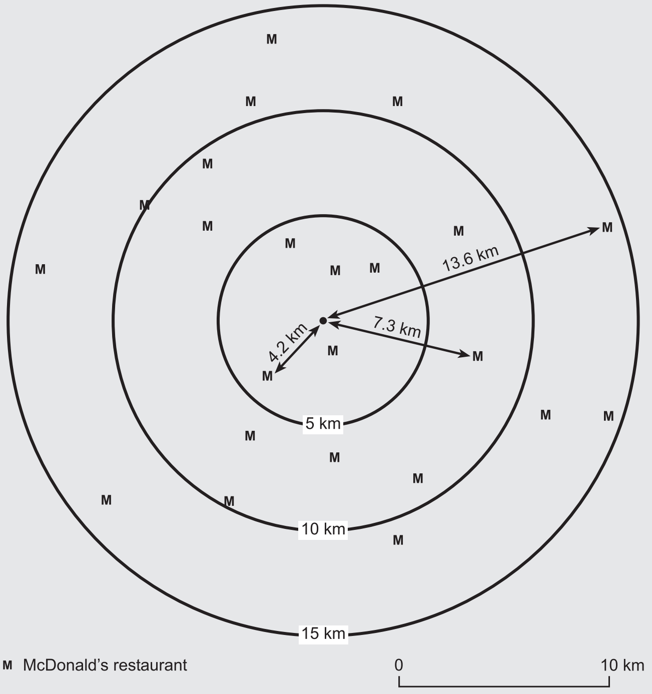
The spatial focus of Geography and the Earth Sciences has two significant implications. First, the location of the phenomena or observations units, in other words where they are on or near the Earth's surface is regarded as important. For instance investigations into landforms associated with calcareous rocks need to recognize whether these are in temperate or tropical environments. The nature of such landforms including their features and structures is in part dependent upon prevailing climatic conditions either now or in the past. Second, the spatial arrangement of phenomena may in itself be important, which implies that a means of numerically quantifying the position of different occurrences of the same category of observation may be required. In this case, spatial variables such as area, proximity, slope angle, aspect and volume may form part of the data items captured for the phenomena under investigation.

1.2 Nature of numerical data

We have already seen that quantitative analysis uses numbers in two ways, as shorthand labels for qualitative characteristics to save dealing with long textual descriptions or as actual measurements denoting differences in magnitude. Underlying this distinction is a division of data items into **attributes** and **variables**. Williams (1984, p. 4) defines an attribute as 'a quality … whereby items, individuals, objects, locations, events, etc. differ from one another.' He contrasted these with variables that 'are *measured* … assigned numerical values relative to some standard – the *unit of measurement*' (Williams, 1984, p. 5). Examples of attributes and variables from Geography and other Earth Sciences are seemingly unbounded in number and diversity, since these fields of investigation cover such a wide range of subject areas. Relevant attributes include such things as rock hardness, soil type, land use, ethnic origin and housing tenure, whereas stream discharge, air temperature, population size, journey time and number of employees are examples of geographical variables. The terms attribute and variable are sometimes confused and applied interchangeably, although we will endeavour to keep to the correct terminology here.

The subdivision of numerical data into different types can also be taken further, into what are known as the **scales of measurement**. The four scales are usually known as **nominal**, **ordinal**, **interval** and **ratio** and can be thought of as a sequence implying a greater degree of detail as you progress from the nominal to the ratio. However, strictly speaking the nominal is not a scale of measurement, since it only applies to attributes and therefore is an assessment of qualitative difference rather than magnitude between observations. The other three scales provide a way of measuring difference between observations to determine whether one is smaller, larger or the same as any other in the set. Box 1.1 summarizes the features of each of these measurement scales and shows how it is possible to discard information by moving from the interval/ratio to the nominal scale. Such collapsing or recoding of the values for data

Box 1.1a: Scales of measurement: location of a selection of McDonalds restaurants in Pittsburgh, USA.



Box 1.1b: Scales of measurement: definitions and characteristics.

Nominal scale attributes record a qualitative difference between observations.

Ordinal scale variables rank one observation against others in the set with respect to the ascending or descending data values of a particular characteristic. These are particularly useful when analysing information obtained from secondary sources where the raw data are unavailable.

Interval scale variables record differences of magnitude between observations in a set where the measurement units have an arbitrary zero.

Ratio scale variables record differences of magnitude between observations in a set where the ratio between the values for any two observations remains constant, irrespective of the measurement units which have an absolute zero.

For most practical purposes the interval and ratio scales, sometimes referred to as scalar measurements, are treated as equivalent with respect to various analytic procedures. The four scales of measurement are listed in order of increasing sophistication and it is possible collapse the detail for a characteristic measured on the interval/ratio scales to the ordinal and to the nominal. This process results in a loss of detail as information is discarded at each stage, but this simplification may be necessary when analysing data from across different measurement scales. Box 1.1c illustrates that progressive loss of detail occurs with respect to the distance measurements between the centre of Pittsburgh and a selection of the McDonalds restaurants in the city. Distance is measured on the ratio scale, with an absolute zero in this case indicating that a restaurant is located exactly at the centre. Sorting these raw data measurements in ascending order and assigning rank scores converts the information into an ordinal scale variable. There are two restaurants closest to the centre at 2.2 km from this point and the furthest is 18.4 km. Further loss of detail results from allocating a qualitative code to denote whether the restaurant is located in the inner, middle or outer concentric zone around the city centre. Although it is feasible to collapse the detailed raw data of ratio/interval variables in this way, it is not possible to go in the other direction from qualitative attributes, through ordinal to interval/ratio scale variables.

Box 1.1c: Collapsing between scales of measurement: distance of a selection of McDonalds restaurants from centre of Pittsburgh, USA.

Restaurant	Distance to centre	Sorted	Rank score	Concentric zone
No.	Ratio variable	distance	Ordinal variable	Nominal attribute
1	17.5	2.2	1	1 (Inner)
2	15.1	2.2	1	1 (Inner)
3	14.2	2.8	3	1 (Inner)
4	10.3	3.6	4	1 (Inner)
5	6.0	4.2	5	1 (Inner)
6	8.2	6.0	6	2 (Middle)
7	9.2	6.4	7	2 (Middle)
8	10.8	7.0	8	2 (Middle)
9	4.2	7.2	9	2 (Middle)
10	3.6	7.3	10	2 (Middle)
11	2.2	8.1	11	2 (Middle)
12	7.0	8.2	12	2 (Middle)
13	2.2	9.2	13	2 (Middle)
14	2.8	10.3	14	3 (Outer)
15	11.2	10.8	15	3 (Outer)
16	8.1	11.2	16	3 (Outer)
17	6.4	11.2	16	3 (Outer)
18	7.2	13.6	18	3 (Outer)
19	7.3	13.9	19	3 (Outer)
20	11.2	14.2	20	3 (Outer)
21	13.9	15.1	21	3 (Outer)
22	13.6	17.5	22	3 (Outer)
23	18.4	18.4	23	3 (Outer)

items may be carried out for various reasons, but may be necessary when combining a qualitative data item, such as housing tenure, with a quantitative one, such as household income in your analysis.

In most investigations the data items will come from more than one of these scales of measurement, and may possibly include all four. This makes it even more important to be able to recognize the different categories, because some types of statistical quantity and technique are appropriate for data items on one measurement scale and not others. In other words, statistical analysis is not just a question of 'pick and mix', more 'horses for courses'. An important distinction is between discrete and continuous measurements, respectively separating the nominal and ordinal from the interval and ratio scales. Discrete data values occur in situations where observations can only take certain values, usually integers (whole numbers), whereas theoretically any value is allowed with continuous measurements except where there is an absolute zero. Information technology has helped investigators to carry out research that applies statistical analysis to numerical data, but generally computer software is not intelligent enough to know whether a given set of integers are category labels for qualitative attributes, an ordinal sequence of integers or real data values on the interval or ratio scales. Investigators need to intervene to select appropriate types of analysis and not just tell the software to produce the same set of analyses for all their attributes and variables.

Investigations using attributes and variables usually involve holding the data in a numerical format. In the case of nominal attributes this usually involves simple integer numbers such as 1, 2, 3 up to *J*, where *J* equals the number of categories and there is a one-to-one correspondence between the number of categories and integers. However, in some cases negative integers may be used. An example would be attitudinal scales, where respondents might be asked to say how they felt about a series of statements in terms of Strongly Disagree, Disagree, Neutral, Agree or Strongly Agree. These qualitative labels may be linked to the numbers −2, −1, 0, 1 and 2, rather than 1, 2, 3, 4 and 5. The ordinal scale is based on ranking observations in order from smallest to largest, or vice versa. These are often converted into a rank score (first, second, third, etc.) corresponding to either the largest or smallest value depending upon whether they are sorted in ascending or descending numerical order. These rank scores are usually stored as integers, although values are repeated if two or more observations are tied (see Box 1.1c). Thus, there may be three observations tied at 15 and no 16 or 17, and the next observation is 18, although arguably the three observations could all be ranked 16. Interval and ratio scale variables are usually recorded to a certain number of decimal places, 2 or 3 for example, even if particular values happen to be integers (e.g. 5.0 km). Computers used to analyse your data may store the values with a far higher degree of precision, for example to 25 decimal places. However, such precision may be spurious, since you may only have captured the data values to one decimal place or perhaps only as whole numbers. An example will illustrate the problem. This book follows the convention of using the International System of metric units (kilometres, degrees Celsius, hectares, etc.) employed in scientific research. However, many people in the USA, UK and some other countries will be unfamiliar with this system and will give distances in miles, temperatures in Fahrenheit and areas in acres. It is possible to convert between these two sets of scales, for instance 1 hectare equals 2.4691 acres (4 decimal places). If we apply this conversion to a specific acreage we may end up with what appears to be a highly accurate number of hectares, but in reality is an entirely spurious level of precision, for example 45 acres divided by 2.4691 equals 18.2252642661138618119962739 ha to 25 decimal places, which could for most practical purposes be shortened to 18.2 ha without any significant loss of information.

Before leaving our consideration of the characteristics of numerical data, it is as well to recognize that there are other types of data item that we may wish to use in our analysis. Dates, for example relating to a specific day, month or year, are another useful category of information that provides investigations with a temporal dimension. It is worth distinguishing between two different types of date depending upon how this temporal reference is obtained. One is recorded by the investigator without directly consulting the observation units and refers to the date when the specific data were collected. An example of this would be the day, month and year, and the start and end times when an individual survey questionnaire was completed or when a field measurement was recorded. The other type of date is obtained from the data subjects themselves and is more common in investigations involving people. Again in a survey of company directors respondents may be asked the date when it was founded, farmers may be asked when they acquired or disposed of land, or car drivers may be asked the time when they started their journey. Although the subjects of an Earth scientific investigation are unlikely to be capable of responding to such questions, this kind of date variable may still be determined. An example would be the dating of lake sediment cores by means of microfossils for the purpose of environmental reconstruction or perhaps more simply by counting the annual growth rings from cores bored in tree trunks.

Another type of measurement requiring special consideration is those recording financial values in a currency (e.g. US $, £ Sterling, € Euro, etc.), which are often needed in investigations concerned with people and institutions. In some respects a currency is just another unit of measurement like metres or litres, which are expressed to two decimal places. Many currencies are made up of units with 100 subunits, for example £8.52 (i.e. 8 whole pounds and 52 pence), which is notionally similar to 8.52 m (i.e. 8 whole metres and 52 centimetres). However, it is notoriously difficult to get accurate monetary figures of this type despite their obvious importance. Individuals and households may not only be reluctant to give details of their income and expenditure, but also may not know the precise figures. How many households in the UK could say **exactly** how much their total gross income was in the last six months? Or equally how many US corporations could report precisely how much they had spent on staff expenses in the last 3 months? Calculations based on rather general figures, such as to the nearest £1000 may produce results that are seemingly more accurate than they really are, for instance an average household income of £15 475.67.

Finally, before leaving our consideration of the characteristics of numerical data, the various spatial variables that are used in geographical and Earth scientific investigations such as distance, length, surface area, volume and so on may not only be analysed on their own, but also be combined with 'standard' nonspatial variables to create composite indicators, such as population per km^2, cubic metres of water discharged per second or point pollution sources per km of a river. Such indicators are similar to rates expressed as per 100 (percentages), 1000 or 1 000 000 insofar as they standardize data values according to some fixed unit, which makes comparison easier. It is simpler to compare population density in different countries or the concentration of chemicals dissolved in water, if the variables representing these concepts are expressed as relative values rather than absolute figures, even if the latter are collected as raw data in the first place. In other words, investigators should be prepared to analyse not only the raw data that they collect, but also additional variables that can be calculated from these, possibly in conjunction with some information obtained from another published source.

1.3 Simplifying mathematical notation

It is difficult to avoid the fact that quantitative analytical techniques depend on the calculation of mathematical values and on underlying statistical theory. It would no doubt please many students of Geography, Earth and Environmental Science if this were not the case, since it is likely to be such substantive topics and themes as those mentioned previously that drew them to these subjects rather than the prospect of encountering mathematical equations. Unfortunately, it is difficult to understand what these statistical techniques are doing, when they should be used without some understanding of their underlying mathematical basis and how they contribute to our understanding of various substantive topics. This inevitably raises the question of how to explain these principles without providing an excess of equations and formulae that might deter some readers from progressing beyond the first chapter and that others might choose to ignore. This text employs two strategies to try and overcome this problem: first, a simple basic principles introduction to the mathematical notation and calculation processes that will be encountered in later chapters; and secondly, wherever possible, restricting equations and formulae to Boxes to illustrate how the calculations work.

The notation used in mathematical equations can be thought of as a language, just as it is possible to translate from English to French, and French into German, it is also feasible to convert mathematical notation into English or any other language. The problem is that a typical equation expressed in prose will use many more words and characters than standard mathematical notation. For example:

$Y = \sum X^2 - \sum X$	translates as	the value of *Y* equals the result of adding together each value of *X* multiplied by itself and then subtracting the total produced by adding up all the values of *X*.
5 'words'		31 words

There is clearly a saving in terms of both words and characters (excluding spaces) between these two 'languages', with the equation requiring 5 'words' (8 characters) and the English translation 31 words (133 characters). The purpose of a language is to enable communication so that concepts and information can be conveyed from one person to another, although with the development of software capable of recognizing speech, this should perhaps be amended to say the transfer may also occur between people and computers. When taking notes in a lecture or from published literature, students will often develop their own abbreviated language to represent common words or phrases. For example, hydrology may be shortened to hyd. or government to gov., which are relatively easy to guess what they mean, but incr. and ↑ might be used interchangeably to denote increase. So a language is most useful when both parties understand what the symbols and abbreviations mean. Some mathematical and statistical notation has made the transition into everyday language. So books, newspapers and other media will often use %, for instance in respect to the national unemployment rate, to indicate a percentage. Many people will understand what the % symbol means, if not necessarily how the figure was calculated. However, much mathematical and statistical notation is like foreign language to many people.

The basic principle of a mathematical equation is that the value on the left side is produced once the calculations denoted on the right have been carried out. In one sense it does not really matter what numbers are substituted for the symbols on the right-hand side, since you can always perform the calculation and obtain a matching value on the left-hand side. In simple mathematics the numbers do not have any intrinsic meaning, you could just as easily derive 11 by adding 3 to 8, or subtracting 3 from 14. However, when using numerical data and applying statistical techniques in particular subject areas, such as Geography and the Earth Sciences, the numbers quantifying our attributes and variables have a meaning with respect to each of the different observation units in the set. If one of the variables in a survey of farms is land area, then each observation could have a different value for this variable. Variations in these values between the different farms actually have meaning in the sense that they convey something about farm size. In other words they are not just a series of numbers, but they signify something seemingly important, in this case that there is variability in the size of farms and some farms have more land than others.

The four sections in Box 1.2 provide a simple guide to explain how to read the different mathematical symbols used in later chapters of this text. First, some general symbols are presented to indicate labelling conventions for variables, attributes and statistical quantities. Secondly, the standard elementary mathematical operators indicate how to compute one number from another. Thirdly, mathematical functions specify particular computations to be carried out on a set of numbers. Finally, comparison or relational operators dictate how the one value on one side of an equation compares with that on the other.

The second strategy for introducing and explaining the mathematical background to the statistical techniques presented in this text is the use of Boxes. These are included to provide a succinct definition and explanation of the various techniques.

Box 1.2: Guide of mathematical notation and conventions and their demonstration in boxes.

Symbol	Meaning	Examples
General		
Decimal places	Number of digits to the right of the right of the decimal point	3 decimal places = 0.123, 5 decimal places 0.000 45
X, Y, Z	Attributes or variables relating to a statistical population	X represents the values of the variable distance to city centre measured in kilometres for each member of the statistical population
x, y, z	Attributes or variables relating to a statistical sample	x represents the values of the variable customers per day for each member of a sample
Greek letters	Each letter represents a different parameter calculated with respect to a statistical population	α (alpha), β (beta), μ (mu) and λ (lambda)
Arabic letters	Each letter represents a different statistic calculated with respect to a statistical sample	R^2, d and s
()	Brackets used, possibly in a hierarchical sequence to subdivide the calculation – work outwards from innermost calculations	$10 = 4 + (3 \times 2)$ $10 = 4 + ((1.5 \times 2) \times 2)$
Mathematical Operators		
+	Add	$10 = 4 + 6$
–	Subtract	$-2 = 4 - 6$
×	Multiply	$24 = 4 \times 6$
/or ———	Divide	$0.667 = 4/6$ $0.667 = \frac{4}{6}$
Mathematical Functions		
X^2	Multiply the values specified variable by themselves once – the square of X	$5^2 = 25.00$, $9.3^2 = 86.49$

Symbol	Meaning	Examples
X^3	Multiply the values of the specified variable by themselves thrice – the cube of *X*	$5^3 = 125.00$, $9.3^3 = 804.507$
X^{n}	Multiply the values of the specified variable by themselves n times – exponentiation of *X* to *n*	$5^5 = 3125.00$, $9.3^5 = 69\,568.837$
$\sqrt{X}$	Determine the number which, when multiplied by itself, will produce *X* – the square root of *X*	$\sqrt{25.00} = 5$, $\sqrt{86.49} = 9.3$
$\sum X$	Add together all the values of the specified variable	3 data values for *X* (4.6, 11.4, 6.1) $\sum X = 4.6 + 11.4 + 6.1 = 22.1$
$X!$	Factorial of the integer value *X*	$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$
$\lvert(X - Y)\rvert$	Take the absolute difference between the value of *X* and the value of *Y* (i.e. ignore any negative sign)	$\lvert (12 - 27) \rvert = 15$ (not −15)
Comparison or Relational Operators		
$=$	Equal to	$10 = 4 + 6$
$\neq$	Not equal to	$11 \neq 4 + 6$
$\approx$	Approximately equal to	$7 \approx 40/6$
$<$	Less than	$8 < 12$
$\leq$	Less than or equal to	$8 \leq 8$
$>$	Greater than	$5 > 3$
$\geq$	Greater than or equal to	$5 \geq 5$

They illustrate the calculations involved in these techniques by using a series of case-study datasets relating to various countries and to topics from Geography and the Earth Sciences. These datasets include attributes and variables referring to the spatial and nonspatial characteristics of the particular observation units. Box 1.3 provides an example of the general design features of the Boxes used in subsequent chapters, although the figures (data) are purely illustrative. Most Boxes are divided into different sections labelled (a), (b) and (c). If appropriate, the first of these comprises a diagram, map or photographic image to provide a geographical context for the case-study data. The second section is explanatory text describing the key features and outline interpretation of the particular statistical concept. The third section defines the statistical quantity in question in terms of its equation, tabulates the data values for the observation units and details the calculations required to produce a particular statistic (see example in Box 1.3c). Inevitably some of the Boxes vary from this standard format, and some of the datasets introduced in the early chapters are reused and possibly joined by 'companion' data in later chapters to enable comparisons to be made and to allow for techniques involving two or more attributes/variables (bivariate and multivariate techniques) to be covered. In addition to the Boxes and other text,

Box 1.3: Guide to conventions in demonstration in boxes.

Population symbol; sample symbol.

The symbols used to represent the statistical measure covered in the box will be given in the title section, if appropriate.

(a) Graphic image – photograph, map or other diagram
Boxes introducing some of the case study data will often include a graphic image to provide a context for the data and to help with visualizing the significance of the results.

(b) Explanatory text
This section provides a relatively short two or three paragraph explanation of how the particular statistical technique 'works', and a summary of how it should be calculated and interpreted.

(c) Calculation table
This section explains how to calculate particular statistical measure(s), in some instances comparing different methods. The first part provides the equation for the statistical quantity the elements of which are then used as column/row headings.

The following example, although based on the calculation of a particular statistic (the variance), is provided simply to illustrate the general format of these computation tables.

Identification of individual observation units	$s^2=\sqrt{\frac{n(\Sigma x^2)-(\Sigma x)^2}{n(n-1)}}$		Definitional equation
	x	x^2	Column headings: x = the values of variable X x^2 = the values of X squared
1	3	9	Arrows here indicate that mathematical operations applied to values in one column produce new values in another column
2	5	25	
3	6	36	Shading to lead the eye down a column of numbers to the calculations
$N = 3$	$\Sigma x = 14$	$\Sigma x^2 = 70$	Column totals
	$s^2=\frac{3(70)-(14^2)}{3(3-1)}$		Numbers substituted into equation
	$s^2 = 2.33$		Result
Sample mean	$\frac{\Sigma x}{n}=4.67$		Other related statistics

there are a series of self assessment or reflective questions scattered through the chapters. The aim of these is to encourage readers to reflect on the points discussed. These questions also provide the opportunity for tutors and students to discuss these issues in workshops or tutorials.

1.4 Introduction to case studies and structure of the book

The majority of this text is intended as an introduction to statistical techniques and their application in contemporary investigations of the type typically undertaken by undergraduate students during the final year of study in Geography and other Earth Sciences. Since the days of hand calculation, albeit using calculators has long since passed, the book assumes students will undertake statistical analysis in a digital environment (i.e. using computer software). However, it is not intended as a substitute for consulting online help or other texts that provide detailed instruction on how to carry out particular procedures in specific software (e.g. Bryman and Cramer, 1999).

The book has three main sections. The present chapter is in the *First Principles* section and has provided an introduction to quantitative analysis and outlined the features of numerical data and elementary mathematical notation. Also in this section Chapters 2 and 3 focus on collecting and accessing data for your analysis and considers the differences between statistical populations and samples, and strategies for acquiring data. Chapters 4 and 5 examine ways of reducing numerical data into manageable chunks either as summary measures or as frequency distributions. Chapter 5 also considers probability and how the outcome of events when expressed in numerical terms can be related to probability distributions. The *First Principles* section concludes with defining research questions and devising hypotheses accompanied by an introduction to inferential statistics, in other words statistical procedures from which you want to be able to reach some conclusion about your research questions. Separating the first and second sections is a pair of diagrams and a short series of questions that guide you through the decisions that need to be taken when carrying out your research investigation.

These diagrams refer to the statistical techniques covered by the chapters in the second and third sections. The second section, *Testing Times*, includes two chapters with a common structure. These, respectively, concentrate on parametric and nonparametric statistical tests and progress through situations in which the analysis focuses on one, two or three or more different attributes or variables. The final section, as its title suggests, *Forming Relationships*, goes beyond testing whether samples are significantly different from their parent populations or each other, and examines ways of expressing and measuring relationships between variables in terms of the nature, strength and direction of their mathematical connections. Examination of the explanatory and predictive roles of statistical analyses covers both bivariate and multivariate situations. Spatial and nonspatial analytical techniques are incorporated into the chapters in both of these sections and make use of a series of case-study datasets. The attributes and variables in these datasets are, by and large entirely genuine 'real' data values, although details that would possibly have permitted the identification of individual observations have been removed and in some cases the number of cases has been restricted in order to limit the Boxed displays. Nevertheless, the datasets can reasonably be regarded as illustrating different types of investigation that might be carried out by undergraduate students either during the course of their independent studies or during group-based field investigations.