Chapter 1

Why you need to use statistics in your research

This chapter explains the importance of statistics, and why you need to use statistics to analyse your data.

What is statistics?

Put simply, statistics is a range of procedures for gathering, organising, analysing and presenting quantitative data. 'Data' is the term for facts that have been obtained and subsequently recorded, and, for statisticians, 'data' usually refers to quantitative data that are numbers. Essentially therefore, statistics is a scientific approach to analysing numerical data in order to enable us to maximise our interpretation, understanding and use. This means that statistics helps us turn data into information; that is, data that have been interpreted, understood and are useful to the recipient. Put formally, for your project, statistics is the systematic collection and analysis of numerical data, in order to investigate or discover relationships among phenomena so as to explain, predict and control their occurrence. The possibility of confusion comes from the fact that not only is statistics the techniques used on quantitative data, but the same word is also used to refer to the numerical results from statistical analysis.

In very broad terms, statistics can be divided into two branches – descriptive and inferential statistics.

1. **Descriptive statistics** is concerned with quantitative data and the methods for describing them. ('Data' (facts) is the plural of 'datum' (a fact), and therefore always needs a plural verb.) This

branch of statistics is the one that you will already be familiar with because descriptive statistics are used in everyday life in areas such as government, healthcare, business, and sport.

2. **Inferential** (analytical) **statistics** makes inferences about populations (entire groups of people or firms) by analysing data gathered from samples (smaller subsets of the entire group), and deals with methods that enable a conclusion to be drawn from these data. (An inference is an assumption, supposition, deduction or possibility.) Inferential statistics starts with a **hypothesis** (a statement of, or a conjecture about, the relationship between two or more variables that you intend to study), and investigates whether the data are consistent with that hypothesis.

Because statistical processing requires mathematics, it is an area that is often approached with discomfort and anxiety, if not actual fear. Which is why this book tells you which statistics to use, why those statistics, and when to use them, and ignores the explanations (which are often expressed mathematically) of the formulae in which they tend to be articulated, though it does give advice on what you should bear in mind when planning your data collection.

One of the major problems any researcher faces is reducing complex situations or things to manageable formats in order to describe, explain or model them. This is where statistics comes in. Using appropriate statistics, you will be able to make sense of the large amount of data you have collected so that you can tell your research story coherently and with justification. Put concisely, statistics fills the crucial gap between information and knowledge.

A very brief history of statistics

The word ‘statistics’ derives from the modern Latin term *statisticum collegium* (council of state) and the Italian word *statista* (statesman or politician). ‘Statistics’ was used in 1584 for a person skilled in state affairs, having political knowledge, power or influence by Sir William Petty, a seventeenth-century polymath and statesman, used the phrase ’political arithmetic’ for ‘statistics’. (A book entitled *Sir William Petty, 1623–1687*, written by Lord Edmond Fitzmaurice, and published in

London in 1895, quotes Petty as saying that 'By political arithmetic, we mean the art of reasoning by figures upon things relating to government'.) By 1787, 'statistic' (in the singular), meant the science relating to the branch of political science dealing with the collection, classification and discussion of facts bearing on the condition of a state or a community.

'Statists' were specialists in those aspects of running a state which were particularly related to numbers. This encompassed the tax liabilities of the citizens as well as the state's potential for raising armies. The word 'statistics' is possibly the descendant of the word 'statist'.

By 1837, statistics had moved into many areas beyond government. Statistics, used in the plural, were (and are) defined as numerical facts (data) collected and classified in systematic ways. In current use, statistics is the area of study that aims to collect and arrange numerical data, whether relating to human affairs or to natural phenomena.

The importance of statistics

It is obvious that society can't be run effectively on the basis of hunches or trial and error, and that in business and economics much depends on the correct analysis of numerical information. Decisions based on data will provide better results than those based on intuition or gut feelings.

What applies to this wider world applies to undertaking research into the wider world. And learning to use statistics in your studies will have a wider benefit than helping you towards a qualification. Once you have mastered the language and some of the techniques in order to make sense of your investigation, you will have supplied yourself with a knowledge and understanding that will enable you to cope with the information you will encounter in your everyday life. Statistical thinking permeates all social interaction. For example, take these statements:

- 'The earlier you start thinking about the topic of your research project, the more likely it is that you will produce good work.'

- 'You will get more reliable information about that from a refereed academic journal than a newspaper.'
- 'On average, my journey to work takes 1 hour and 40 minutes.'
- 'More people are wealthier now than ten years ago.'

Or these questions:

- 'Which university should I go to?'
- 'Should I buy a new car or a second-hand one?'
- 'Should the company buy this building or just rent it?'
- 'Should we invest now or wait till the new financial year?'
- 'When should we launch our new product?'

All of these require decisions to be made, all have costs and benefits (either financial or emotional), all are based upon different amounts of data, and all involve or necessitate some kind of statistical calculation. This is where an understanding of statistics and knowledge of statistical techniques will come in handy.

Why you need to use statistics

Much of everyday life depends on making forecasts, and business can't progress without being able to audit change or plan action. In your research, you may be looking at areas such as purchasing, production, capital investment, long-term development, quality control, human resource development, recruitment and selection, marketing, credit risk assessment or financial forecasts or others.

And that is why the informed use of statistics is of direct importance to you while you are collecting your data and analysing them. If nothing else, your results and findings will be more accurate, more believable and, consequently, more useful.

Some of the reasons why you will be using statistics to analyse your data are the same reasons why you are doing the research. Ignoring the possibility that you are researching because the project or dissertation element of your qualification is compulsory, rather than because you very much want to find something out, you are likely to be researching because you want to:

- measure things;
- examine relationships;
- make predictions;
- test hypotheses;
- construct concepts and develop theories;
- explore issues;
- explain activities or attitudes;
- describe what is happening;
- present information;
- make comparisons to find similarities and differences;
- draw conclusions about populations based only on sample results.

If you didn't want to do at least one of these things, there would be no point to doing your research at all.

What statistical language actually means

Like other academic disciplines, statistics uses words in a different way than they are used in everyday language. You will find a fuller list of the words you need to understand and use in the Glossary.

Variable and constant

In everyday language, something is variable if it has a tendency to change. In statistical language, any attribute, trait or characteristic that can have more than one value is called a **variable**.

In everyday language, something that does not change is said to be constant. In statistical language, an attribute, trait or characteristic that only has one value is a **constant**. Confusingly, something may be a variable in one context and a constant in another. For example, if you are looking at the spending patterns of a number of households, the number of children (which will vary) in a particular household is a variable, because we are likely to want to know how household spending depends on the number of children. But, if you are looking at the spending patterns of households which have, say, three children, then the number of children is a constant.

Strictly speaking, in statistical language, when your variables and constants are categorical (we will discuss this more in Chapter 2), for example, eye colour or nationality, they are known as **attributes**.

Discrete and continuous

Quantitative variables are divided into 'discrete' and 'continuous'. A **discrete** variable is one that can only take certain values, which are clearly separated from one another – for instance, a sales department can have 2 or 15 or 30 people within it. It cannot, however, contain 3⅔ or 48.1 people. A **continuous** variable is one that could take any value in an interval. Examples of continuous variables include body mass, height, age, weight or temperature. Where continuous variables are concerned, whatever two values you mention, it is always possible to have more values (in the interval) between them. An example of this is height – a child may be 1.21 metres tall when measured on 27th September this year, and 1.27 metres on 27 September next year. In the intervening 12 months, however, the child will have been not just 1.22 or 1.23 or 1.24 and so on up to 1.27 metres, but will have been all the measurements possible, however small they might be, between 1.21 and 1.27.

Sometimes the distinction between discrete and continuous is less clear. An example of this is a person's age, which could be discrete (the stated age at a particular time, 42 in 2007) or continuous, because there are many possible values between the age today (42 years, 7 weeks and 3 days) and the age next week (42 years, 8 weeks and 3 days).

Cardinal and ordinal

We will deal with cardinal and ordinal numbers later in Chapter 2, but here we want to highlight that a 'cardinal' in statistics is not a person with high-rank in the Roman Catholic church. **Cardinal** numbers are 1, 2, 3 and so on, and they can be added, subtracted, multiplied and divided. An **ordinal** number describes position 1st, 2nd, 3rd and so on), and they express order or ranking, and can't be added, subtracted, multiplied or divided. Most of the statistical techniques created for the analysis of quantitative are not applicable to ordinal data. It is therefore meaningless (and misleading) to use these statistical techniques on rankings.

Population and sample

In statistics, the term ‘population’ has a much wider meaning than in everyday language. The complete set of people or things that is of interest to you in its own right (and not because the collection may be representative of something larger) is a **population**. The number of items, known as **cases**, in such a collection is its size. For example, if you are interested in all the passengers on a particular plane in their own right and not as representatives of the passengers using the airline which owns that particular plane, then those particular plane passengers are your population.

But if you do a statistical analysis of those particular plane passengers in order to reach some conclusion about, say, (1) all plane passengers heading to that destination, or (2) all plane passengers on any route on that day and at that time, then the passengers are a sample. They are being used to indicate something about the population (1) or (2). A **sample** is therefore a smaller group of people or things selected from the complete set (the population).

It hardly goes without saying that you need to be clear about whether your data are your population or a sample. Most of statistics concerns using sample data to make statements about the population from which the sample comes.

Misuses of statistics

Statistics consists of tests used to analyse data. You have decided what your research question is, which group or groups you want to study, how those groups should be put together or divided, which variables you want to focus on, and what are the best ways to categorise and measure them. This gives you full control of your study, and you can manipulate it as you wish. Statistical tests provide you with a framework within which you can pursue your research questions. But such tests can be misused, either by accident or design, and this can result in potential misinterpretation and misrepresentation. You could, for instance, decide to:

- alter your scales to change the distribution of your data;
- ignore or remove high or low scores which you consider to be inconvenient so that your data can be presented more coherently;
- focus on certain variables and exclude others;
- present correlation (the relationship between two variables, for example, height and weight – the taller people in the sample are thinner than the shorter people) as causation (tallness results in or is a cause of thinness).

It goes without saying that, because research is based on trust, you must undertake your research in an ethical manner, and present your findings truthfully. Deliberately misusing your statistics is inexcusable and unacceptable, and if it is discovered by your supervisor or examiner, retribution will be severe.

Because you are inexperienced in research, the main errors which you might make are bias, using inappropriate tests, making improper inferences, and assuming you have causation from correlations.

Bias

In ordinary language, the term 'bias' refers simply to prejudice. It could be that when the data you are using were collected, the respondents were prejudiced in their responses. You might get this kind of thing if you are eliciting attitudes or opinions. In statistical language, **bias** refers to any systematic error resulting from the collection procedures you used. For example, in a questionnaire, if the non-respondents (those who haven't answered the questionnaire) are composed of, say, a large percentage of a higher socio-economic group, it could introduce bias (systematic error) because you would have an under-representation of that group in your study. Often the people with the strongest opinions, or those who have a greater interest in the results of the research, who may derive some benefit from the results, or who have a loyalty or allegiance to express, are more likely to respond to the questionnaire than those without those views or interests. There are procedures that can deal with nonresponse in questionnaires and interviews. It would benefit your research if you read up about these and included them in your research design if you are collecting data specifically for your research project (**primary data**) rather than reanalysing data that have already been collected for some other purpose (**secondary data**).

Using inappropriate tests

We'll come back to this, but here we need to warn you that one of the ways in which to misuse statistics is to use the wrong tests on your data. All statistics textbooks will tell you that non-parametric tests are to be used on nominal and ordinal variables (we'll explain these terms more fully in Chapter 2) and that parametric tests are reserved for interval and ratio variables. You will find, however, that researchers, who should know better, use parametric tests on ordinal variables. But now you know better, and you won't do that.

Improper inferences

Much of statistical reasoning involves inferences about populations from data observed in samples. The reasoning may be **inductive**, in other words, reasoning from the particular (the sample) to the general (the population). However, to avoid improper inferences, you'll need to define the population carefully and use an appropriate probability sampling technique.

Concluding causation from correlations

It is a great temptation to conclude that because two factors are correlated (co related); one of these factors caused the variations in the other. You need to be careful not to fall into this trap and not to try to draw cause-and-effect conclusions from statistical data concerning correlated factors. For example, just because sales of coal are higher when the temperature is higher does not mean these sales are caused by an increase in temperature. The real reason is that, when temperature is higher, the price of coal is reduced, resulting in more coal being purchased.

The advice about avoiding these four errors is that you should question every stage of your statistical investigation, from the design of your project, through the collection and analysis of your data, to the presentation of your findings.

Data collection

We will consider issues of primary data collection in Chapter 3 when we discuss why you need to link the creation of your questionnaire or interview to your analysis plans. It is sufficient to point out here that how you collect your primary data and how you make sense of what you have collected in order to come up with credible results are not so much connected as intertwined.

You will avoid a great deal of anxiety and anguish if you undertake the planning of your analysis at the same time as you undertake the planning of your data requirements. It will prevent a number of potential difficulties such as selecting secondary data that will not enable you to answer your questions fully or, in the case of primary data, asking the wrong questions, or asking the right questions in the wrong way, or leaving out questions you should have asked. In addition, it should help you avoid finding that you don't know how to analyse or interpret the data you have got!