2ND EDITION

# BEGINNING STATISTICS

## AN INTRODUCTION FOR SOCIAL SCIENTISTS

LIAM FOSTER
IAN DIAMOND
JULIE JEFFERIES

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

# ONE

# INTRODUCTION – ARE STATISTICS RELEVANT TO REAL LIFE?

## Introduction

The first edition of *Beginning Statistics* by Ian Diamond and Julie Jefferies stemmed from a belief that there were some excellent general statistics texts available at the time but none of them had the right tone or coverage of social statistics. Widely used across social science courses in particular, the first edition received much positive feedback. Despite its success, it is evident that it was becoming dated given that it is now over 10 years old and, as such, a number of the examples and features required updating. The decision to write the second edition, with the assistance of Liam Foster, also comes at a time when we are witnessing an increasing emphasis on the need for social science students to possess statistical research skills. This new edition provides an opportunity to bring data sources up to date, to clarify and add some explanations, provide chapter learning objectives and conclusions and a glossary. It also contains an expanded introduction, new sections on sampling and presentation conventions, and a conclusion. Despite these changes we still start with the basic assumption that you may have done no or limited previous work with statistics or may require a refresher. Maths may also be a distant (and painful) memory. By making use of interesting examples (well, we think so!) from the social sciences this book introduces a number of statistical approaches and techniques which will be invaluable to the development of your statistical skills. It does this in a systematic way, taking you through the steps required to use statistics in the social sciences, from introducing data and their presentation, to correlation and regression.

## Did you know you already use statistics?

Believe it or not, you can *already* think **statistically** – in fact, it is something you do all the time. If you don't believe this, think about the number of times you have thought about the average number of hours you are studying or working in a week, how many times you have decided to take a coat with you because it is likely to be cold this time of year, or noted the fact that you tend to do better when you start revising for your exams earlier! In each case, you have acted on the basis of statistical concepts and used data that you had stored in your brain. You just didn't realise it! You are making a statistical statement even though you are performing at best rough calculations. In the first example you are summarising data; in the second and third you are generalising from previous experiences of weather patterns and exam performance to make a prediction or statement.

So our observations often involve, for instance, the process of counting how many times something has occurred, and measuring the length of time since something has happened. As if, by instinct, we look for patterns and connections among the things we notice, often in a rather imprecise manner. This might be something as simple as whether you are more likely to score a goal when wearing lucky pants or have a more successful date when you go for a meal, bowling or ice-skating! We constantly question data and use them to influence our decisions and behaviour. This is where statistics are crucial. They are a way of making sense of our observations.

Learning about statistics helps you to look for reliable patterns and associations in both the short and long term (Rowntree, 2003). It also teaches us caution in expecting these to hold true in all situations (for instance, unfortunately Liam doesn't always score a goal with his lucky pants on). There are often important limitations to data which need to be considered, including whether they are biased, unrepresentative or totally meaningless (it does happen!). It teaches us to think critically about our techniques, samples and claims we make. Statistics is basically about understanding and knowing how to use **data**. So we all use statistical concepts intuitively in our daily lives. Learning statistics is simply a case of learning how to express things more accurately. By using statistical concepts this enables us to summarise and predict more precisely than we normally would in our everyday observations.

## How are statistics used?

Data provide information which governments and organisations use to make policy decisions (and to evaluate the effectiveness of existing policies). Statistics about institutions such as schools and hospitals are increasingly collected and made available to the general public. This may be useful in deciding where a child is going to have the best chance of getting their A levels, or where you are least likely to die in the operating theatre. Such league tables can increase transparency and choice, but they may also be seen to fulfil a political purpose.

This century has been called the century of 'big data', and the techniques in this book are ever more important in this context. 'Big data' refers to every piece of knowledge that has or will be digitalised and stored on a computer hard drive, a database, or in the 'cloud'. It is vital to millions of companies such as Facebook, Google and Twitter, as well as Tesco, Nectar and Amazon, which harvest user information, making use of big data in targeted sales and advertising. For instance, if you have been looking for a nice pair of walking boots you are more likely to receive adverts about tents or camping holidays because statistically those who buy walking boots are more likely to go camping than those who don't. Simple stuff, I know, in an example such as this, but statistics are being used to look at these kinds of relationships in buying behaviour. So it is not only governments that use data to monitor and analyse behaviour.

There are two main kinds of statistics that we are going to concentrate on in this book. **Descriptive statistics** is a set of methods used to describe data and their characteristics. For example, if you were investigating the number of visitors to a beach in August (nice job if you can get it!), you might draw a graph to see how the number of visitors varied each day, work out how many people visit on an average day and calculate the proportion of visitors who were male/female or children/adults. These would all be descriptive data.

**Inferential statistics** involves using what we know to make inferences (estimates or predictions) about what we don't know. For example, if we asked 200 people who they were going to vote for on the day before a local election we could try to predict which party would win the election. Or if we asked 50 injecting drug users whether they share injecting equipment such as needles with other users, we could try to estimate the proportion of all injecting drug users who share equipment.

We would never be able to say for sure who would definitely win the election or what proportion of injecting drug users share equipment, but we *are* able to predict the *likely* outcome or proportion. Statistics is all about weighing up the chances of something happening or being true. Statistics *are* relevant to real life because without real life we wouldn't need statistics. If everybody always voted for the same party and all injecting drug users shared equipment, we wouldn't need to predict the outcome of an election or estimate the proportion of drug users sharing equipment because it would be obvious from asking one voter or one injecting drug user. Only in a world of clones would statistics about people be unnecessary! Life would be pretty boring if everybody were exactly the same.

So in real life everybody is different and, in social science, statistics are frequently used to highlight the differences between groups of people or places. For example, we might want to investigate how smoking behaviour varies by socio-economic group or how unemployment varies by local authority. Knowledge of statistical methods is crucial for answering many research questions like these.

## The emergence of statistics

We know that statistics are commonly used in contemporary society, but where did they come from? The use of statistics is nothing new. It goes back to at least the earliest city states. For instance, Babylonians and Egyptians collected numerical data on crops and growing conditions. The word 'statistics' is derived from the Latin term for 'state' or 'government'. In the UK and the Westernised world it was particularly the birth of industrialism that led to an interest in social data. In Europe and the USA censuses started to be taken in the eighteenth century, with the first in the UK in 1801. In the UK a census has been conducted every 10 years (with the exception of 1941 as a result of the Second World War).

Sapsford (2007), commenting on the history of statistics in the UK, stated that, for early Victorians, the role of statisticians was to collect information about people in the emerging capitalist societies. It was thought that we needed to 'map' the human population to make the best use of people in industry as well as providing the services they needed. It was important to know about things like where people lived, their ages (including how many were able to work), what children were living in the family, whether there were many older or disabled people who needed support, and what types of housing were available.

The potential of these kinds of data for social scientists was shown by people such as Charles Booth (1886), whose work on occupation patterns was derived from analysis of the 1801–1881 UK censuses. The first journal of the Royal Statistical Society in 1838 had lots of articles describing the social conditions of the time. Large-scale social surveys looking at urban poverty were pioneered by Seebohm Rowntree (1901) in York and Charles Booth (1902) in London at the turn of the twentieth century, as statistics started to play a greater role in exploring social problems. Following the Second World War, statistical methods have enjoyed increasing popularity in the social sciences. There has also been an increase in the number of large scale national datasets available, covering a wealth of areas such as health, crime, employment, housing and social attitudes. These have been used by academics and policy-makers alike.

This increase in the number of datasets available, both national and international, longitudinal and cross-sectional, has been accompanied by advances in technology, in particular the influx of computers, especially from the 1980s and 1990s. With this came further access to data analysis for a larger number of people and further possibilities in its use. Specific computer programmes such as IBM SPSS, SAS, STATA and R have been developed which make the analysis of statistical data more accessible to social scientists. Analysis is now possible at the click of a button (although you need to know which buttons to press and what the results mean!). Statistical tests have continued to be developed enabling higher levels of analysis. This has been supported by the UK Data Service (formerly the UK Data Archive), which helps researchers to acquire data and supports their use. Despite this, unfortunately it is not uncommon for statistics courses taken by students in the social sciences to be treated essentially as maths courses with examples used which do not relate to the social sciences. This can be off-putting to those learning statistics and something this book is mindful of.

## Do we really need to know about statistics?

Basically, if data are to be useful, they have to be processed and analysed. This requires statistical skills, which will become ever more important in the social sciences and beyond. Many of the concepts that underlie statistical analysis are not particularly complex and are things which people do on a daily basis, as we have already shown. The majority of people have some understanding of figures (even if they don't admit it). They know a 10% pay rise is better than a 5% one or can work out the odds on the horse they bet on winning. This will stand you in good stead for this book and for developing your statistical capabilities.

There are two major reasons why learning about statistics will be useful to you:

- You are constantly exposed to statistics every day of your life, as you have already seen. Marketing surveys, voting polls and findings from social research appear in daily newspapers and popular magazines. By learning about statistics you will (hopefully) become a more effective consumer of statistical information. For example, if a hair advert tells you that 75% of participants say their hair is silky and smooth following the use of a particular shampoo you would start to question what they were using before (or not using before!) and think about the sample size (if the small print says a sample of 10 you might be a bit worried!).
- You need to be able to understand and interpret statistics at university or in the workplace. Even if conducting research is not part of your job or you don't do a quantitative project at university, you will be expected to understand and learn from other people's research based on statistical analysis.

Modern society is driven by statistics which frequently influence our behaviour, from reviews on TripAdvisor which tell us when a place really is a dive which we need to avoid like the plague, to crime statistics which tell us which areas we are most likely to be burgled in (and might not want to move to if we have a choice in the matter!). Even if you never go on to do research, a good grasp of statistics will help you to understand the figures that you read or hear about and to avoid being misled by people who (mis)use statistics to their own advantage.

The need to promote statistical skills among social scientists is perhaps greater than ever. Recent discussion has recognised that many graduates in the social sciences lack the quantitative skills required by the social research industry and that quantitative skills are in demand by employers (Wiles et al., 2009). This has led to lots of institutions working to improve their quantitative research teaching and research. This has also been encouraged by Q-Step, a £19.5 million programme funded by the Nuffield Foundation, ESRC and HEFC to set up 15 quantitative research centres at UK universities designed to promote a step-change in quantitative social science training.

## 'There are three kinds of lies: lies, damned lies and statistics'

The expansion of data available has also led to increasing debate about how figures are constructed. While numbers are often thought of as hard facts, they are actually the result of different decisions about how something should be categorised or counted. There is much cynicism about statistics and how they are used. Mark Twain reported that the British statesman Benjamin Disraeli once said, 'there are three kinds of lies: lies, damned lies and statistics'. More recently the postmodern theorist Jean Baudrillard (1990: 147) said, ‘like dreams, statistics are a form of wish fulfilment’. However, to reject statistics would involve blinding oneself to much important information.

It is crucial to be aware that statistical data can be misinterpreted, distorted or selected to serve particular ends. This is not the inherent fault of statistics per se; rather the fault of analysis which does not carefully examine the logic of an argument and how data support this (Dietz and Kalof, 2009). Not all inaccurate use of statistics is deliberate (not everybody has sinister motives – I can assure you!). Things can go wrong when figures are written down inaccurately, files not backed up, research poorly designed and inappropriate statistical tests used. Sometimes it is how the research is reported that is problematic. For instance, it has been known for the media to report statistics in a way that comes across as misleading or politicians to be very selective about the figures they present.

It is also common for inaccurate conclusions about findings to be made (and bits that don’t fit left out!) without looking at whether it actually means what is stated or whether there could be other possible explanations! So, for instance, several years ago research by Halpern and Coren (1991) found that the average age of death of left-handers was about 8 years less than for right-handers. Data were collected by using a sample of 987 deaths in Southern California (where lists are published of everyone who has died) and interviewing relatives or friends to find out the hand they used most often. As a left-handed person, Liam was particularly interested (and worried) about this research and had a closer look at it. The authors of the study speculated why left-handed people may be more likely to die at a younger age, but Strang (1991) stated that they failed to sufficiently account for possible bias given that the frequency of left-handedness is likely to have changed appreciably over time. For instance, if many people born 60 or more years ago (such as Liam’s mum) were forced to use their right hands, then this would result in a large number of right-handed people being reported as dying in older age, with the probability of being left-handed greater among younger people, resulting in a lower average age of death for left-handed people. Whether this explains the eight year difference is another question! Perhaps a bigger issue is that the study was based on a sample of deaths and didn’t actually measure the chance of left or right handers in the population dying at different ages.

Another example commonly referred to is the idea that children who eat breakfast (not a chocolate bar from the local corner shop though) perform better at school in general than those who don’t eat their bowl of Weetabix or turn the milk chocolatey with their Coco Pops. While it is true that statistics do indicate a correlation between having breakfast and academic achievement (leading to parents making their children hear that snap, crackle and pop in the morning), when the trend was investigated, although food can aid concentration, it was found that it was largely the types of children who have things going on in their lives which stop them from eating breakfast who are also more likely to struggle at school.

A second problematic example involving children (Levitas, 2012) is the coalition government’s ‘Troubled Families’ initiative, targeted at those families who ‘fail to take responsibility for their own lives’. Current government policy on social justice claims that there are 120,000 ‘troubled families’ in Britain. The Department for Communities and Local Government (DCLG, 2013) identifies the research the figure is based upon (though not the details of the costing). The original research is a report carried out for the Social Exclusion Task Force in 2007 using secondary analysis of the Family and Children Study, a longitudinal survey. It focused on families with troubles rather than 'troubled families'. This analysis showed that in 2004 about 2% of the families in the survey had five or more of seven characteristics, and were severely multiply disadvantaged. The characteristics were: no parent in the family is in work; the family lives in overcrowded housing; no parent has any qualifications; the mother has mental health problems; at least one parent has a long-standing limiting illness, disability or infirmity; the family has low income (below 60% of median income); and the family cannot afford a number of food and clothing items. That 2% of families generated an estimate of 140,000 for Britain, later calculated as 117,000 for England, rounded to 120,000. The DCLG website makes a jump from families that have troubles, through families that are 'troubled', to families that are or cause trouble (Levitas, 2012). Portes (2012) points out that none of these criteria, in themselves, have anything at all to do with disruption, irresponsibility, or crime. While a family meeting five criteria is likely to be disadvantaged and poor, are the criteria identified a source of wider social problems? At another point the DCLG (2012) stated that these families are characterised by being involved in crime and anti-social behaviour; have children not in school; have an adult on out-of-work benefits and cause high costs to the public purse, but still used the same figure of 120,000 despite it being based on data from 2004 and changing in definition! It is highly unlikely that the figures remain exactly the same (see Levitas, 2012). Therefore, these figures should be used with much caution (if at all)!

A final example that we like comes from Huck and Sandler (1979) and relates to a billboard advertising campaign and Miss America. In an attempt to prove that billboard advertising is the best form of advertising, the Institute of Outdoor Advertising conducted a research study. Using 10,000 billboard panels, they placed a poster showing a large picture of Miss America with her crown and the simple message, 'Shirley Cothran, Miss America, 1975'. Before they did this a series of studies were conducted to determine public awareness of Miss America's name prior to it going on the billboards, with a random sample of over 15,000 adults questioned in 1975. Despite previous exposure which Miss America had received on TV and radio and in print, only 1.6% of respondents gave the correct answer when asked, 'what is the name of Miss America 1975?' The billboards then went up and two months later a second wave of over 15,000 interviews was conducted by the same research teams (with different participants). This time, 16.3% of the respondents knew who Miss America was. This was said to indicate that the billboard advertising had a very positive effect on promoting the name of Miss America and was therefore a highly effective form of advertising. However, could there be any other reasons for this trend? Well, first, the study fails to acknowledge that other forms of media also existed during the time the billboards were up. Second, the unique nature of the billboard study led other forms of media to cover the story of the research, with Miss America's name discussed more than usual in the media. So even the results from a simple statistical question like the one asked here can be difficult to interpret and require some critical thinking, something this book is going to help you with. So an awareness of statistics helps to identify these types of concerns and mistakes, and many more like them!

## Overview of the chapters

Statistics is rather like building a house (but not as physically demanding). It involves laying the foundations first, establishing the basics of using and presenting statistics, before more advanced statistics can be used (or the rest of the house can be built). Without this secure foundation and knowledge of descriptive statistics it is difficult to know what inferential statistics are appropriate, and this is likely to result in inaccurate statistics to interpret. Similarly, if you were building a house and the basic foundations were not established you risk the house falling down (and getting sued if it caused other damage!). This is where the book starts, gradually moving from simple topics to the more complicated ones.

Chapters 2 and 3 focus on how data are measured, or how the levels of measurement affect how they can and should be presented, with a particular focus on different forms of tables and graphs. The correct type of table or graph to use depends on the type of data you have, not just because it looks pretty! There are a number of rules or conventions you need to be familiar with in the design and presentation of tables and graphs which are covered in these chapters.

Chapters 4 and 5 move on to look at ways of describing data. Rather than concentrating on investigating the distribution of data by drawing graphs and tables, these chapters explore how to describe a dataset statistically, so that we can summarise the features of a distribution (descriptive statistics). They focus on measures of central tendency including the mean, median and mode, and calculating and presenting percentiles, terms you will become familiar with. These are useful in helping us to identify key patterns such as the most common response to a question or the average number of times an event occurs over a period of time.

Sometimes in order to use data more effectively it is necessary to transform them. Chapter 6 explores how and when this can be useful. It may be something as simple as standardising measurements when you are provided with information in both miles and kilometres and you want to undertake research on distance travelled to work. Transforming variables can be a useful thing to do when dealing with statistics. This includes the process of standardising the data and producing *Z*-scores.

Chapter 7 introduces the process of calculating whether data are normally distributed and, building on the previous chapter, how measures of standard deviation and *Z*-scores are used to do this. Inferential statistics are used as we show how we can predict the probability of an event occurring when data are normally distributed. In Chapter 8 the focus turns to sampling and how to select a representative sample in order to make estimates of the things we are trying to find out about in the population. Using a sample can save vast amounts of time and money. This chapter highlights different forms of sampling and when they might be used. It shows how it is not always as easy as it sounds and takes you through strategies to help with representativeness.

Chapters 9 and 10 continue to focus on the use of samples and their relationship with populations, using confidence intervals. However, Chapter 9 uses continuous data whereas Chapter 10 deals with proportions.

Chapters 11 and 12 provide a brief introduction to hypothesis testing in various contexts. This involves using confidence intervals and *Z*-scores, when you have sample means or proportions, to test hypotheses (in Chapter 11). It also shows the difference between two-sided (or two-tailed) tests and one-sided (or one-tailed) tests in calculating whether a hypothesis should be proved or disproved. Chapter 12 introduces the *t* test, used to assess the statistical significance of the difference between the means of two sets of scores and whether the average score for one set of scores differs significantly from the average score for another set of scores. It also provides details about when Z-scores, one-sample *t* tests and one-sample sign tests in particular should be used.

The process of comparing two variables with each other is the main focus of Chapter 13. For instance, is a student's performance at university related to how much paid work they are doing? To assist in answering this kind of question, correlation can be used to measure the association between two continuous variables. Regression takes things a little further, enabling us to predict the values of one variable from the values of another variable. For instance, the data on the number of qualified doctors in countries per 1000 people could be used to try to predict the death rate. It is also possible to use inferential statistics to analyse categorical data such as using the chi-square test, which allows us to test the association between two variables in which the expected values are compared with the observed values. These processes are shown in Chapter 14. Chapter 15 provides you with more information about how and where to develop your statistical skills further, something we hope you will want to do when you have completed this book.