PART 1

# Elementary statistical analysis

CHAPTER 1

# Introduction

## 1.1 Aims of the book

This text has three principal objectives. The first is to provide an elementary and very informal introduction to the fundamental concepts and techniques of modern quantitative methods. A primer cannot be comprehensive, but we will cover many of the procedures most widely used in research in the historical and social sciences. The book is deliberately written at a very basic level. It does not include any statistical theory or mathematics, and there is no attempt to prove any of the statistical propositions. It has been planned on the assumption that those reading it have no retained knowledge of statistics, and very little of mathematics beyond simple arithmetic.

It is assumed that the material in the book will be taught in conjunction with one of the several statistical packages now available for use with computers, for example, *SPSS for Windows, STATA, MINITAB*, or *SAS*. By using the computer to perform all the relevant statistical calculations and manipulations it is possible to eliminate both the need to learn numerous formulae, and also the tedious work of doing laborious calculations. However, if the computer is going to provide the results, then it is absolutely essential that the student should be able to understand and interpret the content and terminology of the printouts, of which figure 1.1 is a typical specimen, and the second objective of the book is to achieve this.

This leads naturally to the third and most important objective. The book is throughout concerned to relate the quantitative techniques studied to examples of their use by historians and social scientists, and by doing this to promote the understanding and use of these methods. In the following section we introduce four specific studies that will be deployed throughout the book, but – at appropriate points in the text – we also refer readers to other examples of the application of quantitative methods to historical and other issues. A student who studies this text will not be able

**Variables Entered/Removed[b]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | FARMERS, CHILDALL, LONDON, WEALTH, GRAIN[a] | | Enter |

a. All requested variables entered.

b. Dependent Variable: RELIEF

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .532[a] | .283 | .271 | 6.84746 |

a. Predictors: (Constant), FARMERS, CHILDALL, LONDON, WEALTH, GRAIN

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 5638.346 | 5 | 1127.669 | 24.050 | .000[a] |
| | Residual | 14300.769 | 305 | 46.888 | | |
| | Total | 19939.116 | 310 | | | |

a. Predictors: (Constant), FARMERS, CHILDALL, LONDON, WEALTH, GRAIN

b. Dependent Variable: RELIEF

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 11.352 | 1.718 | | 6.609 | .000 |
| | GRAIN | .301 | .089 | .204 | 3.363 | .001 |
| | CHILDALL | 5.609 | 1.030 | .271 | 5.445 | .000 |
| | WEALTH | -.276 | .179 | -.081 | -1.544 | .124 |
| | LONDON | -4.156E-02 | .010 | -.234 | -4.110 | .000 |
| | FARMERS | 5.873 | 2.131 | .158 | 2.756 | .006 |

a. Dependent Variable: RELIEF

**Figure 1.1** Illustration of SPSS print-out of regression results

to claim that she is a fully qualified statistician. However, she should have the confidence to read chapters or articles which use quantitative methods, to understand what the authors have done and why they have done it, and to make her own critical evaluation of the procedures used and the historical conclusions drawn from the statistical results.

Students should also be able to see from the case studies and other examples how the use of quantitative methods can open up new aspects of an enquiry and can supplement and strengthen other methods of research. We hope that they might then appreciate how their own research projects might benefit from the application of these methods, and take their own first steps in this direction.

The book is designed to be used both as the basic text for taught courses and for students working on their own without an instructor. In planning the content and sequence of the chapters, one of our primary considerations has been to keep the material in the early chapters at a very elementary level. Many of those for whom this book is intended will naturally be somewhat wary about taking a course in quantitative methods, but if they find that they can make substantial progress in understanding some of the basic statistical concepts they will gain confidence in their ability to handle the slightly more difficult material in later chapters. There is a small price to be paid for this approach, since it means that correlation and regression are covered twice: first without any statistical theory (in chapters 3 and 4) and then again in greater depth (in chapter 8). However, the text has been used for a number of years to teach a class taken almost exclusively by statistical novices who initially approached the course with deep suspicion, and experience has shown that this strategy is very successful.

It is, of course, also possible for instructors to follow the material in a different sequence, or – depending on the time available and the level it is desired to achieve – to omit certain topics altogether. For example, chapter 7 on non-parametric methods has been included because these procedures are often appropriate for the sort of problems faced by historians; and it has been placed at this point in the text because it makes a useful complement to the discussion of the standard principles of hypothesis testing in chapter 6. However, a course designed primarily to provide a very basic introduction to regression analysis in 10 sessions might skip this. It might, for example, start with chapters 2–6 and 8–9, add material on dummy variables from chapter 10 and on the basic aspects of non-linear regression from chapter 12, and then cover the appropriate applications in the case studies in chapters 14 and 15.

This would be sufficient to give students a good grounding in some of the main aspects of regression methods, and should enable them to cope

with many examples of the use of these methods in the historical literature. However, the omission of chapter 11 would mean that they would not have acquired any knowledge of either the serious problems which can arise when the assumptions underlying the standard regression model (discussed in chapter 9) are violated, or the associated procedures for diagnosing and – where possible – correcting for these violations. This would also be a substantial weakness for any students who wished to apply these methods to their own research. One alternative, for students able to start with some knowledge of very elementary statistics and wishing to aim a little higher (while still limited to 10 sessions), would be to skip chapters 1–4 and 7, and work through chapters 5–6 and 8–15.

The normal text and tables are supplemented by material in boxes and panels. *Boxes* are used to highlight the fundamental definitions and concepts, and should be studied closely. *Panels* are used to provide explanations or information at a slightly more advanced level than the rest of the text. The panels should be helpful for some readers, but those who omit them will not be at any disadvantage in understanding the remainder of the text. In addition to footnotes (referred to by a letter), we have also used endnotes (referred to by a number) where it seems desirable not to burden the main text with lengthy annotations. Endnotes typically consist either of lists of references to further examples of the applications of statistical methods to historical topics, or of technical points which need not distract all readers although they may be useful to some.

We have given a formula for all the basic concepts even though the book is written on the assumption that the computers will provide what is required for any particular statistical operation. It is usually possible to set out these formulae in various ways, but we have always chosen the variant that best explains the essential nature of the concept, rather than one that facilitates computation (and we avoid the complication of alternative versions of the same formula). We recognize that many readers will not be accustomed to working with symbols, and so may be uncomfortable initially with this method of setting out information. However, the multi-part nature of many concepts means that a formula is the most concise and effective 'shorthand' way of showing what is involved, and we would strongly urge all readers to make the small initial effort required to learn to read this language; the rewards for doing so are very substantial.

## 1.2 The four case studies and the data sets

Throughout the book we will make frequent reference to four data sets that were compiled and used by historians to investigate specific historical

issues.[a] In order to illustrate and practise the various techniques discussed in subsequent chapters we will normally draw on one of these data sets, and the results will be reproduced and discussed in depth in chapters 14 and 15, by which time readers will have encountered all the procedures used in these studies. We are extremely grateful to the authors for their agreement to the use of their statistical material and for making available their unpublished data. These data sets are not reproduced in this book but can be accessed without charge on the Cambridge University Press web site <http://uk.cambridge.org/resources/0521806631>. Further instructions for downloading the data are given on the web site.

A brief outline of the issues raised in these studies is given in appendix A, together with a description of each of the individual statistical series that make up the data sets. In all cases we refer to each series in a data set by a short name given in capital letters, and since some computer programs limit series titles to eight letters we have adopted that restriction. Even when abbreviated these names should be reasonably self-explanatory (for example, BRTHRATE or IRISHMIG) and they are much easier to remember and interpret than symbols. Where the authors we are quoting have used a similar convention we have generally adopted their names; where they have not done so we have created our own.

In addition to these primary data sets, most chapters include references to other examples of the application of the given procedures by historians. These illustrations are designed both to reinforce understanding of the techniques and to show the range of issues that have been addressed by historians with the aid of quantitative methods. The text also uses imaginary data where it is desirable to have very simple numbers in order to illustrate basic concepts or to make calculations without the need for a computer.

## 1.3 Types of measurement

Measurement is the assignment of numbers or codes to observations. The measurements we work with can be classified in various ways. In this introductory section we note a number of basic terms and distinctions.

[a] The four studies are George Boyer, *An Economic History of the English Poor Law,* Cambridge University Press, 1990, chapters 4 and 5; Timothy J. Hatton and Jeffrey G. Williamson, 'After the famine: emigration from Ireland, 1850–1913', *Journal of Economic History,* 53, 1993, pp. 575–600; Daniel K. Benjamin and Levis A. Kochin, 'Searching for an explanation for unemployment in interwar Britain', *Journal of Political Economy*, 87, 1979, pp. 441–78; and Richard H. Steckel, 'The age at leaving home in the United States, 1850–1860', *Social Science History*, 20, 1996, pp. 507–32.

### 1.3.1 Cases, variables, and values

A data set consists of a series of units or **cases** each of which has one or more characteristics, known as **variables**. For each variable there is a sequence of varying observations, each with its own particular **value**. The cases are the basic unit for which the measurements are available, and there are two main types (see §1.3.2). They can be individuals, households, firms, towns, churches, votes, or any other subject of analysis; or they can be weeks, quarters, years, or any other period of time for which a data series is available.

In the English Poor Law data set:

*Cases:* The cases are 311 English parishes (identified by a number).

*Variables:* There are 15 variables for each parish. These include, for example, the *per capita* relief expenditure of each parish (RELIEF), the annual income of adult male agricultural labourers in the parish (INCOME), and the proportion of unemployed labourers in the parish (UNEMP).

*Values:* For the RELIEF variable the values for the first five parishes are (in shillings): 20.4, 29.1, 14.9, 24.1, 18.2, etc.[b]

In the data set for annual emigration from Ireland:

*Cases:* The cases are the 37 successive years from 1877 to 1913.

*Variables:* There are five basic variables for each year. These are the rates per 1,000 of the population emigrating from Ireland (IRISHMIG), the foreign and domestic employment rates (EMPFOR and EMPDOM), the foreign wage relative to the domestic wage (IRWRATIO), and the stock of previous emigrants (MIGSTOCK).

*Values:* For the main emigration variable (IRISHMIG), the values of this rate for the first five years are: 7.284, 7.786, 8.938, 18.358, and 15.238.

### 1.3.2 Cross-section and time-series variables

A set of measurements that applies to a single case at different periods of time is referred to as a **time series**. The series in the data sets dealing with migration from Ireland from 1877 to 1913, and with unemployment and benefits in Great Britain each year from 1920 to 1938, are both examples of annual time series.

A set of measurements that applies to different cases at a single point in time is referred to as a **cross-section**. The data set for children leaving home is a cross-section, with values for each variable for each child in the sample in 1850.

[b] The data on relief expenditure shown for the 24 Kent parishes in the data set are given to 3 decimal places (i.e. have 3 figures after the decimal point). In the text the values for the first five parishes are **rounded** to 1 decimal place.

It is also possible to combine cross-section and time-series data to create what is known as a **panel** or **longitudinal** data set. The second of the two data sets for Irish emigration (see part (b) of table A.3 in appendix A), which has values for the variables for a cross-section of 32 counties for each of the four census dates (1881, 1891, 1901, and 1911), is an example of panel data. The process of combining the values for the cross-section over time is referred to as **pooling** the data.

In principle, time series are used to study changes *over time*, and cross-sections to study differences *between cases* (for example, families, parishes, or countries) at a particular date. However, it is not always possible to obtain the necessary time series for periods in the past, and historians sometimes attempt to infer how relationships might have changed over time by looking at differences in a cross-section for a more recent date. For example, the patterns of household expenditure of low-income families in a modern cross-section might be used to infer something about overall patterns of expenditure at an earlier date when all families had lower incomes.[c]

### 1.3.3 Levels of measurement

We can distinguish the following three **levels** or **scales** of measurement. Each has its own properties and the nature of the statistical exercises that can be performed depends on the level of the data.

*Nominal measurement*

This is the lowest level and conveys no information about the relations between the values. Each value defines a distinct category but can give no information other than the label or name (hence **nominal** level) of the category. They are sometimes also referred to as **categorical** variables.

For example, a study of migration to urban areas might include as one of the variables the birthplace of the migrants. These towns or villages cannot be ranked or placed in any order in terms of their value as place of birth (though they could be by other criteria such as size, or distance from the migrants' final destination).

*Ordinal measurement*

This applies when it is possible to **order or rank** all the categories according to some criterion without being able to specify the exact size of the interval between any two categories.

[c] A similar procedure is adopted by Boyer, *Poor Law*, p. 126 to explain the increase in relief expenditure over time on the basis of cross-section variations in expenditure across parishes.

This is a very common situation with historical data and can occur for one of three reasons:

- The relevant variables are often *ordinal by definition*. For example, an analysis of the labour force might classify workers as unskilled, semi-skilled, skilled, and professional. We could agree that skilled workers ranked above semi-skilled, and semi-skilled above unskilled, but could not say anything about the distance between the categories.
- *Lack of data might impose an ordinal scale* even when a higher level would, in principle, be possible. For example, in a study of education it would in principle be possible to measure the number of years spent at school by each member of the labour force, but the surviving records may state only whether or not they completed five years' primary education, and if they did whether they then went on to secondary school. We would thus have to treat 'schooling' as an ordinal variable with three values: 'less than five years', 'five years', and 'secondary or above'.
- The third reason for using ordinal scales might be doubts about the *accuracy of the data*. In a study of wealth, for example, we might decide that the actual values reported by property owners were too unreliable to use, but that we could safely order them on the basis of, say, three values: large, medium, and small.

*Interval or ratio measurement*[1]

These measurements have all the properties of an ordinal scale. In addition it is now possible to measure the *exact distance between any pair of values*. This level of measurement is thus truly quantitative, and any appropriate statistical procedure can be applied to values on an interval or ratio scale.

Values measured on such a scale may be either **continuous** or **discrete** (discontinuous). A continuous variable, such as height or income, is measured in units that can be reduced in size to a theoretically infinite degree, limited only by the sensitivity of our measurement procedures. Discrete variables, by contrast, can take only a limited number of pre-determined values (most commonly whole numbers) and occur when we are counting indivisible units such as cotton mills or people.

### 1.3.4 Populations and samples

The term **population** (or universe) refers to all possible observations. If, for example, we were interested in the number of cotton mills in New England in 1880, the 'population' would consist of all mills in existence at that date.[2] In a few cases historians might have such information if, for example, a

complete census of textile mills had been taken at the relevant date. Normally, however, we would have only a **sample**.

The characteristics of the population variables are known as **parameters**, those of the sample variables as **statistics**. Parameters are fixed values at any point in time and are normally unknown. Statistics, on the other hand, are known from the sample, but may vary with each sample taken from the population. The extent of such variation from sample to sample will depend on the homogeneity (uniformity) of the population from which it is drawn.[d]

A crucial feature of any sample is whether or not it is **random**. A random sample satisfies three basic conditions. First, *every item* in the population (parishes in England and Wales, voters in an election, cards in a deck of cards) has *an equal chance of appearing in the sample*. Secondly, *every combination of items* has an equal chance of selection. Thirdly, there is independence of selection: the fact that any one item in the population has been selected has *absolutely no influence on whether or not any other item will be selected*.

When the sample is drawn **with replacement** the same item can be selected more than once; for example, after a card is drawn it is put back in the deck. If the sample is drawn **without replacement** the item can be selected only once. Almost all use of sampling in historical analysis is sampling without replacement. Each parish, each voter, each cotton mill is sampled only once when the data set is being compiled.

The proper procedures to be followed in constructing samples, whether the sampling should be random or some other type, such as stratified or cluster sampling, and the size of the sample required for any particular project are complex subjects which go beyond the scope of this book and specialized texts should be consulted.[3]

### 1.3.5 Dummy variables

A variable that cannot be measured can still be used in quantitative work by assigning values representing each of two (or more) categories. This is known as a **dummy variable**. In the simplest case, where there are only two possible values (equivalent to yes or no) the dummy variable is formed by setting the positive values equal to 1 and the negative values to 0. There are a number of these in the Poor Law data set – for example, whether or not the parish pays child allowances and whether or not it has a workhouse.

[d] Harold Wilson, Britain's only statistically informed prime minister, once remarked that he needed to sip only one spoonful from a plate of soup to know whether it was too hot to drink.