# MY451 Introduction to Quantitative Analysis

*Jouni Kuha*
*Department of Methodology*
*London School of Economics and Political Science*

*2016-17*

# Contents

# Course information


## MY451: Introduction to Quantitative Analysis

### Course Description

This course is intended for those with little or no past training in quantitative methods. The course is an intensive introduction to some of the principles and methods of statistical analysis in social research. Topics covered in MY451 include descriptive statistics, basic ideas of inference and estimation, contingency tables and an introduction to linear regression models. For those with some quantitative training the slightly more advanced course MY452 (Applied Regression Analysis) would be more appropriate, followed by other Department of Methodology and Department of Statistics courses on quantitative methods, such as MY454 (Applied Statistical Computing), MY455 (Multivariate Analysis and Measurement), MY456 (Survey Methodology), MY457 (Causal Inference for Observational and Experimental Studies), MY459 (Quantitative Text Analysis), ST416 (Multilevel Modelling), and ST442 (Longitudinal Data Analysis).

Please note that up to 2010-11 the course was called MI451, so older materials will bear that code. For example, you will find past examination papers on the Library website under MI451.

### Course Objectives

This course aims to impart a level of familiarity suitable for a moderately critical understanding of the statistical material in the journals commonly used by students in their work and the ability to use some elementary techniques.

### Staff

- **Teachers responsible**: Dominik Hangartner in MT; Jonathan Jackson in LT
- **Other Lecturers**: Benjamin Wilson (MT) and Matthew Mulford (LT)
- **Course Administrator**: Esther Heyhoe, email `e.heyhoe@lse.ac.uk`
- **Class Teachers**: See the MY451 Moodle site for the list of class teachers.

**Teaching**

- **Lectures**: Ten two-hour lectures in weeks 1 - 5 and 7 - 11:
  - Michaelmas Term (MY451M): Thu 2.00pm - 4.00pm in the Peacock Theatre
  - Lent Term (MY451L): Wed 2.00pm - 4.00pm in the Hong Kong Theatre (CLM.G.02)
- **Computer classes**: Each student will attend a one-hour class each week, **starting in Week 2**. Please see on-line timetables for the times and places of the class groups. The allocation of students to classes is done through LSE for You or the Moodle system (see below), depending on your status. This will be exaplained in the first lecture and on the MY451 Moodle page. Please do not change the classes allocated to you without our consent.

**Course Materials**

- **Coursepack**: This coursepack is the main course text. Printed copies of this coursepack will be available for purchase at the beginning of each term, and instructions on how to obtain a hard copy will be given at the first lecture and on the MY451 Moodle page. Digital copies of this coursepack are available for download for free from the Moodle page.
- **Lecture slides**: Copies of most of the slides displayed during the lectures can be downloaded from the MY451 Moodle page.
- **Recommended course texts**:
  - Alan Agresti and Christine Franklin (2013). *Statistics: The Art and Science of Learning from Data* (Third Ed.). Pearson.
  - Alan Agresti and Barbara Finlay (2013). *Statistical Methods for the Social Sciences* (Fourth Ed.). Pearson

  Earlier editions are also suitable. While neither of these books is absolutely required, we recommend that you purchase one if you would like to have additional explanation, examples and exercises to supplement the coursepack. Please note that during the examination you will be allowed to use the coursepack and any other notes or books that you find useful. Of these two, Agresti and Finlay is a bit more advanced. It is also the recommended additional course text for MY452 (which also has a coursepack similar to this one), so you may want to purchase it if you are planning to also take that course.
- **Other text books**: There are hundreds of introductory textbooks on statistics and quantitative methods, many of them covering almost identical material. If you have one which you would like to use, and which looks as if it covers the same material at about the same level as this course, then it is probably suitable as additional reading. If in doubt, please ask one of the lecturers.
  - There are also many books which focus on the SPSS statistical software package used in the computer classes. We do not consider them necessary for this course, or for learning statistics or SPSS. If, however, you feel otherwise, some such books are mentioned at the beginning of Section **??**.

### Homework

There will be homework assignments each week. These are your opportunity to practice interpreting the results of the types of analyses that are introduced in the lectures and then carried out in the computer classes. The homework questions are also broadly similar in style to typical examination questions for MY451. The questions are based on results of the computer class exercises or of further similar analyses. For the homework, you can use output that you produced yourself in the class or afterwards. Alternatively, all the computer outputs that are required to answer the homework questions can be downloaded from the Moodle page.

Homework should be submitted by Monday after its distribution in class, at the latest by 5pm in MT and by 10am in LT. Please see the MY451 Moodle page for instructions on how to submit your homeworks. Homework submitted in time will be corrected and returned at the next class. Please make sure to clearly indicate your name, and the time, class number and class teacher of your computer class on all homework assignments. For the homework assignments, students are encouraged to work together.

### MY451 on Moodle

The course materials are all available on Moodle. Go to `http://moodle.lse.ac.uk/` and login using your *username* and *password* (the same as for your LSE e-mail). Then in the *select courses* dialogue box type in MY451, and in *search results* click on MY451. The site contains the structure of the course week by week, the readings, weekly computer class assignments and the associated data sets, coursepack and other materials, as well as a section on news and announcements.

### Advisory Sessions

There will be a time each week during the Michaelmas and Lent terms during which a member of the teaching team will be available at the Department of Methodology in Columbia House to answer questions about the course. Information on the times of the advisory hours will be given at the first lecture. These sessions are not intended to be private tutorials and you will be expected to have made a genuine attempt to work through the relevant sections of the coursepack prior to coming to an advisory session. Also, questions addressing material from that week's lecture will be given priority. If you are having difficulty with a topic, come and see us at that time. If you wait until the end of term, when the advisory sessions can be very busy, we cannot guarantee that your questions from earlier sessions will be covered. **There will be no advisory sessions after the end of Lent Term.**

### Notes on studying for the course

To learn the material from this course you must do the work every week since it is cumulative; if you miss a week or two (or sleep through it!) there is a chance that you will find yourself hopelessly lost. So this is definitely not a "pick and choose" course! Also bear in mind that most people cannot learn quantitative techniques passively by just turning up to the lectures and reading the occasional chapter in a textbook. To learn statistics you have to do it; there are no shortcuts. Thus in addition to a two-hour weekly lecture there will be one-hour computer classes (in which you do some data analysis and interpretation using SPSS - instructions will be provided) and there will be weekly homework (which will be corrected but not graded by your class teacher). Doing the assignments in the computer classes and the homework demonstrate

whether you have understood and can apply what was covered in the lectures. If you are having any trouble this will reveal what the problem is. Thus the course is designed to have multiple, reinforcing ways of helping you get to grips with this material.

### Examinations/assessment

There will be a **two-hour examination in the Summer Term 2016**. This is an **open-book examination**, to which you are allowed to bring in any written material, including the coursepack and any text book with your notes. You are required to bring your own calculators for the examination. Examination papers from previous years are available for revision purposes at the LSE library web site. Students should understand that past examinations should only be used as rough guides to the types of questions that are likely to appear on the examination.

For many of you, MY451 is only one part of a package of methods training with a course code such as MC4M1 or MY4M1. If this is the case, your result for MY451 will contribute only part of the final mark for the package, with a weight determined by the regulations.

### Computing

Students must know their Username and Password in time for the first class. This information can be obtained from IT Help Desk (Library, 1st floor). The course uses Microsoft Windows-based software. If you are not familiar with the program, you might want to attend introductory courses in Windows during the first two weeks of the term. The statistical package being used is SPSS, which will be introduced in the first computing class of the course.

### Software availability

A personal copy of the program SPSS for Windows can be bought by course participants (maximum 1 copy per person) for the price of £10 from IT services, **not** the Department of Methodology.

### Feedback

We would welcome any comments you have on the course. If there are any problems that we can deal with, we will attempt to do so as quickly as possible. Speak to any member of the course team, or to your departmental supervisor if you feel that would be easier for you. Also please let us know if you find any errors or omissions in the coursepack, so that we can correct them for next year.

### Acknowledgements

The computer classes were extensively revised for 2012-13. The hard work of designing the new exercises and writing the revised instructions was done by Sally Stares.

This coursepack bears many traces of previous materials and all of their authors, Colm O'Muircheartaigh, Colin Mills, Matt Mulford, Fiona Steele, Paul Mitchell, and Sally Stares. Many thanks to Farimah Daftary, Sue Howard, Jon Jackson, Paul Mitchell, Indraneel Sircar, and many students of previous years for comments and suggestions which are incorporated in the current revision.

**Course Programme**

| | |
|---|---|
| **Week 1** | |
| Lecture | Course overview and organisation. Introduction to basic concepts |
| Class | **No class** |
| Coursepack | Chapter **??** |
| **Week 2** | |
| Lecture | Descriptive statistics for categorical variables |
| Class | Introduction to SPSS. Descriptive statistics for categorical variables |
| Coursepack | Sections **??**–**??** and **??** |
| **Week 3** | |
| Lecture & Class | Descriptive statistics for continuous variables |
| Coursepack | Sections **??**–**??** |
| **Week 4** | |
| Lecture & Class | Analysis of two-way contingency tables |
| Coursepack | Chapters **??** and **??** |
| **Week 5** | |
| Lecture & Class | Inference for means in two populations |
| Coursepack | Chapters **??** and **??** |
| **Week 6** | |
| Reading Week | **No lecture, no class**. Revision quiz available on Moodle. |
| **Week 7** | |
| Lecture & Class | Inference for proportions in one and two populations |
| Coursepack | Chapter **??** |
| **Week 8** | |
| Lecture & Class | Correlation and simple linear regression as descriptive methods |
| Coursepack | Sections **??**–**??** |
| **Week 9** | |
| Lectures | |
| Hour 1 | Inference for the simple linear regression model |
| Hour 2 | Three-way contingency tables |
| Class | Simple linear regression |
| Coursepack | Section **??** (Hour 1); Section **??** and Chapter **??** (Hour 2) |
| **Week 10** | |
| Lecture & Class | Multiple linear regression |
| Coursepack | Sections **??**–**??** |
| **Week 11** | |
| Lecture | Review and further statistical methods |
| Class | Discussion and additional exercise on multiple linear regression |
| Coursepack | Chapter **??** |

**FAQ: Frequently Asked Questions**

**Why do we use SPSS? I've heard that SAS/STATA/MINITAB/R/LIMDEP is much better.** At this level it does not matter which program you use since we are learning standard procedures that are common to all programs. In favour of SPSS is that it has an easy to learn interface and that it is widely available at sites both within and outside the UK.

**Can I get a copy of the SPSS software to use on my home computer?** Yes, for a small fee from IT Services helpdesk, not from the Department of Methodology.

**I'm taking MY451 because I want to learn how to use SPSS but we don't seem to learn very much about the program. Why is that?** MY451 is not a course about learning to use SPSS. We use the program merely to facilitate data analysis and interpretation. Some options for learning more about SPSS will be mentioned in the first lecture.

**I'm taking MY451 to help me analyse data for my dissertation. Can I discuss my data and my specific problems with the lecturers?** Yes, but not during the course. Staff of the Department of Methodology will be happy to talk to you about problems specific to your dissertation during the weekly sessions of the Methodology Surgery (see the website of the department for more information).

**Does the coursepack contain everything I need to know for the exam?** Yes. However, you will stand by far the best chance in the exam if you also attend the lectures, where the lecturers emphasise and explain the key parts of the material.

**The lecturer introduced some material that was not in the coursepack. Do I need to know that material?** This is almost certainly an illusion. The lectures will not introduce any genuinely new material not included in the course pack. However, sometimes the lecturer may of course use different words or a different example to further explain some topic. Copies of the most relevant notes displayed at the lectures will be posted in the MY451 Moodle site. All of the material required for the exam is contained in the coursepack, with the posted lecture notes as additional clarification.

**Do I need to bring my coursepack to lectures and classes?** Yes, because some of the material displayed during the lectures will be graphs or tables straight from the coursepack. However, an even more important way of using the coursepack is to read the relevant parts of it *before* each lecture. In addition, you will find the instructions for the class exercises at the end of the coursepack.

**I can't make it to the advisory hour. Why aren't there more of them?** Advisory hours are offered in addition to the normal support arrangements for School courses and rely on the goodwill and enthusiasm of the course team. Given that most team members also offer their time to support advisory hours for other courses there is a limit to how much time they can volunteer. If you genuinely are unable to attend an advisory hour, but desperately need advice, lecturers will always see you in their personal office hours. If these are inconvenient you may be able to set up an appointment at another time by special arrangement.

**There was a long queue at advisory and I didn't get much/any time to discuss my particular problem. Isn't there a better way to arrange things?** Advisory sessions are meant for quick feedback and clarification of problems on a first-come first served basis. They are not meant to be personal tutorials, seminars or impromptu lectures. They are also not meant to be a substitute for attending the lectures; reading the course-pack; doing the homework; thinking. Priority is always given to problems related to the previous week's lecture material. If several people have the same problem it will usually be possible to talk to everyone together, but there is a finite limit to how many people can be crammed into an office at the same time!

**I want to bring my homework to the advisory session because I couldn't understand the class teacher's comments. Is that OK?** Yes, but it is unlikely that the person that wrote the comments on your homework will be taking the advisory hour. Usually it will be much better to discuss homework comments directly with your class teacher in class.

**Can I work together on the homework with my friends?** Yes, we positively encourage you to discuss the homework assignments with your colleagues. If you do this, please submit **one** copy to be marked rather than multiple copies of the same answers.

**If I get less than 50% in the exam what happens now?** Candidates sit MY451 under different examination regulations depending on which department they are in and what degree they are registered for. For some degrees a pass in MY451 is required, for others the mark from MY451 is just one component of an overall course mark (implying a "pass" mark in MY451 is not required). The regulations for your degree are not set by Methodology, but by your home department. To find out whether a pass in MY451 is required, consult your degree regulations or ask the tutor responsible for your program in your home department. Candidates who fail at the first attempt and whose degree regulations require a pass are, as per the usual School examination rules, entitled, at the next available opportunity, to ONE more attempt. This will be in the Summer Term of the next Session.

**Why don't we get our examination scripts returned after they have been marked?** Written work is of two sorts, formative and summative. The former, for example the homework exercises, is meant to give you feedback on how you are doing and will, where appropriate, be returned with comments. The latter, for example the examination, is meant to evaluate your performance, not explain to you where you have gone wrong, or for that matter where you have done exceptionally well.

**If I don't agree with the mark, how do I appeal?** The cultural norm in the UK is that marks are not arrived at by a process of teacher-student negotiation. You can make a formal appeal through the School's appeal process (see the appropriate section of the website for details). NB: appeals cannot be made on grounds of academic substance, only on grounds of procedural irregularities. In other words an appeal will not be allowed if the only grounds you have for appealing is that you/your friend/your personal advisor/your spiritual guru think your script was worth more marks than the examiners did.

**The class teachers are always very busy during class helping people with problems and I can't get enough personal attention. Why can't there be more class teachers in the classroom?** For MY451 there will usually be two class teachers per classroom i.e. double the normal complement. Even so we realise that you may not be able to get attention at the exact moment you want it. Please be patient and if help is not immediately available you can always try asking the person who is sitting next to you!

**I'm not registered at the LSE but at another University of London college. Can I attend this course?** Normally yes, but you will have to complete an intercollegiate enrolment form.

**I would like to audit the course without taking the exam. Is that OK?** Yes, you are welcome to attend the lectures providing you are an LSE/University of London student and there is room for you.

**MY451 is not challenging enough for me. Is there a more difficult course?** Yes, MY452 and numerous other courses offered by the Department of Methodology and the Statistics department.

# Contents

# Chapter 1

# Introduction

## 1.1 What is the purpose of this course?

The title of any course should be descriptive of its contents. This one is called

**MY451: Introduction to Quantitative Analysis**

Every part of this tells us something about the nature of the course:

The **M** stands for *Methodology* of social research. Here *research* refers to activities aimed at obtaining new knowledge about the world, in the case of the social sciences the *social* world of people and their institutions and interactions. Here we are concerned solely with *empirical* research, where such knowledge is based on information obtained by *observing* what goes on in that world. There are many different ways (*methods*) of making such observations, some better than others for deriving valid knowledge. "Methodology" refers both to the methods used in particular studies, and the study of research methods in general.

The word **analysis** indicates the area of research methodology that the course is about. In general, any empirical research project will involve at least the following stages:

1. Identifying a research *topic*
2. Formulating *research questions*
3. Deciding what kinds of *information* to collect to try to answer the research questions, and deciding how to collect it and where to collect it from
4. Collecting the information
5. *Analysing* the information in appropriate ways to answer the research questions
6. *Reporting* the findings

The empirical information collected in the research process is often referred to as *data*. This course is mostly about some basic methods for step 5, the *analysis* of such data.

Methods of analysis, however competently used, will not be very useful unless other parts of the research process have also been carried out well. These other parts, which (especially steps 2–4 above) can be broadly termed *research design*, are covered on other courses, such as MY400 (Fundamentals of Social Science Research Design) or comparable courses at your own department. Here we will mostly not consider research design, in effect assuming that we start at a point where we want to analyse some data which have been collected in a sensible way to

answer meaningful research questions. However, you should bear in mind throughout the course that in a real research situation both good design and good analysis are essential for success.

The word **quantitative** in the title of the course indicates that the methods you will learn here are used to analyse quantitative data. This means that the data will enter the analysis in the form of *numbers* of some kind. In social sciences, for example, data obtained from administrative records or from surveys using structured interviews are typically quantitative. An alternative is *qualitative* data, which are not rendered into numbers for the analysis. For example, unstructured interviews, focus groups and ethnography typically produce mostly qualitative data. Both quantitative and qualitative data are important and widely used in social research. For some research questions, one or the other may be clearly more appropriate, but in many if not most cases the research would benefit from collecting both qualitative and quantitative data. This course will concentrate solely on quantitative data analysis, while the collection and analysis of qualitative data are covered on other courses (e.g. MY421, MY426 and MY427), which we hope you will also be taking.

All the methods taught here, and almost all approaches used for quantitative data analysis in the social sciences in general, are *statistical* methods. The defining feature of such methods is that randomness and probability play an essential role in them; some of the ways in which they do so will become apparent later, others need not concern us here. The title of the course could thus also have included the word *statistics*. However, the Department of Methodology courses on statistical methods (e.g. MY451, MY465, MY452, MY455 and MY459) have traditionally been labelled as courses on "quantitative analysis" rather than "statistics". This is done to indicate that they differ from classical introductory statistics courses in some ways, especially in the presentation being less mathematical.

The course is called an "**Introduction** to Quantitative Analysis" because it is an introductory course which does not assume that you have learned any statistics before. MY451 or a comparable course should be taken before more advanced courses on quantitative methods. Statistics is a cumulative subject where later courses build on material learned on earlier ones. Because MY451 is introductory, it will start with very simple methods, and many of the more advanced (and powerful) ones will only be covered on the later courses. This does not, however, mean that you are wasting your time here even if it is methods from, say, MY452 that you will eventually need most: understanding the material of this course is essential for learning more advanced methods.

Finally, the course has an **MY** code, rather than GV, MC, PS, SO, SP, or whatever is the code of your own department. MY451 is taken by students from many different degrees and departments, and thus cannot be tailored to any one of them specifically. For example, we will use examples from many different social sciences. However, this generality is definitely a good thing: the reason we *can* teach all of you together is that statistical methods (just like the principles of research design or qualitative research) are generic and applicable to the analysis of quantitative data in all fields of social research. There is not, apart from differences in emphases and priorities, one kind of statistics for sociology and another for political science or economics, but one coherent set of principles and methods for all of them (as well as for psychiatry, epidemiology, biology, astrophysics and so on). After this course you will have taken the first steps in learning about all of that.

At the end of the course you should be familiar with certain methods of statistical analysis. This will enable you to be both a user and a consumer of statistics:

- You will be able to use the methods to analyse your own data and to report the results of the analyses.
- Perhaps even more importantly, you will also be able to understand (and possibly criticize)

their use in other people's research. Because interpreting results is typically somewhat easier than carrying out new analyses, and because all statistical methods use the same basic ideas introduced here, you will even have some understanding of many of the techniques not discussed on this course.

Another pair of different but complementary aims of the course is that MY451 is both a self-contained unit and a prerequisite for courses that follow it:

- If this is the last statistics course you will take, it will enable you to understand and use the particular methods covered here. This includes the technique of linear regression modelling (described in Chapter **??**), which is arguably the most important and commonly used statistical method of all. This course can, however, introduce only the most important elements of linear regression, while some of the more advanced ones are discussed only on MY452.
- The ideas learned on this course will provide the conceptual foundation for any further courses in quantitative methods that you may take. The basic ideas will then not need to be learned from scratch again, and the other courses can instead concentrate on introducing further, ever more powerful statistical methods for different types of data.

## 1.2 Some basic definitions

Like any discipline, statistics involves some special terminology which makes it easier to discuss its concepts with sufficient precision. Some of these terms are defined in this section, while others will be introduced later when they are needed.

You should bear in mind that all terminology is arbitrary, so there may be different terms for the same concept. The same is true of notation and symbols (such as $n$, $\mu$, $\bar{Y}$, $R^2$, and others) which will be introduced later. Some statistical terms and symbols are so well established that they are almost always used in the same way, but for many others there are several versions in common use. While we try to be consistent with the notation and terminology within this coursepack, we cannot absolutely guarantee that we will not occasionally use different terms for the same concept even here. In other textbooks and in research articles you will certainly occasionally encounter alternative terminology for some of these concepts. If you find yourself confused by such differences, please come to the advisory hours or ask your class teacher for clarification.

### 1.2.1 Subjects and variables

Table **??** shows a small set of quantitative data. Once collected, the data are typically arranged and stored in this kind of spreadsheet-type rectangular table, known as a **data matrix**. In the computer classes you will see data in this form in SPSS.

Table 1.1: An example of a small data matrix based on data from the U.S. General Social Survey (GSS), showing measurements of seven variables for 20 respondents in a social survey. The variables are defined as *age*: age in years; *sex*: sex (1=male; 2=female); *educ*: highest year of school completed; *wrkstat*: labour force status (1=working full time; 2=working part time; 3=temporarily not working; 4=unemployed; 5=retired; 6=in education; 7=keeping house; 8=other); *life*: is life exciting or dull? (1=dull; 2=routine; 3=exciting); *income4*: total annual family income (1=$24,999 or less; 2=$25,000–$39,999; 3=$40,000–$59,999; 4=$60,000 or more; 99 indicates a missing value); *pres92*: vote in the 1992 presidential election (0=did not vote or not eligible to vote; 1=Bill Clinton; 2=George H. W. Bush; 3=Ross Perot; 4=Other).

| Id | *age* | *sex* | *educ* | *wrkstat* | *life* | *income4* | *pres92* |
|---|---|---|---|---|---|---|---|
| 1 | 43 | 1 | 11 | 1 | 2 | 3 | 2 |
| 2 | 44 | 1 | 16 | 1 | 3 | 3 | 1 |
| 3 | 43 | 2 | 16 | 1 | 3 | 3 | 2 |
| 4 | 78 | 2 | 17 | 5 | 3 | 4 | 1 |
| 5 | 83 | 1 | 11 | 5 | 2 | 1 | 1 |
| 6 | 55 | 2 | 12 | 1 | 2 | 99 | 1 |
| 7 | 75 | 1 | 12 | 5 | 2 | 1 | 0 |
| 8 | 31 | 1 | 18 | 1 | 3 | 4 | 2 |
| 9 | 54 | 2 | 18 | 2 | 3 | 1 | 1 |
| 10 | 23 | 2 | 15 | 1 | 2 | 3 | 3 |
| 11 | 63 | 2 | 4 | 5 | 1 | 1 | 1 |
| 12 | 33 | 2 | 10 | 4 | 3 | 1 | 0 |
| 13 | 39 | 2 | 8 | 7 | 3 | 1 | 0 |
| 14 | 55 | 2 | 16 | 1 | 2 | 4 | 1 |
| 15 | 36 | 2 | 14 | 3 | 2 | 4 | 1 |
| 16 | 44 | 2 | 18 | 2 | 3 | 4 | 1 |
| 17 | 45 | 2 | 16 | 1 | 2 | 4 | 1 |
| 18 | 36 | 2 | 18 | 1 | 2 | 99 | 1 |
| 19 | 29 | 1 | 16 | 1 | 3 | 3 | 1 |
| 20 | 30 | 2 | 14 | 1 | 2 | 2 | 1 |

The rows (moving downwards) and columns (moving left to right) of a data matrix correspond to the first two important terms: the rows to the *subjects* and the columns to the *variables* in the data.

- A **subject** is the smallest unit yielding information in the study. In the example of Table **??**, the subjects are individual people, as they are in very many social science examples. In other cases they may instead be families, companies, neighbourhoods, countries, or whatever else is relevant in a particular study. There is also much variation in the term itself, so that instead of "subjects", a study might refer to "units", "elements", "respondents" or "participants", or simply to "persons", "individuals", "families" or "countries", for example. Whatever the term, it is usually clear from the context what the subjects are in a particular analysis.

  The subjects in the data of Table **??** are uniquely identified only by a number (labelled

"Id") assigned by the researcher, as in a survey like this their names would not typically be recorded. In situations where the identities of individual subjects are available and of interest (such as when they are countries), their names would typically be included in the data matrix.

- A **variable** is a characteristic which varies between subjects. For example, Table **??** contains data on seven variables — age, sex, education, labour force status, attitude to life, family income and vote in a past election — defined and recorded in the particular ways explained in the caption of the table. It can be seen that these are indeed "variable" in that not everyone has the same value of any of them. It is this variation that makes collecting data on many subjects necessary and worthwhile. In contrast, research questions about characteristics which are the same for every subject (i.e. *constants* rather than variables) are rare, usually not particularly interesting, and not very difficult to answer.

  The labels of the columns in Table **??** (*age*, *wrkstat*, *income4* etc.) are the names by which the variables are uniquely identified in the data file on a computer. Such concise titles are useful for this purpose, but should be avoided when reporting the results of data analyses, where clear English terms can be used instead. In other words, a report should not say something like "The analysis suggests that WRKSTAT of the respondents is…" but instead something like "The analysis suggests that the labour force status of the respondents is…", with the definition of this variable and its categories also clearly stated.

Collecting quantitative data involves determining the values of a set of variables for a group of subjects and assigning numbers to these values. This is also known as **measuring** the values of the variables. Here the word "measure" is used in a broader sense than in everyday language, so that, for example, we are measuring a person's sex in this sense when we assign a variable called "Sex" the value 1 if the person is male and 2 if she is female. The value assigned to a variable for a subject is called a **measurement** or an **observation**. Our data thus consist of the measurements of a set of variables for a set of subjects. In the data matrix, each row contains the measurements of all the variables in the data for one subject, and each column contains the measurements of one variable for all of the subjects.

The number of subjects in a set of data is known as the **sample size**, and is typically denoted by $n$. In a survey, for example, this would be the number of people who responded to the questions in the survey interview. In Table **??** we have $n = 20$. This would normally be a very small sample size for a survey, and indeed the real sample size in this one is several thousands. The twenty subjects here were drawn from among them to obtain a small example which fits on a page.

A common problem in many studies is **nonresponse** or **missing data**, which occurs when some measurements are not obtained. For example, some survey respondents may refuse to answer certain questions, so that the values of the variables corresponding to those questions will be missing for them. In Table **??**, the income variable is missing for subjects 6 and 18, and recorded only as a *missing value code*, here "99". Missing values create a problem which has to be addressed somehow before or during the statistical analysis. The easiest approach is to simply ignore all the subjects with missing values and use only those with complete data on all the variables needed for a given analysis. For example, any analysis of the data in Table **??** which involved the variable *income4* would then exclude all the data for subjects 6 and 18. This method of "complete-case analysis" is usually applied automatically by most statistical software packages, including SPSS. It is, however, not a very good approach. For example, it means that a lot of information will be thrown away if there are many subjects with some observations missing. Statisticians have developed better ways of dealing with missing data, but they are unfortunately beyond the scope of this course.

### 1.2.2 Types of variables

Information on a variable consists of the observations (measurements) of it for the subjects in our data, recorded in the form of numbers. However, not all numbers are the same. First, a particular way of measuring a variable may or may not provide a good measure of the concept of interest. For example, a measurement of a person's weight from a well-calibrated scale would typically be a good measure of the person's true weight, but an answer to the survey question "How many units of alcohol did you drink in the last seven days?" might be a much less accurate measurement of the person's true alcohol consumption (i.e. it might have *measurement error* for a variety of reasons). So just because you have put a number on a concept does not automatically mean that you have captured that concept in a useful way. Devising good ways of measuring variables is a major part of research design. For example, social scientists are often interested in studying attitudes, beliefs or personality traits, which are very difficult to measure directly. A common approach is to develop *attitude scales*, which combine answers to multiple questions ("items") on the attitude into one number.

Here we will again leave questions of measurement to courses on research design, effectively assuming that the variables we are analysing have been measured well enough for the analysis to be meaningful. Even then we will have to consider some distinctions between different kinds of variables. This is because the type of a variable largely determines which methods of statistical analysis are appropriate for that variable. It will be necessary to consider two related distinctions:

- Between different measurement levels
- Between continuous and discrete variables

#### Measurement levels

When a numerical value of a particular variable is allocated to a subject, it becomes possible to relate that value to the values assigned to other subjects. The **measurement level** of the variable indicates how much information the number provides for such comparisons. To introduce this concept, consider the variables obtained as answers to the following three questions in the former U.K. General Household Survey:

1

*Are you*

| | |
|---|---|
| *single, that is, never married?* | (coded as 1) |
| *married and living with your husband/wife?* | (2) |
| *married and separated from your husband/wife?* | (3) |
| *divorced?* | (4) |
| *or widowed?* | (5) |

2

*Over the last twelve months, would you say your health has on the whole been good, fairly good, or not good?*
("Good" is coded as 1, "Fairly Good" as 2, and "Not Good" as 3.)

3

*About how many cigaretters A DAY do you usually smoke on weekdays?*
(Recorded as the number of cigarettes)

These variables illustrate three of the four possibilities in the most common classification of measurement levels:

- A variable is measured on a **nominal scale** if the numbers are simply labels for different possible values (*levels* or *categories*) of the variable. The only possible comparison is then to identify whether two subjects have the *same* or *different* values of the variable. The marital status variable

  1

  is measured on a nominal scale. The values of such *nominal-level variables* are not in any order, so we cannot talk about one subject having "more" or "less" of the variable than another subject; even though "divorced" is coded with a larger number (4) than "single" (1), divorced is not more or bigger than single in any relevant sense. We also cannot carry out arithmetical calculations on the values, as if they were numbers in the ordinary sense. For example, if one person is single and another widowed, it is obviously nonsensical to say that they are on average separated (even though $(1 + 5)/2 = 3$).

  The only requirement for the codes assigned to the levels of a nominal-level variable is that different levels must receive different codes. Apart from that, the codes are arbitrary, so that we can use any set of numbers for them in any order. Indeed, the codes do not even need to be numbers, so they may instead be displayed in the data matrix as short words ("labels" for the categories). Using successive small whole numbers $(1, 2, 3, \ldots)$ is just a simple and concise choice for the codes.

  Further examples of nominal-level variables are the variables *sex*, *wrkstat*, and *pres92* in Table **??**.

- A variable is measured on an **ordinal scale** if its values do have a natural ordering. It is then possible to determine not only whether two subjects have the same value, but also whether one or the other has a *higher* value. For example, the self-reported health variable

  2

  is an ordinal-level variable, as larger values indicate worse states of health. The numbers assigned to the categories now have to be in the correct order, because otherwise information about the true ordering of the categories would be distorted. Apart from the order, the choice of the actual numbers is still arbitrary, and calculations on them are still not strictly speaking meaningful.

  Further examples of ordinal-level variables are *life* and *income4* in Table **??**.

- A variable is measured on an **interval scale** if *differences* in its values are comparable. One example is temperature measured on the Celsius (Centigrade) scale. It is now meaningful to state not only that 20°C is a *different* and *higher* temperature than 5°C, but also that the *difference* between them is 15°C, and that that difference is of the same size as the difference between, say, 40°C and 25°C. Interval-level measurements are "proper" numbers in that calculations such as the average noon temperature in London over a year are meaningful. What we *cannot* do is to compare *ratios* of interval-level variables. Thus 20°C is not four times as warm as 5°C, nor is their real ratio the same as that of 40°C and 10°C. This is because the zero value of the Celcius scale (0°C) is not the lowest possible temperature but an arbitrary point chosen for convenience of definition.

- A variable is measured on a **ratio scale** if it has all the properties of an interval-level variable and also a true zero point. For example, the smoking variable

  3

  is measured on a ratio level, with zero cigarettes as its point of origin. It is now possible to carry out all the comparisons possible for interval-level variables, and also to compare ratios. For example, it is meaningful to say that someone who smokes 20 cigarettes a day smokes *twice* as many cigarettes as one who smokes 10 cigarettes, and that that ratio is equal to the ratio of 30 and 15 cigarettes.

  Further examples of ratio-level variables are *age* and *educ* in Table **??**.

The distinction between interval-level and ratio-level variables is in practice mostly unimportant, as the same statistical methods can be applied to both. We will thus consider them together throughout this course, and will, for simplicity, refer to variables on either scale as interval level variables. Doing so is logically coherent, because ratio level variables have all the properties of interval level variables, as well the additional property of a true zero point.

Similarly, nominal and ordinal variables can often be analysed with the same methods. When this is the case, we will refer to them together as nominal/ordinal level variables. There are, however, contexts where the difference between them matters, and we will then discuss nominal and ordinal scales separately.

The simplest kind of nominal variable is one with only *two* possible values, for example sex recorded as "male" or "female" or an opinion recorded just as "agree" or "disagree". Such a variable is said to be **binary** or **dichotomous**. As with any nominal variable, codes for the two levels can be assigned in any way we like (as long as different levels get different codes), for example as 1=Female and 2=Male; later it will turn out that in some analyses it is most convenient to use the values 0 and 1.

The distinction between ordinal-level and interval-level variables is sometimes further blurred in practice. Consider, for example, an attitude scale of the kind mentioned above, let's say a scale for happiness. Suppose that the possible values of the scale range from 0 (least happy) to 48 (most happy). In most cases it would be most realistic to consider these measurements to be on an ordinal rather than an interval scale. However, statistical methods developed specifically for ordinal-level variables do not cope very well with variables with this many possible values. Thus ordinal variables with many possible values (at least more than ten, say) are typically treated as if they were measured on an interval scale.

### Continuous and discrete variables

This distinction is based on the possible values a variable can have:

- A variable is **discrete** if its basic unit of measurement cannot be subdivided. Thus a discrete variable can only have certain values, and the values between these are logically impossible. For example, the marital status variable

  1

  and the health variable

  2

  defined under "Measurement Levels" in Section **??** are discrete, because values like marital status of 2.3 or self-reported health of 1.7 are impossible given the way the variables are defined.

- A variable is **continuous** if it can in principle take infinitely varied fractional values. The idea implies an unbroken scale or continuum of possible values. Age is an example of a continuous variable, as we can in principle measure it to any degree of accuracy we like — years, days, minutes, seconds, micro-seconds. Similarly, distance, weight and even income can be considered to be continuous.

You should note the "in principle" in this definition of continuous variables above. Continuity is here a pragmatic concept, not a philosophical one. Thus we will treat age and income as continous even though they are in practice measured to the nearest year or the nearest hundred pounds, and not in microseconds or millionths of a penny (nor is the definition inviting you to start musing on quantum mechanics and arguing that nothing is fundamentally continuous). What the distinction between discrete and continuous really amounts to in practice is the difference between variables which in our data tend to take relatively few values (discrete variables) and ones which can take lots of different values (continuous variables). This also implies that we will sometimes treat variables which are undeniably discrete in the strict sense as if they were really continuous. For example, the number of people is clearly discrete when it refers to numbers of registered voters in households (with a limited number of possible values in practice), but effectively continuous when it refers to populations of countries (with very many possible values).

The measurement level of a variable refers to the way a characteristic is recorded in the data, not to some other, perhaps more fundamental version of that characteristic. For example, annual income recorded to the nearest dollar is continuous, but an income variable (c.f. Table **??**) with values

- if annual income is $24,999 or less;
- if annual income is $25,000–$39,999;
- if annual income is $40,000–$59,999;
- if annual income is $60,000 or more

is discrete. This kind of variable, obtained by grouping ranges of values of an initially continuous measurement, is common in the social sciences, where the exact values of such variables are often not that interesting and may not be very accurately measured.

The term **categorical variable** will be used in this coursepack to refer to a discrete variable which has only a finite (in practice quite small) number of possible values, which are known in advance. For example, a person's sex is typically coded simply as "Male" or "Female", with no other values. Similarly, the grouped income variable shown above is categorical, as every income corresponds to one of its four categories (note that it is the "rest" category 4 which guarantees that the variable does indeed cover all possibilities). Categorical variables are of separate interest because they are common and because some statistical methods are designed specifically for them. An example of a non-categorical discrete variable is the population of a country, which does not have a small, fixed set of possible values (unless it is again transformed into a grouped variable as in the income example above).

### Relationships between the two distinctions

The distinctions between variables with different measurement levels on one hand, and continuous and discrete variables on the other, are partially related. Essentially all nominal/ordinal-level variables are discrete, and almost all continous variables are interval-level variables. This leaves one further possibility, namely a discrete interval-level variable; the most common exam-

ple of this is a **count**, such as the number of children in a family or the population of a country. These connections are summarized in Table **??**.

Table 1.3: Relationships between the types of variables discussed in Section @ref(ss-intro-def-vartypes.

| | *Measurement level* **Nominal/ordinal** | *Measurement level* **Interval/ratio** |
|---|---|---|
| **Discrete** | Many<br>- Always **categorical**, i.e. having a fixed set of possible values (categories)<br>- If only two categories, variable is **binary** (**dichotomous**) | *Counts*<br>- If many different observed values, often treated as effectively continuous |
| **Continuous** | None | Many |

In practice the situation may be even simpler than this, in that the most relevant distinction is often between the following two cases:

1. Discrete variables with a small number of observed values. This includes both categorical variables, for which all possible values are known in advance, and variables for which only a small number of values were actually observed even if others might have been possible[1]. Such variables can be conveniently summarized in the form of tables and handled by methods appropriate for such tables, as described later in this coursepack. This group also includes all nominal variables, even ones with a relatively large number of categories, since methods for group 2. below are entirely inappropriate for them.

2. Variables with a large number of possible values. This includes all continuous variables and those interval-level or ordinal discrete variables which have so many values that it is pragmatic to treat them as effectively continuous.

Although there are contexts where we need to distinguish between types of variables more carefully than this, for practical purposes this simple distinction is often sufficient.

### 1.2.3 Description and inference

In the past, the subtitle of this course was "Description and inference". This is still descriptive of the contents of the course. These words refer to two different although related tasks of statistical analysis. They can be thought of as solutions to what might be called the "too much and not enough" problems with observed data. A set of data is "too much" in that it is very difficult to understand or explain the data, or to draw any conclusions from it, simply by staring at the numbers in a data matrix. Making much sense of even a small data matrix like the one in Table **??** is challenging, and the task becomes entirely impossible with bigger ones. There is thus a clear need for methods of statistical description:

- **Description**: summarizing some features of the data in ways that make them easily understandable. Such methods of description may be in the form of numbers or graphs.

[1] ESS Round 5: European Social Survey Round 5 Data (2010). Data file edition 2.0. Norwegian Social Science Data Services, Norway Data Archive and distributor of ESS data. The full data can be obtained from `http://ess.nsd.uib.no/ess/round5/`.

The "not enough" problem is that quite often the subjects in the data are treated as representatives of some larger group which is our real object of interest. In statistical terminology, the observed subjects are regarded as a **sample** from a larger **population**. For example, a pre-election opinion poll is not carried out because we are particularly interested in the voting intentions of the particular thousand or so people who answer the questions in the poll (the sample), but because we hope that their answers will help us draw conclusions about the preferences of all of those who intend to vote on election day (the population). The job of statistical inference is to provide methods for generalising from a sample to the population:

- **Inference**: drawing conclusions about characteristics of a population based on the data observed in a sample. The two main tools of statistical inference are **significance tests** and **confidence intervals**.

Some of the methods described on this course are mainly intended for description and others for inference, but many also have a useful role in both.

### 1.2.4 Association and causation

The simplest methods of analysis described on this course consider questions which involve only one variable at a time. For example, the variable might be the political party a respondent intends to vote for in the next general election. We might then want to know what proportion of voters plan to vote for the Labour party, or which party is likely to receive the most votes.

However, considering variables one at a time is not going to entertain us for very long. This is because most interesting research questions involve associations between variables. One way to define an association is that

- There is an **association** between two variables if knowing the value of one of the variables will help to predict the value of the other variable.

(A more careful definition will be given later.) Other ways of referring to the same concept are that the variables are "related" or that there is a "dependence" between them.

For example, suppose that instead of considering voting intentions overall, we were interested in *comparing* them between two groups of people, homeowners and people who live in rented accommodation. Surveys typically suggest that homeowners are more likely to vote for the Conservatives and less likely to vote for Labour than renters. There is then an association between the two (discrete) variables "type of accommodation" and "voting intention", and knowing the type of a person's accommodation would help us better predict who they intend to vote for. Similarly, a study of education and income might find that people with more education (measured by years of education completed) tend to have higher incomes (measured by annual income in pounds), again suggesting an association between these two (continuous) variables.

Sometimes the variables in an association are in some sense on an equal footing. More often, however, they are instead considered asymmetrically in that it is more natural to think of one of them as being used to predict the other. For example, in the examples of the previous paragraph it seems easier to talk about home ownership predicting voting intention than vice versa, and of level of education predicting income than vice versa. The variable used for prediction is then known as an **explanatory variable** and the variable to be predicted as the **response variable** (an alternative convention is to talk about **independent** rather than explanatory variables and **dependent** instead of response variables). The most powerful statistical techniques for analysing associations between explanatory and response variables are known as **regression** methods. They are by far the most important family of methods of quantitative data analysis.

On this course you will learn about the most important member of this family, the method of **linear regression**.

In the many research questions where regression methods are useful, it almost always turns out to be crucially important to be able to consider several different explanatory variables simultaneously for a single response variable. Regression methods allow for this through the techniques of **multiple regression**.

The statistical concept of association is closely related to the stronger concept of **causation**, which is at the heart of very many research questions in the social sciences and elsewhere. The two concepts are not the same. In particular, association is not *sufficient* evidence for causation, i.e. finding that two variables are statistically associated does not prove that either variable has a causal effect on the other. On the other hand, association is almost always *necessary* for causation: if there is no association between two variables, it is very unlikely that there is a direct causal effect between them. This means that analysis of associations is a necessary part, but not the only part, of the analysis of causal effects from quantitative data. Furthermore, statistical analysis of associations is carried out in essentially the same way whether or not it is intended as part of a causal argument. On this course we will mostly focus on associations. The kinds of additional arguments that are needed to support causal conclusions are based on information on the research design and the nature of the variables. They are discussed only briefly on this course, and at greater length on courses of research design such as MY400 (and the more advanced MY457, which considers design and analysis for causal inference together).

## 1.3 Outline of the course

We have now defined three separate distinctions between different problems for statistical analysis, according to (1) the types of variables involved, (2) whether description or inference is required, and (3) whether we are examining one variable only or associations between several variables. Different combinations of these elements require different methods of statistical analysis. They also provide the structure for the course, as follows:

- **Chapter ??**: Description for single variables of any type, and for associations between categorical variables.
- **Chapter ??**: Some general concepts of statistical inference.
- **Chapter ??**: Inference for associations between categorical variables.
- **Chapter ??**: Inference for single dichotomous variables, and for associations between a dichotomous explanatory variable and a dichotomous response variable.
- **Chapter ??**: More general concepts of statistical inference.
- **Chapter ??**: Description and inference for associations between a dichotomous explanatory variable and a continuous response variable, and inference for single continuous variables.
- **Chapter ??**: Description and inference for associations between any kinds of explanatory variables and a continuous response variable.
- **Chapter ??**: Some additional comments on analyses which involve three or more categorical variables.

As well as in Chapters **??** and **??**, general concepts of statistical inference are also gradually introduced in Chapters **??**, **??** and **??**, initially in the context of the specific analyses considered

in these chapters.

## 1.4 The use of mathematics and computing

Many of you will approach this course with some reluctance and uncertainty, even anxiety. Often this is because of fears about mathematics, which may be something you never liked or never learned that well. Statistics does indeed involve a lot of mathematics in both its algebraic (symbolical) and arithmetic (numerical) senses. However, the understanding and use of statistical concepts and methods can be usefully taught and learned even without most of that mathematics, and that is what we hope to do on this course. It is perfectly possible to do well on the course without being at all good at mathematics of the secondary school kind.

### 1.4.1 Symbolic mathematics and mathematical notation

Statistics *is* a mathematical subject in that its concepts and methods are expressed using mathematical formalism, and grounded in a branch of mathematics known as probability theory. As a result, heavy use of mathematics is essential for those who develop these methods (i.e. statisticians). However, those who only *use* them (i.e. you) can ignore most of it and still gain a solid and non-trivialised understanding of the methods. We will thus be able to omit most of the mathematical details. In particular, we will not show you how the methods are derived or prove theorems about them, nor do we expect you to do anything like that.

We will, however, use mathematical notation whenever necessary to state the main results and to define the methods used. This is because mathematics is the language in which many of these results are easiest to express clearly and accurately, and trying to avoid all mathematical notation would be contrived and unhelpful. Most of the notation is fairly simple and will be explained in detail. We will also interpret such formulas in English as well to draw attention to their most important features.

Another way of explaining statistical methods is through applied examples. These will be used throughout the course. Most of them are drawn from real data from research in a range social of social sciences. If you wish to find further examples of how these methods are used in your own discipline, a good place to start is in relevant books and research journals.

### 1.4.2 Computing

Statistical analysis involves also a lot of mathematics of the numerical kind, i.e. various calculations on the numbers in the data. Doing such calculations by hand or with a pocket calculator would be tedious and unenlightening, and in any case impossible for all but the smallest samples and simplest methods. We will mostly avoid doing that by leaving the drudgery of calculation to computers, where the methods are implemented in statistical software packages. This also means that you can carry out the analyses without understanding all the numerical details of the calculations. Instead, we can focus on trying to understand when and why certain methods of analysis are used, and learning to interpret their results.

A simple pocket calculator is still more convenient than a computer for some very simple calculations. You will also need one for this purpose in the examination, where computers are not allowed. Any such calculations required in the examination will be extremely simple to do (assuming you know what you are trying to do, of course). For more complex analyses, the exam questions will involve interpreting computer output rather than carrying out the calculations.

The homework questions that follow the computer classes contain examples of both of these types of questions.

The software package used in the computer classes of this course is called SPSS. There are other comparable packages, for example SAS, Minitab, Stata and R. Any one of them could be used for the analyses on this course, and the exact choice does not matter very much. SPSS is convenient for our purposes, because it is widely used, has a reasonably user-friendly menu interface, and is available on a cheap licence even for the personal computers of LSE students.

Sometimes you may see a phrase such as "SPSS course" used apparently as a synonym for "Statistics course". This makes as little sense as treating an introduction to Microsoft Word as a course on how to write good English. It is not possible to learn quantitative data analysis well by just sitting down in front of SPSS or any other statistics package and trying to figure out what all those menus are for. On the other hand, using SPSS to apply statistical methods to analyse real data is an effective way of strengthening the understanding of those methods *after* they have first been introduced in lectures. That is why this course has weekly computer classes.

The software-specific questions on how to carry out statistical analyses are typically of a lesser order of difficulty once the methods themselves are reasonably well understood. In other words, once you have a clear idea of what you want to do, finding out how to do it in SPSS tends not to be that difficult. For example, in the next chapter we will discuss the mean, one simple tool of descriptive statistics. Suppose that you then want to calculate the mean of a variable called *Age* in a data set. Learning how to do this in SPSS is then a matter of (1) finding the menu item where SPSS can be told to calculate a mean, (2) finding out which part of that menu is used to tell SPSS that you want the mean of *Age* specifically, and (3) finding the part of the SPSS output where the calculated mean of *Age* is reported. Instructions for steps like this for techniques covered on this course are given in the descriptions of the corresponding computer classes.

There are, however, some tasks which have more to do with specific software packages than with statistics in general. For example, the fact that SPSS has a menu interface, and the general style of those menus, need to be understood first. You also need to learn how to get data into SPSS in the first place, how to manipulate the data in various ways, and how to export output from the analyses to other packages. Some instructions on how to do such things are given in the first computer class. The introduction to the computer classes also includes details of some SPSS guidebooks and other sources of information which you may find useful if you want to know more about the program.

# Chapter 2

# Descriptive statistics

## 2.1 Introduction

This chapter introduces some common descriptive statistical methods. It is organised around two dichotomies:

- Methods that are used only for variables with small numbers of values, vs. methods that are used also or only for variables with many values (see the end of Section **??** for more on this distinction). The former include, in particular, descriptive methods for categorical variables, and the latter the methods for continuous variables.
- **Univariate** descriptive methods which consider only one variable at a time, vs. **bivariate** methods which aim to describe the association between *two* variables.

Section **??** describes univariate methods for categorical variables and Section **??** bivariate methods for cases where both variables are categorical. Sections **??** and **??** cover univariate methods which are mostly used for continuous variables. Section **??** lists some bivariate methods where at least one variable is continuous; these methods are discussed in detail elsewhere in the coursepack. The chapter concludes with some general guidelines for presentation of descriptive tables and graphs in Section **??**.

## 2.2 Example data sets

Two examples are used to illustrate the methods throughout this chapter:

*Example: Country data*

Consider data for 155 countries on three variables:

- The **region** where the country is located, coded as 1=Africa, 2=Asia, 3=Europe, 4=Latin America, 5=Northern America, 6=Oceania.
- A measure of the level of **democracy** in the country, measured on an 11-point scale from 0 (lowest level of democracy) to 10 (highest).
- Gross Domestic Product (**GDP**) per capita, in thousands of U.S. dollars.

Further information on the variables is given in the appendix to this chapter (Section **??**), together with the whole data set, shown in Table **??**.