

Module 1: Using Quantitative Data in Research: Concepts and Definitions

*Antony Fielding*¹
University of Birmingham & Centre for Multilevel Modelling

Rebecca Pilling
Centre for Multilevel Modelling

Contents

Introduction.....	2
C1.1 The uses of statistical analysis in research.....	3
C1.1.1 Why is there a need for detailed data collection?.....	3
C1.1.2 Quantitative methods and other approaches to social research	4
C1.1.3 The nature of social science relationships and evidence.....	5
C1.1.4 What does quantitative investigation do?.....	6
C1.2 Research design and generalisation.....	7
C1.2.1 Statistical generalisation	8
C1.2.2 Theoretical generalisation	9
C1.3 Data: units of analysis and statistical variables.....	11
C1.3.1 Example of data from the 2002 European Social Surveys	11
C1.3.2 Units of analysis.....	12
C1.3.3 Operational definition of research concepts	13
C1.3.4 Variables	13
C1.3.5 Naming variables.....	14
C1.3.6 Variable description	14
C1.3.7 Categories and values	17
C1.3.8 A classification scheme for variables	18
C1.3.9 Measurement error	21
C1.4 Data hierarchies.....	22

¹ With contributed material from Kelvyn Jones and Fiona Steele and extensive comment by Harvey Goldstein

There is an online quiz at the end of the module for you to test your understanding. To find the quiz:

EXAMPLE

From within the LEMMA learning environment

- Go down to the Lesson for **Module 1: Using Quantitative Data in Research**
- Click "[1.4 Data hierarchies](#)" to open Lesson 1.4
- Click  to open the first question

Introduction

This set of modules is not aimed at the complete beginner in statistics and quantitative analysis; we rather expect most users will have some familiarity with many basic ideas and some prior experience of traditional elementary statistical methods, but may need these ideas refreshing- and there may be some topics which users feel they have not yet fully understood. In Modules 1 and 2 therefore we present quite selective overviews of some crucial topics rather than re-invent a basic statistics course. We will initially rarely go into any technical detail and the presentation will be mainly on ideas supported by concrete examples. The focus is on key topics that will help to contextualise the ultimate task of learning about multilevel models.

For readers requiring a thorough preliminary coverage of basic statistical ideas or more in-depth treatment of some of the key ideas to be presented in this module and the next there are a number of on-line resources and texts that may be consulted, which are listed in the Resources section

In Module 1 we will look at quantitative research and how we collect data, in order to provide a firm foundation for the analyses covered in later modules (ultimately, of course, for multilevel modelling, introduced in Modules 4 & 5). In more detail, the aims of Module 1 are:

- To give a broad overview of how research questions might be answered through quantitative analysis. Such questions as the following are explored: How does quantitative analysis relate to other methods of inquiry? Why is it required and what sorts of evidence can it supply?
- To introduce the vocabulary of quantitative analysis and specify the common terminology to be used in later modules. Of particular importance is the operational definition of research concepts (how we get from real world characteristics to numbers in our data set) and how this leads to

observable variables at different levels of measurement.

- To introduce sources of data and concepts relating to how it may be possible to generalise results from samples of various kinds to the populations they are drawn from.
- To discuss how variables are defined, what different types there are, and how this may influence how they are analysed.
- To give some emphasis to certain ideas such as the nature of variability or the recognition of hierarchical units of analysis that are central to multilevel modelling.

C1.1 The uses of statistical analysis in research

This course presents various methods of analysis used in quantitative research. So we will first take a step back and consider the role of quantitative research as a method of inquiry: why might we want to collect and analyse quantitative data? We concentrate on what sort of questions quantitative analysis, and in particular modelling, can answer and what sort of data is required. We will also discuss briefly what steps we can take, in terms of the design of a study and during the analysis, to ensure we get valid and reliable results.

C1.1.1 Why is there a need for detailed data collection?

We start with an example of a research aim: to develop an understanding of the nature and extent of racial discrimination against staff working in the legal profession. You might wonder why we need to go to the trouble of carrying out this research - surely everyone knows that certain groups are discriminated against? The answer is that subjective 'conventional wisdom' may often be misleading. We need credible concrete evidence of the size and nature of such things as the discrimination effect, particularly because results can often surprise us. For instance, contrary to 'conventional wisdom', it may be that legislation has meant that white males are now the disadvantaged group, as a process of over-compensation has gone on. Though there may be incredulity at this idea, at least it is a possibility that is admissible and must be investigated.

There are numerous examples of such counter-intuitive quantitative research findings. A famous example is Stouffer et al's (1949)² analysis of the American soldier during World War II. He lists some aspects of what was then the prevailing

² Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Starr, S.A. and Williams, R.M. (1949). *The American Soldier*. Princeton University Press.

knowledge: for instance, that blacks are less ambitious to become officers and blacks from the South are even less ambitious. He then proceeds to show via detailed statistical evidence that the data point to conclusions that are completely and unequivocally to the contrary. Similarly and, it might seem, paradoxically it is also found that blacks stationed in the southern United States are more satisfied with their situation than those in the northern ones. Stouffer suggests that understanding the concept of relative deprivation is the key issue here. Blacks in the army compare themselves to blacks who are not in the army rather than to white soldiers in the army, but those stationed in the south compare themselves to local southern blacks.

The relations we study in the social sciences are often complex. Understanding them requires attention to detailed quantitative evidence.

C1.1.2 Quantitative methods and other approaches to social research

The motivation for our example research question concerning racial discrimination in the legal profession comes from the claim 'Lawyers not without blame in racist Britain' in a recent report by the 1990 Trust for Human Rights and Racial Equality³. This report says, among other things, that the highest incidence of claims to the London Race Discrimination Unit is generated by the legal profession itself. Another report draws attention to the under-representation of black people in firms and agencies in the Criminal Justice system and the general failures to promote suitably qualified black candidates⁴. Both reports comment on the lack of the kind of information which would be needed to enable a detailed assessment of the extent and implications of this possible institutional racism.

Very general questions are involved which might be tackled via a variety of research approaches. At one extreme we could use 'participant observation' and work in a legal firm. This would give the researcher a great deal of insight into the specific barriers that are being enacted during recruitment, promotion and reward for that particular firm. But, unless some agency is providing very generous funding for the research project and there is the luxury of a lot of time, it is unlikely that this intensive approach could be used in more than one or two firms. There is thus the caveat that any findings from this exercise, whilst very detailed, are highly specific to these few cases, which might have unique characteristics affecting the presence and extent of discrimination. We will therefore be faced with difficulties in generalising our findings and be confronted with those, in other firms for instance, who will argue that the results are not applicable to them.

Another possibility would be to undertake open-ended and interactive interviews, for example with those involved in choosing who to promote. These case studies will provide important information about how the processes of promotion work out

³ <http://www.blink.org.uk/pdescription.asp?key=1002&grp=16>

⁴ *Race, Discrimination and the Criminal Justice System*. Black Information Link, 21/5/2002: <http://www.blink.org.uk/docs/abpoa.pdf>

in practice, but even if we are exceptionally industrious we will still have a problem in generalising beyond the handful of cases that our resources will permit us to interview.

The ability to generalise is a reason for using a quantitative approach. We can collect large scale quantitative data on a considerable number of legal firms and/or their employees, and analyse these to discover what the scale and types of discrimination are and how it differs between firms. The strength of this approach is the weakness of the other approaches and *vice versa*. The more extensive survey of a large number of firms, provided attention is given to their selection through careful design, gives one the potential to apply findings to wider situations. On the other hand, by necessity, the nature of information collected on each is more limited than intensive case studies allow.

For this reason each type of approach may have a role in different phases of research. Starting with intensive interactive interviews at the outset, we can get some understanding of the realities of the situation. We might for instance note some black employees reported that they faced insults from fellow employees, or that they believed that they were underpaid relative to white colleagues doing similar work. This will allow some formal questions to be drawn up about these issues which, along with other questions, might then be put to respondents in a subsequent survey and data gathering exercise on a much larger number of legal firms. We could then analyse these data to provide evidence of the existence and/or extent and nature of these problems and possibly uncover factors affecting them. We might even then follow this with participant observation in firms that have been identified by the analysis as particularly discriminatory. Amongst firms where black employees perceive pay barriers can we study in detail the pay allocation machinery? For instance do particular individuals in the firm who have control over pay and promotions exhibit features making them discriminatory? Such a follow up exercise might provide added explanatory depth when it comes to understanding the issues involved.

In later modules we will see how we can use quantitative analysis to answer questions about explanations for the phenomena that we observe. Again, the possibility of scientific generalisation from the samples we use in our analysis to a broader population is an advantage of the quantitative over the qualitative approach, whilst the qualitative approach can be useful in suggesting possible explanations for testing.

C1.1.3 The nature of social science relationships and evidence

If we apply enough heat to water, it will eventually turn to steam. This is an example of a deterministic cause and effect relationship of a type that rarely arises in social science: we will not be able to say, for example, that a pupil who gets 83% in an exam taken at age 11 will get exactly 87% in an exam taken at age 16. Nonetheless, we are often trying to gather a range of detailed evidence that is supportive of causal-type relationships, and we will be able to say that pupils who do well at age 11 are more likely to do well at age 16. We will also be able to find a range of values for our pupil who scored 83%, within which it is very likely that

their exam score at 16 will lie.

A classic example arises in epidemiology. Many lung cancer victims have never smoked, i.e. smoking is not a necessary condition for the occurrence of lung cancer. Also there are heavy smokers, of up to 60 cigarettes a day over a period of many years, without any illness and thus smoking is not a sufficient condition. (Note that adding heat is not necessary to turn water into steam, but is usually sufficient!). The estimate is that one cigarette reduces your life *on average* by 11 minutes (*British Medical Journal*, 2000; 320:53). But of course, for a specific person smoking a specific cigarette, that cigarette will almost certainly not reduce their life by exactly 11 minutes. What we can say is that a reduction in smoking will reduce the *risk* of cancer. Another important point to note is that we cannot say this just on the basis of an observed link between smoking and cancer. For example, some other factor might be both causing a person to smoke more and increasing their risk of cancer. Then a reduction in smoking will not reduce the risk of cancer. Initially, the link we observe can only suggest to us that there is a possibility that smoking acts in a causal way on the risk of cancer. We then have to follow a process of testing and thus strengthening this claim in order to accept that smoking does indeed increase the risk of cancer.

In general, this process may involve animal studies or the discovery of plausible biological mechanisms and even possibly randomised experiments, although these are rare outside the constrained area of clinical trials (and would not be suitable in the example of smoking and cancer). But the kind of evidence for causality that we focus on here is *observational studies*, consisting of the measurement of naturally occurring data from the real world, which are then analysed using *statistical modelling*. This can provide us with a quantitative assessment of the plausibility of various possible explanations for the relationships that we observe.

We can thus use quantitative research to observe a link between two properties and then as a key part of amassing convincing evidence that one property causes the other. The same data may be used in each case (though it doesn't have to be), but when we attempt to establish whether a relationship is causal we will analyse the data in a different way and we will generally require some additional data, usually measurements of some more properties of the individuals we are studying (in the smoking example, these might include age, socioeconomic status, diet, and so on).

C1.1.4 What does quantitative investigation do?

Using our main example as an illustration, we summarise what quantitative approaches might allow us to do. In the modules to come we will introduce methods of statistical analysis that will enable us to answer questions like these, which we can group according to three 'modes' of quantitative investigation:

- **descriptive mode** (what are the mean salaries for different ethnic groups? Do these means change with the size of the firm?).

- **'hypothesis-evaluation' or inferential mode** (does being black result in a lower salary?). These are matters related to traditional ideas of '*statistical significance*'. Typically this will involve determining the size of the average effect after taking account of other factors. For example, in general black legal workers may be younger on average than their white colleagues. Salaries tend to increase with age and it may be that black workers earn on average less than white colleagues simply because they are younger, or on the other hand it may be that their age alone does not explain the difference in salaries. Quantitative analysis will help us decide which of these alternatives is best supported by the data.
- **predictive mode** (what salary would we expect a black employee aged 35 to earn in a large firm dealing with legal aid work?).

C1.2 Research design and generalisation

Data which lend themselves to research by quantitative analysis can arise from diverse sources, including nowadays large administrative databases from government agencies. However, often they will have arisen from a planned research design involving a *survey* or *experiment*. Carefully designed experiments are frequent in industry, agricultural research and in the physical and biological sciences and randomised controlled trials in medical and pharmaceutical research. In social sciences, although these kinds of design are sometimes used, surveys are more prevalent.

The design of trials, experiments or surveys can sometimes exhibit very complex features which we will not pursue with any vigour here.⁵ In general, however, research design is governed by an interest in the *generalisability* of the findings or results of analysis, beyond the particular set or sample of study units considered and the specific circumstances in which the data were collected. What can we say, for instance, about the size of differences between black and white salaries for the legal profession as a whole when we only have knowledge of this difference on a sample of legal employees? For instance, if in our sample we observe that black salaries are on average lower than white salaries, do we have the confidence to assert that this will also be the case across the whole legal profession? Or has it arisen from the way the sample has been selected? The ability to answer questions like this and to generalise beyond the sample is often referred to as *external validity*. When samples have been selected in certain ways, or we can make

⁵ Good but very brief introductions to the basic principles of sample design can be found in Chapters 23 and 24 by Peter Lynn in T. Greenfield (2002) *Research Methods for Postgraduates*, London: Arnold. A common design in social survey research is the multistage or clustered design. For instance, samples of teachers may be selected by a two-stage mechanism of first selecting a set of schools and then including all or a subsample of teachers within each school. This type of two-stage design will become very relevant for later modules where the two stages set up the two-level structure of a multilevel model.

This document is only the first few pages of the full version.

To see the complete document please go to learning materials and register:

<http://www.cmm.bris.ac.uk/lemma>

The course is completely free. We ask for a few details about yourself for our research purposes only. We will not give any details to any other organisation unless it is with your express permission.