2

Multilevel Theories, Multistage Sampling, and Multilevel Models

Phenomena and data sets in the social sciences often have a multilevel structure. This may be reflected in the design of data collection: simple random sampling is often not a very cost-efficient strategy, and multistage samples may be more efficient instead. This chapter is concerned with the reasons why it is important to take account of the clustering of the data, also called their multilevel structure, in the data analysis phase.

OVERVIEW OF THE CHAPTER

First we discuss how methods of inference failing to take into account the multilevel data structure may lead to erroneous conclusions, because independence assumptions are likely to be violated. The next two sections sketch the reasons for interest in a multilevel approach from the applications point of view. In many cases the multilevel data structure reflects essential aspects of the social (biological, etc.) world, and important research questions can be formulated about relations between variables at different layers in a hierarchical system. In this case the dependency of observations within clusters is of focal interest, because it reflects the fact that clusters differ in certain respects. In either case, the use of single-level statistical models is no longer valid. The fallacies to which their use can lead are described in the next chapter.

2.1 Dependence as a nuisance

Textbooks on statistics tell us that observations should be sampled *independently* of each other as standard. Thus the standard sampling design on which statistical models are based is simple random sampling with replacement from an infinite population: the result of one selection is independent of the result of any other selection, and all single units in the population have the same chances of being selected into the sample.

Textbooks on sampling, however, make it clear that there are more cost-efficient sampling designs, based on the idea that probabilities of selection should be known but do not have to be constant. One of those cost-efficient sampling designs is the *multistage sample*: the population of interest consists of subpopulations, also called *clusters*, and selection takes place via those subpopulations.

If there is only one subpopulation level, the design is a *two-stage sample*. Pupils, for instance, are grouped in schools, so the population of pupils consists of subpopulations of schools that contain pupils. Other examples are: families in neighborhoods, teeth in jawbones, animals in litters, employees in firms, and children in families. In a random two-stage sample, a random sample of the primary units (schools, neighborhoods, jawbones, litters, firms, families) is taken in the first stage, and then the secondary units (pupils, families, teeth, animals, employees, children) are sampled at random from the selected primary units in the second stage. A common mistake in research is to ignore the fact that the sampling scheme was a two-stage one, and to pretend that the secondary units were selected independently. The mistake in this case would be that the researcher is overlooking the fact that the secondary units were not sampled independently of each other: having selected a primary unit (e.g., a school) increases the chances of selection of secondary units (e.g., pupils) from that primary unit. In other words, the multistage sampling design leads to *dependent* observations, and failing to deal with this properly in the statistical analysis may lead to erroneous inferences. An example of the grossly inflated type I error rates that may then occur is given by Dorman (2008).

The multistage sampling design can be depicted graphically as in Figure 2.1. This shows a population that consists of 10 subpopulations, each containing 10 micro-units. A sample of 25% is taken by randomly selecting 5 out of 10 subpopulations and within these – again at random of course – 5 out of 10 micro-units.

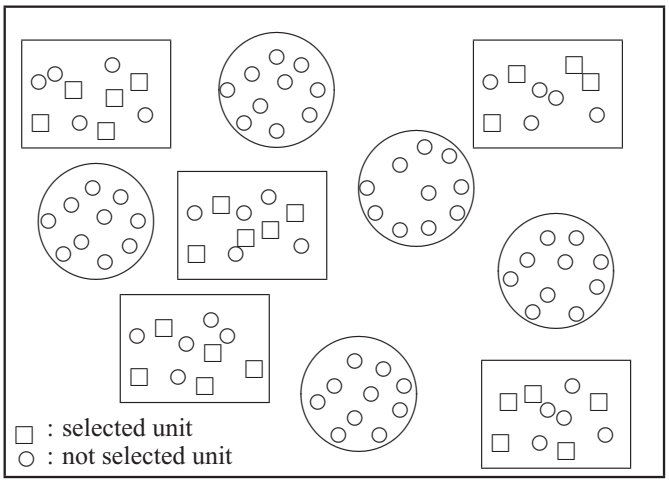

Figure 2.1: Multistage sample.

Multistage samples are preferred in practice, because the costs of interviewing or testing persons are reduced enormously if these persons are geographically or organizationally grouped. Such sample designs correspond to the organization of the social world. It is cheaper to travel to 100 neighbourhoods and interview 10 persons per neighbourhood on

their political preferences than to travel to 1,000 neighbourhoods and interview one person per neighbourhood. In the next chapters we will see how we can make adjustments to deal with these dependencies.

2.2 Dependence as an interesting phenomenon

The previous section implies that, if we want to make inferences on, for example, the earnings of employees in the for-profit sector, it is cost-efficient to use a multistage sampling design in which employees are selected via the firms in which they work. A common feature in social research, however, is that in many cases we wish to make inferences on the firms as well as on the employees. Questions that we seek to answer may be: Do employees in multinationals earn more than employees in other firms? Is there a relation between the performance of pupils and the experience of their teacher? Is the sentence differential between black and white suspects different between judges, and if so, can we find characteristics of judges to which this sentence differential is related? In this case a variable is defined at the primary unit level (firms, teachers, judges) as well as at the secondary unit level (employees, pupils, cases). Henceforth we will refer to primary units as *macro-level units* (or macro-units for short) and to secondary units as *micro-level units* (or micro-units for short). The micro level is called the *lower level* (first) and the macro level is called the *higher level* (second). For the time being, we will restrict ourselves to the two-level case, and thus to two-stage samples only. Table 2.1 gives a summary of the terminology.

Table 2.1: Summary of terms to describe units at either level in the two-level case.

macro-level units	micro-level units
macro-units	micro-units
primary units	secondary units
clusters	elementary units
level-two units	level-one units

Examples of macro-units and the micro-units nested within them are presented in Table 2.2. Most of the examples presented in the table have been dealt with in the text already. It is important to note that what is defined as a macro-unit or a micro-unit depends on the theory at hand. Teachers are nested within schools, if we study organizational effects on teacher burn-out then teachers are the micro-units and schools the macro-units. But when studying teacher effects on student achievement, teachers are the macro-units and students the micro-units. The same goes, *mutatis mutandis*, for neighborhoods and families (e.g., when studying the effects of housing conditions on marital problems), and for families and children (e.g., when studying effects of income on educational performance of siblings).

In all these instances the dependency of the observations on the micro-units within the macro-units is of focal interest. If we stick to the example of schools and pupils, then the dependency (e.g., in mathematics achievement of pupils within a school) may stem from:

Table 2.2: Some examples of units at the macro and micro level.

Macro level	Micro level
schools	teachers
classes	pupils
neighbourhoods	families
firms	employees
jawbones	teeth
families	children
litters	animals
doctors	patients
subjects	measurements
interviewers	respondents
judges	suspects

1. pupils within a school sharing the same school environment;
2. pupils within a school sharing the same teachers;
3. pupils within a school affecting each other by direct communication or shared group norms;
4. pupils within a school coming from the same neighborhood.

The more the achievement levels of pupils within a school are alike (as compared to pupils from other schools), the more likely it is that causes of the achievement have to do with the organizational unit (in this case, the school). Absence of dependency in this case implies absence of institutional effects on individual performance.

A special kind of nesting is defined by longitudinal data, represented in Table 2.2 as 'measurements within subjects'. The measurement occasions here are the micro-units and the subjects the macro-units. The dependence of the different measurements for a given subject is of primary importance in longitudinal data, but the following section on relations between variables defined at either level is not directly intended for the nesting structure defined by longitudinal data. Because of the special nature of this nesting structure, Chapter 15 is specifically devoted to it.

The models treated in this book are for situations where the dependent variable is at the lowest level. For models with nested data sets where the dependent variable is defined at a higher level one may consult Croon and van Veldhoven (2007), Lüdtke et al. (2008), and van Mierlo et al. (2009).

2.3 Macro-level, micro-level, and cross-level relations

In the study of hierarchical or multilevel systems, Lazarsfeld and Menzel (1971) made important distinctions between properties and propositions connected to the different levels.

In his summary of this work, Tacq (1986) distinguished between three kinds of propositions: on micro-units (e.g., 'employees have on average 4 effective working hours per day'; 'boys lag behind girls in reading comprehension'), on macro-units (e.g., 'schools have on average a budget of $20,000 to spend on resources'; 'in neighborhoods with bad housing conditions crime rates are above average'), or on macro–micro relations (e.g., 'if firms have a salary bonus system, the productivity of employees will be greater'; 'a child suffering from a broken family situation will affect the climate in the classroom').

Multilevel statistical models are always[1] called for if we are interested in propositions that connect variables defined at different levels, the micro and the macro, and also if a multistage sample design has been employed. The use of such a sampling design is quite obvious if we are interested in macro–micro relations, less obvious (but often necessary from a cost-effectiveness point of view) if micro-level propositions are our primary concern, and hardly obvious at all (but sometimes still applicable) if macro-level propositions are what we are focusing on. These three instances will be discussed below. To facilitate comprehension, following Tacq (1986) we use figures with the following conventions: a dotted line indicates that there are two levels; below the line is the micro level; above the line is the macro level; macro-level variables are denoted by capitals; micro-level variables are denoted by lower-case letters; and arrows denote presumed causal relations.

Multilevel propositions

Multilevel propositions can be represented as in Figure 2.2. In this example we are interested in the effect of the macro-level variable Z (e.g., teacher efficacy) on the micro-level variable y (e.g., pupil motivation), controlling for the micro-level variable x (e.g., pupil aptitude).

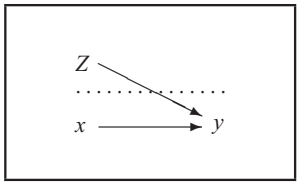

Figure 2.2: The structure of a multilevel proposition.

Micro-level propositions

Micro-level propositions are of the form indicated in Figure 2.3. In this case the line indicates that there is a macro level which is not referred to in the hypothesis that is put to the test, but which is used in the sampling design in the first stage. In assessing the strength of the relation between occupational status and income, for instance, respondents may have been selected for face-to-face interviews by zip-code area. This then may cause dependency (as a nuisance) in the data.

[1] As with any rule, there are exceptions. If the data set is such that for each macro-unit only one micro-unit is included in the sample, single-level methods still can be used.

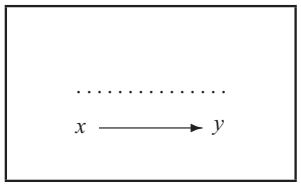

Figure 2.3: The structure of a micro-level proposition.

Macro-level propositions

Macro-level propositions are of the form of Figure 2.4. The line separating the macro level from the micro level seems superfluous here. When investigating the relation between the long-range strategic planning policy of firms and their profits, there is no multilevel situation, and a simple random sample may have been taken. When either or both variables are not directly observable, however, and have to be measured at the micro level (e.g., organizational climate measured as the average satisfaction of employees), then a two-stage sample is needed nevertheless. This is the case *a fortiori* for variables defined as aggregates of micro-level variables (e.g., the crime rate in a neighborhood).

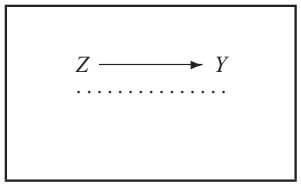

Figure 2.4: The structure of a macro-level proposition.

Macro–micro relations

The most common situation in social research is that macro-level variables are supposed to have a relation with micro-level variables. There are three obvious instances of macro-to-micro relations, all of which are typical examples of the multilevel situation (see Figure 2.5). The first case is the macro-to-micro proposition. The more explicit the religious norms in social networks, for example, the more conservative the views that individuals have on contraception. The second proposition is a special case of this. It refers to the case where there is a relation between Z and y, given that the effect of x on y is taken into account. The example given may be modified to: 'for individuals of a given educational level'. The last case in the figure is the *macro–micro-interaction*, also known as the cross-level interaction: the relation between x and y is dependent on Z. To put it another way, the relation between Z and y is dependent on x. The effect of aptitude on achievement, for instance, may be small in case of ability grouping of pupils within classrooms but large in ungrouped classrooms.

Next to these three situations there is the so-called emergent, or micro–macro, proposition (Figure 2.6). In this case, a micro-level variable x affects a macro-level variable Z (student achievement may affect teachers' experience of stress).

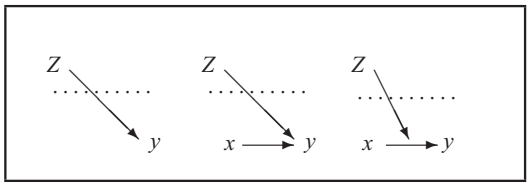

Figure 2.5: The structure of macro–micro propositions.

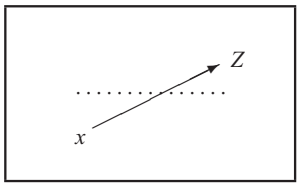

Figure 2.6: The structure of a micro–macro proposition.

It is of course possible to form combinations of the various examples given. Figure 2.7 contains a causal chain that explains through which micro-variables there is an association between the macro-level variables W and Z (cf. Coleman, 1990). As an example of this chain, we may be interested in why the qualities of a football coach affect his social prestige. The reason is that good coaches are capable of motivating their players, thus leading the players to good performance, thus to winning games, and this of course leads to more social prestige for the coach. Another instance of a complex multilevel proposition is the contextual effects proposition. For example, low socio-economic status pupils achieve less in classrooms with a low average aptitude. This is also a cross-level interaction effect, but the macro-level variable, average aptitude in the classroom, is now an aggregate of a micro-level variable.

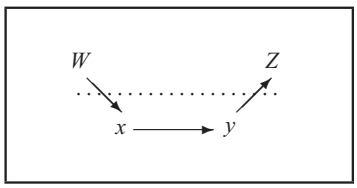

Figure 2.7: A causal macro–micro–micro–macro chain.

The methodological advances in multilevel modeling are now also leading to theoretical advances in contextual research: suitable definitions of context and 'levels', meaningful ways of aggregating variables to higher levels, conceptualizing and analyzing the interplay between characteristics of lower- and higher-level units. Some examples in various disciplines are the following. Following up on the initial work of Hauser (1970, 1974), in which he stated that group composition effects may be artifacts of underspecification of the micro-level model, Harker and Tymms (2004) discuss the issue of group composition effects in education. Sampson et al. (2002) give a review of theoretical work in the analysis of neighborhood effects. Diez-Roux (2000), Blakely and Woodward (2000), and O'Campo (2003) comment on advances along these lines in epidemiology and public health.

In the next chapters the statistical tools to handle multilevel structures will be introduced for outcome variables defined at the micro level.

2.4 Glommary

Multilevel data structures. Many data sets in the social sciences have a multilevel structure, that is, they constitute hierarchically nested systems with multiple levels. Much of our discussion focuses on two-level structures, but this can be generalized to three or more nested levels.

Sampling design. Often the multilevel nature of the social world leads to the practical efficiency of multistage samples. The population then consists of a nested system of subpopulations, and a nested sample is drawn accordingly. For example, when employing a random two-stage sample design, in the first stage a random sample of the primary units is taken, and in the second stage the secondary units are sampled at random from the selected primary units.

Levels. The levels are numbered such that the most detailed level is the first. For example, in a two-level structure of individuals nested in groups the individuals are called level-one units and the groups level-two units. (Note the different terminology compared to the words used in theories of survey sampling: in the preceding example, the primary units are the level-two units and the secondary units the level-one units.)

Units. The elements of a level are called units. Higher-level units are also called clusters. We talk about level-one units, level-two units, etc.

Dependence as a nuisance. Not taking account of the multilevel data structure, or of the multistage sampling design, is likely to lead to the use of statistical procedures in which independence assumptions are violated so that conclusions may be unfounded.

Dependence as an interesting phenomenon. The importance of the multilevel structure of social (biological, etc.) reality implies that research can often become more interesting when it takes account of the multilevel structure.

Multilevel propositions. Illustrations were given of scientific propositions involving multiple levels: micro-level propositions, macro-level propositions, macro–micro relations, cross-level interaction, and emergent propositions or micro–macro relations.