2nd Edition

MULTILEVEL ANALYSIS

An Introduction to Basic and Advanced Multilevel Modeling

Tom A B **SNIJDERS**
Roel J **BOSKER**

MULTILEVEL ANALYSIS

2nd Edition

MULTILEVEL ANALYSIS

An Introduction to Basic and Advanced Multilevel Modeling

Tom A B **SNIJDERS**
Roel J **BOSKER**

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

First edition published 1999
Reprinted 2002, 2003, 2004

SAGE Publications Ltd
1 Oliver’s Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
33 Pekin Street #02-01
Far East Square
Singapore 048763

Library of Congress Control Number: 2011926498

British Library Cataloguing in Publication data

A catalogue record for this book is available from the British Library

ISBN 978-1-84920-200-8
ISBN 978-1-84920-201-5

Typeset by C&M Digitals (P) Ltd, Chennai, India
Printed by MPG Books Group, Bodmin, Cornwall
Printed on paper from sustainable resources

Contents

Preface to the Second Edition

Multilevel analysis is a domain of data analysis that has been developing strongly before as well as after the publication of the first edition of our book in 1999. This second edition has been seriously revised. It contains more material, it has been updated and corrected, and a number of explanations were clarified.

The main new material is in three new chapters. A chapter was added about missing data, and another about the use of multilevel modeling for surveys with nonconstant inclusion probabilities ('survey weights'). Also a chapter was added in which three special techniques are briefly treated: Bayesian estimation, cluster-robust standard errors (the so-called sandwich standard error), and latent class models. The topics of these new chapters all belong to the 'advanced' part of the title. Among what was not covered in the first edition, these are some of the topics which we believe are most frequently encountered in the practice of multilevel research.

New material has been added also to existing chapters. The main example (starting in Chapter 4) has been renewed because the treatment of missing data in the old version was inadequate. Various other new examples also have been added. Further, there now is a more elaborate treatment of the combination of within-group evidence without using full-blown multilevel modeling (Section 3.7); more detailed considerations are discussed for choosing between fixed and random effects models (Section 4.3); diagnostic and comparative standard errors of posterior means are explained (Section 4.8.1); the treatment of tests for parameters of the random part was corrected and confidence intervals for these parameters are discussed (Sections 6.2 and 6.3); multiple membership models are treated in Chapter 13; and there has been an overhaul of the treatment of estimation methods for hierarchical generalized linear models in Chapter 17. Chapter 18 about multilevel software was totally rewritten, keeping it relatively short because this is the part of any textbook that ages most rapidly. Throughout all chapters new developments have been mentioned, pointers are given to the recent literature, various difficulties now are explained in more elaborate ways, and errors have been corrected.

All chapters (from the second on) now start by an overview, and are concluded (except for the last) by a 'glommary'. As every reader will know after reading this book, this is a summary of the main concepts treated in the chapter in a form akin to a glossary. Our intention is that these new elements will improve the didactical qualities of this textbook. Having said this, we think that the understanding of the book, or parts of it, may be further enhanced by going through the examples using the data that we made available (as far as this was allowed) at http://www.stats.ox.ac.uk/∼snijders/mlbook.htm. This website will also contain our comments on remarks made on the book by industrious readers, as well as our corrections for errors if any will be discovered.

We are very grateful for stimulating discussions (over the years or in the recent period of revising the text), comments on drafts of chapters, and help with software, to many people: Marnie Bertolet, sir David Cox, Roel de Jong, Jon Fahlander, Mark Huisman, Johan Koskinen, Catalina Lomos, Mayra Mascareño, Paulina Preciado, Roy Stewart, Anneke Timmermans, and Marijtje van Duijn. For help with new data sets we are grateful to some of these people and also to Hennie Brandsma, Simone Doolaard, Sonja Drobnič, Anja Knuver, Hans Kuyper, Sascha Peter, Stijn Ruiter, Greetje van der Werf, and Frank van Tubergen.

Tom Snijders
Roel Bosker
March, 2011

Preface to the First Edition

This book grew out of our teaching and consultation activities in the domain of multilevel analysis. It is intended for the absolute beginner in this field as well as for those who have already mastered the fundamentals and are now entering more complicated areas of application. The reader is referred to Section 1.2 for an overview of this book and for some reading guidelines.

We are grateful to various people from whom we got reactions on earlier parts of this manuscript and also to the students who were exposed to it and helped us realize what was unclear. We received useful comments from, and benefited from discussions about parts of the manuscript with, among others, Joerg Blasius, Marijtje van Duijn, Wolfgang Langer, Ralf Maslowski, and Ian Plewis. Moreover we would like to thank Hennie Brandsma, Mieke Brekelmans, Jan van Damme, Hetty Dekkers, Miranda Lubbers, Lyset Rekers-Mombarg and Jan Maarten Wit, Carolina de Weerth, Beate Völker, Ger van der Werf, and the Zentral Archiv (Cologne) who kindly permitted us to use data from their respective research projects as illustrative material for this book. We would also like to thank Annelies Verstappen-Remmers for her unfailing secretarial assistance.

Tom Snijders
Roel Bosker
June, 1999

1
Introduction

1.1 Multilevel analysis

Multilevel analysis is a methodology for the analysis of data with complex patterns of variability, with a focus on nested sources of such variability – pupils in classes, employees in firms, suspects tried by judges in courts, animals in litters, longitudinal measurements of subjects, etc. In the analysis of such data, it is usually illuminating to take account of the fact that each level of nesting is associated with variability that has a distinct interpretation. There is variability, for example, between pupils but also between classes, and one may draw incorrect conclusions if no distinction is made between these different sources of variability. Multilevel analysis is an approach to the analysis of such data, including the statistical techniques as well as the methodology for their use. The term 'multilevel analysis' is used mainly in the social sciences (in the wide sense: sociology, education, psychology, economics, criminology, etc.), but also in other fields such as the biomedical sciences. Our focus will be on the social sciences. Other terms, referring to the technical aspects, are hierarchical linear models, mixed models, and random coefficient models.

In its present form, multilevel analysis is a stream which has two tributaries: contextual analysis and mixed effects models. *Contextual analysis* is a development in the social sciences which has focused on the effects of the social context on individual behavior. Some landmarks before 1980 are the paper by Robinson (1950) which discussed the ecological fallacy (which refers to confusion between aggregate and individual effects), the paper by Davis et al. (1961) on the distinction between within-group and between-group regression, the volume edited by Dogan and Rokkan (1969), and the paper by Burstein et al. (1978) on treating regression intercepts and slopes on one level as outcomes on the higher level.

Mixed effects models are statistical models in the analysis of variance (ANOVA) and in regression analysis where it is assumed that some of the coefficients are fixed and others are random. This subject is too vast even to mention some landmarks. A standard reference book on random effects models and mixed effects models is Searle et al. (1992), Chapter 2 of which gives an extensive historical overview. The name 'mixed model' seems to have been used first by Eisenhart (1947).

Contextual modeling until about 1980 focused on the definition of appropriate variables to be used in ordinary least squares regression analysis. Until the 1980s the main focus in the development of statistical procedures for mixed models was on random effects (i.e., random differences between categories in some classification system) more than on random coefficients (i.e., random effects of numerical variables). Multilevel analysis as we now know it was formed by these two streams coming together. It was realized that, in contextual modeling, the individual and the context are distinct sources of variability, which should both be modeled as random influences. On the other hand, statistical methods and algorithms were developed that allowed the practical use of regression-type models with nested random coefficients. There was a cascade of statistical papers: Aitkin et al. (1981), Laird and Ware (1982), Mason et al. (1983), Goldstein (1986), Aitkin and Longford (1986), Raudenbush and Bryk (1986), de Leeuw and Kreft (1986), and Longford (1987) proposed and developed techniques for calculating estimates for mixed models with nested coefficients. These techniques, together with the programs implementing them which were developed by a number of these researchers or under their supervision, allowed the practical use of models of which until that moment only special cases were accessible for practical use. By 1986 the basis of multilevel analysis was established. The first textbook appeared (by Goldstein, now in its fourth edition) and was followed by a few others in the 1990s and many more in the 2000s. The methodology has been further elaborated since then, and has proved to be quite fruitful in applications in many fields. On the organizational side, there are stimulating centers such as the 'Multilevel Models Project' in Bristol with its Newsletter and its website http://www.mlwin.com/, and there is an active internet discussion group at http://www.jiscmail.ac.uk/lists/multilevel.html.

In the biomedical sciences mixed models were proposed especially for longitudinal data; in economics mainly for panel data (Swamy, 1971), the most common longitudinal data in economics. One of the issues treated in the economics literature was the pooling of cross-sectional and time series data (e.g., Maddala, 1971; Hausman and Taylor, 1981), which is closely related to the difference between within-group and between-group regressions. Overviews are given by Chow (1984) and Baltagi (2008).

A more elaborate history of multilevel analysis is presented in the bibliographical sections of Longford (1993) and in Kreft and de Leeuw (1998). For an extensive bibliography of the older literature, see Hüttner and van den Eeden (1995). A more recent overview of much statistical work in this area can be found in the handbook by de Leeuw and Meijer (2008a).

1.1.1 Probability models

The main statistical model of multilevel analysis is the hierarchical linear model, an extension of multiple linear regression to a model that includes nested random coefficients. This model is explained in Chapter 5 and forms the basis of most of this book.

There are several ways to argue why it makes sense to use a probability model for data analysis. In sampling theory a distinction is made between *design-based inference* and *model-based inference*. This is discussed further in Chapter 14. The former means that the researcher draws a probability sample from some finite population, and wishes to make inferences from the sample to this finite population. The probability model then follows from how the sample is drawn by the researcher. Model-based inference means that the researcher postulates a probability model, usually aiming at inference to some large and sometimes hypothetical population such as all English primary school pupils in the 2000s or all human adults living in a present-day industrialized culture. If the probability model is adequate then so are the inferences based on it, but checking this adequacy is possible only to a limited extent.

It is possible to apply model-based inference to data collected by investigating some entire research population, such as all 12-year-old pupils in Amsterdam at a given moment. Sometimes the question arises as to why one should use a probability model if no sample is drawn but an entire population is observed. Using a probability model that assumes statistical variability, even though an entire research population was investigated, can be justified by realizing that conclusions are sought which apply not only to the investigated research population but also to a wider population. The investigated research population is assumed to be representative of this wider population – for pupils who are older or younger, in other towns, perhaps in other countries. This is called a *superpopulation* in Chapter 14, where the relation between model-based and design-based inference is further discussed. Applicability of the statistical inference to such a wider population is not automatic, but has to be carefully argued by considering whether indeed the research population may be considered to be representative of the larger (often vaguely outlined) population. This is the 'second span of the bridge of statistical inference' as discussed by Cornfield and Tukey (1956).[1] The inference then is not primarily about a given delimited set of individuals but about social, behavioral, biological, etc., mechanisms and processes. The random effects, or residuals, playing a role in such probability models can be regarded as resulting from the factors that are not included in the explanatory variables used. They reflect the approximative nature of the model used. The model-based inference will be adequate to the extent that the assumptions of the probability model are an adequate reflection of the effects that are not explicitly included by means of observed variables.

As we shall see in Chapters 3–5, the basic idea of multilevel analysis is that data sets with a nesting structure that includes unexplained variability at each level of nesting, such as pupils in classes or employees in firms, are usually not adequately represented by the probability model of multiple linear regression analysis, but are often adequately represented by the hierarchical linear model. Thus, the use of the hierarchical linear model in multilevel analysis is in the tradition of model-based inference.

1.2 This book

This book is intended as an introductory textbook and as a reference book for practical users of multilevel analysis. We have tried to include all the main points that come up when applying multilevel analysis. Most of the data sets used in the examples, and corresponding commands to run the examples in various computer programs (see Chapter 18), are available on the website http://www.stats.ox.ac.uk/∼snijders/mlbook.htm.

After this introductory chapter, the book proceeds with a conceptual chapter about multilevel questions and a chapter on ways to treat multilevel data that are not based on the hierarchical linear model. Chapters 4–6 treat the basic conceptual ideas of the hierarchical linear model, and how to work with it in practice. Chapter 4 introduces the random intercept model as the primary example of the hierarchical linear model. This is extended in Chapter 5 to random slope models. Chapters 4 and 5 focus on understanding the hierarchical linear model and its parameters, paying only very limited attention to procedures

[1] We are indebted to Ivo Molenaar for this reference.

and algorithms for parameter estimation (estimation being work that most researchers delegate to the computer). Chapter 6 is concerned with testing parameters and specifying a multilevel model.

An introductory course on multilevel analysis could cover Chapters 1–6 and Section 7.1, with selected material from other chapters. A minimal course would focus on Chapters 4–6. The later chapters cover topics that are more specialized or more advanced, but important in the practice of multilevel analysis.

The text of this book is not based on a particular computer program for multilevel analysis. The last chapter, 18, gives a brief review of computer programs that can be used for multilevel analysis.

Chapters 7 (on the explanatory power of the model) and 10 (on model assumptions) are important for the interpretation of the results of statistical analyses using the hierarchical linear model. Researchers who have data sets with many missing values, or who plan to collect data sets that may run this type of risk, will profit from reading Chapter 9. Chapter 11 helps the researcher in setting up a multilevel study, and in choosing sample sizes at the various levels.

Some multilevel data sets come from surveys done according to a complex design, associated with survey weights reflecting the undersampling and oversampling of parts of the population. Ways to analyze such data sets using the hierarchical linear model are covered in Chapter 14.

Several methods and models have been developed that can sometimes be useful as additions or alternatives to the more commonly used methods for the hierarchical linear model. Chapter 12 briefly presents three of these: Bayesian procedures, the sandwich estimator for standard errors, and latent class models.

Chapters 8 and 13–17 treat various extensions of the basic hierarchical linear model that are useful in practical research. The topic of Chapter 8, heteroscedasticity (nonconstant residual variances), may seem rather specialized. Modeling heteroscedasticity, however, is easily done within the hierarchical linear model and can be very useful. It also allows model checks and model modifications that are used in Chapter 10. Chapter 13 treats data structures where the different sources of variability, the 'levels' of the multilevel analysis, are not nested but related in different ways: crossed classifications and multiple memberships. Chapter 15 is about longitudinal data, with a fixed occasion design (i.e., repeated measures data) as well as those with a variable occasion design, where the time moments of measurement may differ arbitrarily between subjects. This chapter indicates how the flexibility of the multilevel model gives important opportunities for data analysis (e.g., for incomplete multivariate or longitudinal data) that were unavailable earlier. Chapter 16 is about multilevel analysis for multivariate dependent variables. Chapter 17 describes the multilevel modeling of dichotomous, ordinal, and frequency data.

Each chapter starts with an overview and finishes with a summarizing glossary, which we have called a *glommary*. The glommaries can be consulted to gain rapid overviews of what is treated in the various chapters.

If additional textbooks are sought, one could consider the excellent introductions by Hox (2010) and Gelman and Hill (2007); Raudenbush and Bryk (2002), for an elaborate treatment of the hierarchical linear model; Longford (1993), Goldstein (2011), Demidenko (2004), and de Leeuw and Meijer (2008a) for more detailed mathematical background; and Skrondal and Rabe-Hesketh (2004) for further modeling, especially latent variable models.

1.2.1 Prerequisites

In order to read this textbook, a good working knowledge of statistics is required. It is assumed that you know the concepts of probability, random variable, probability distribution, population, sample, statistical independence, expectation (population mean), variance, covariance, correlation, standard deviation, and standard error. Furthermore, it is assumed that you know the basics of hypothesis testing and multiple regression analysis, and that you can understand formulas of the kind that occur in the explanation of regression analysis.

Matrix notation is used only in a few more advanced sections. These sections can be skipped without loss of understanding of other parts of the book.

1.2.2 Notation

The main notational conventions are as follows. Abstract variables and random variables are denoted by italicized capital letters, such as X or Y. Outcomes of random variables and other fixed values are denoted by italicized lower-case letters, such as x or z. Thus we speak about the variable X, but in formulas where the value of this variable is considered as a fixed, nonrandom value, it will be denoted by x. There are some exceptions to this, for example in Chapter 2 and in the use of the letter N for the number of groups ('level-two units') in the data.

The letter $\mathcal{E}$ is used to denote the *expected value*, or population average, of a random variable. Thus, $\mathcal{E}Y$ and $\mathcal{E}(Y)$ denote the expected value of Y. For example, if P_n is the fraction of tails obtained in n coin flips, and the coin is fair, then the expected value is $\mathcal{E}P_n = \frac{1}{2}$.

Statistical parameters are indicated by Greek letters. Examples are μ, σ^2, and β. The following Greek letters are used.

α	alpha
β	beta
γ	gamma
δ	delta
η	eta
θ	theta
λ	lambda
μ	mu
π	pi
ρ	rho
σ	sigma
τ	tau
φ	phi
χ	chi
ω	omega
Δ	capital Delta
Σ	capital Sigma
T	capital Tau
X	capital Chi

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.

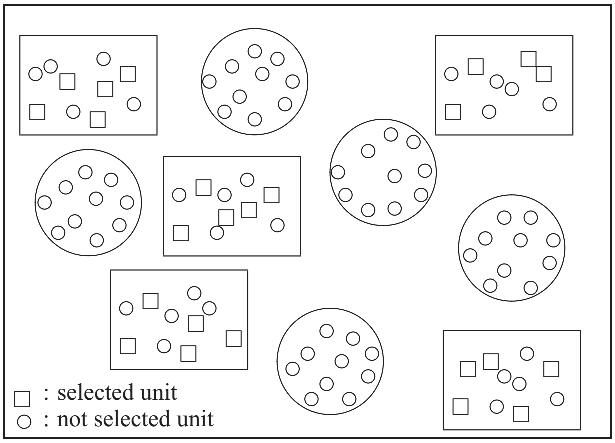

Figure 2.1: Multistage sample.

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).

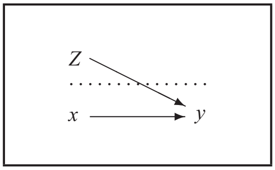

Figure 2.2: The structure of a multilevel proposition.

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.

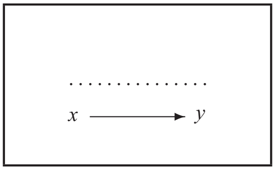

Figure 2.3: The structure of a micro-level proposition.

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).

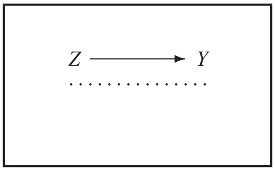

Figure 2.4: The structure of a macro-level proposition.

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).

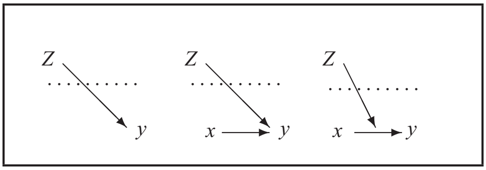

Figure 2.5: The structure of macro–micro propositions.

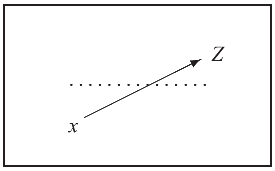

Figure 2.6: The structure of a micro–macro proposition.

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.

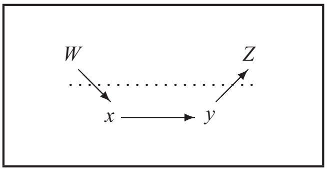

Figure 2.7: A causal macro–micro–micro–macro chain.

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.

3

Statistical Treatment of Clustered Data

Before proceeding in the next chapters to the explanation of the hierarchical linear model, the main statistical model for multilevel analysis, this chapter looks at approaches to analyzing multilevel data sets that are more elementary and do not use this model.

OVERVIEW OF THE CHAPTER

The chapter starts by considering what will happen if we ignore the multilevel structure of the data. Are there any instances where one may proceed with single-level statistical models although the data stem from a multistage sampling design? What kind of errors – so-called ecological fallacies – may occur when this is done? The rest of the chapter is devoted to some statistical methods for multilevel data that attempt to uncover the role played by the various levels without fitting a full-blown hierarchical linear model. First, we describe the intraclass correlation coefficient, a basic measure for the degree of dependency in clustered observations. Second, some simple statistics (mean, standard error of the mean, variance, correlation, reliability of aggregates) are treated for two-stage sampling designs. To avoid ecological fallacies it is essential to distinguish within-group from between-group regressions. These concepts are explained, and the relations are spelled out between within-group, between-group, and total regressions, and similarly for correlations. Finally, we mention some simple methods for combining evidence from several independent studies, or groups, in a combined test or a combined estimate.

3.1 Aggregation

A common procedure in social research with two-level data is to aggregate the micro-level data to the macro level. The simplest way to do this is to work with the averages for each macro-unit.

There is nothing wrong with aggregation in cases where the researcher is only interested in macro-level propositions, although it should be borne in mind that the reliability of an aggregated variable depends, *inter alia*, on the number of micro-level units in a macro-level unit (see later in this chapter), and thus will be larger for the larger macro-units than for the smaller ones. In cases where the researcher is interested in macro–micro or micro-level propositions, however, aggregation may result in gross errors.

The first potential error is the ‘shift of meaning’ (cf. Firebaugh, 1978; Hüttner, 1981). A variable that is aggregated to the macro level refers to the macro-units, not directly to the micro-units. The firm average of an employee rating of working conditions, for example, may be used as an index for ‘organizational climate’. This variable refers to the firm, not directly to the employees.

The second potential error with aggregation is the ecological fallacy (Robinson, 1950). A correlation between macro-level variables cannot be used to make assertions about micro-level relations. The percentage of black inhabitants in a neighborhood could be related to average political views in the neighborhood – for example, the higher the percentage of blacks in a neighborhood, the higher might be the proportion of people with extreme right-wing political views. This, however, does not give us any clue about the micro-level relation between race and political conviction. (The shift of meaning plays a role here, too. The percentage of black inhabitants is a variable that means something for the neighborhood, and this meaning is distinct from the meaning of ethnicity as an individual-level variable.) The ecological and other related fallacies are extensively discussed by Alker (1969), Diez-Roux (1998), and Blakely and Woodward (2000). King (1997), originally focusing on deriving correlates of individual voting behavior from aggregate data, describes a method for making inferences – within certain boundaries – at the micro level, when data are only available in aggregate form at the macro level.

The third potential error is the neglect of the original data structure, especially when some kind of analysis of covariance is to be used. Suppose one is interested in assessing between-school differences in pupil achievement after correcting for intake differences, and that Figure 3.1 depicts the true situation. The figure depicts the situation for five groups, for each of which we have five observations. The groups are indicated by the symbols □, $\times$, $+$, $\Diamond$, and $*$. The five group means are indicated by $\bullet$.

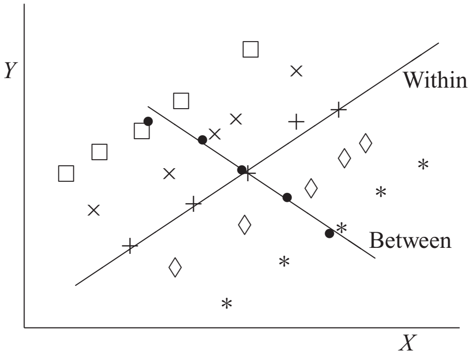

Figure 3.1: Micro-level versus macro-level adjustments.
(X, Y) values for five groups indicated by $*, \Diamond, +, \times,$ □ ; group averages by $\bullet$.

Now suppose the question is whether the differences between the groups on the variable Y, after adjusting for differences on the variable X, are substantial. The micro-level approach, which adjusts for the within-group regression of Y on X, will lead to the regression line with positive slope. In this picture, the micro-units from the group that have the □

symbol are all above the line, whereas those from the group with the $*$ symbol are all below the regression line. The micro-level regression approach will thus lead us to conclude that the five groups do differ given that an adjustment for X has been made.

Now suppose we were to aggregate the data, and regress the average $\bar{Y}$ on the average $\bar{X}$. The averages are depicted by $\bullet$. This situation is represented in the graph by the regression line with negative slope. The averages of all groups are almost exactly on the regression line (the observed average $\bar{Y}$ can be almost perfectly predicted from the observed average $\bar{X}$), thus leading us to the conclusion that there are almost no differences between the five groups after adjusting for the average $\bar{X}$.

Although the situation depicted in the graph is an idealized example, it clearly shows that working with aggregate data 'is dangerous at best, and disastrous at worst' (Aitkin and Longford, 1986, p. 42). When analyzing multilevel data, without aggregation, the problem described in this section can be dealt with by distinguishing between the within-group and the between-group regressions. This is worked out in Sections 3.6, 4.6, and 10.2.1.

The last objection to aggregation is that it prevents us from examining potential cross-level interaction effects of a specified micro-level variable with an as yet unspecified macro-level variable. Having aggregated the data to the macro level one cannot examine relations such as whether the sentence differential between black and white suspects is different between judges, when allowance is made for differences in seriousness of crimes. Or, to give another example, whether the effect of aptitude on achievement, present in the case of whole-class instruction, is smaller or even absent in the case of ability grouping of pupils within classrooms.

3.2 Disaggregation

Now suppose that we treat our data at the micro level. There are two situations:

1. we also have a measure of a variable at the macro level, next to the measures at the micro level;
2. we only have measures of micro-level variables.

In situation (1), disaggregation leads to 'the miraculous multiplication of the number of units'. To illustrate, suppose a researcher is interested in the question of whether older judges hand down more lenient sentences than younger judges. A two-stage sample is taken: in the first stage ten judges are sampled, and in the second stage for each judge ten trials are sampled (in total there are thus 10 × 10 = 100 trials). One might disaggregate the data to the level of the trials and estimate the relation between the experience of the judge and the length of the sentence, without taking into account that some trials involve the same judge. This is like pretending that there are 100 independent observations, whereas in actual fact there are only 10 independent observations (the 10 judges). This shows that disaggregation and treating the data as if they are independent implies that the sample size is dramatically exaggerated. For the study of between-group differences, disaggregation often leads to serious risks of committing type I errors (asserting on the basis of the observations that there is a difference between older and younger judges whereas in the population of judges there is no such relation). On the other hand, when studying within-group differences, disaggregation often leads to unnecessarily conservative tests (i.e., type

I error probabilities that are too low); this is discussed in detail in Moerbeek et al. (2003) and Berkhof and Kampen (2004).

If measures are taken only at the micro level, analyzing the data at the micro level is a correct way to proceed, as long as one takes into account that observations within a macro-unit may be correlated. In sampling theory, this phenomenon is known as the design effect for two-stage samples. If one wants to estimate the average management capability of young managers, while in the first stage a limited number of organizations (say, 10) are selected and within each organization five managers are sampled, one runs the risk of making an error if (as is usually the case) there are systematic differences between organizations. In general, two-stage sampling leads to the situation that the 'effective' sample size that should be used to calculate standard errors is less than the total number of cases, the latter being given here by the 50 managers. The formula will be presented in one of the next sections.

Starting with Robinson's (1950) paper on the ecological fallacy, many papers have been written about the possibilities and dangers of cross-level inference, that is, methods to conclude something about relations between micro-units on the basis of relations between data at the aggregate level, or conclude something about relations between macro-units on the basis of relations between disaggregated data. Discussions and many references are given by Pedhazur (1982, Chapter 13), Aitkin and Longford (1986), and Diez-Roux (1998). Our conclusion is that if the macro-units have any meaningful relation to the phenomenon under study, analyzing only aggregated or only disaggregated data is apt to lead to misleading and erroneous conclusions. A multilevel approach, in which within-group and between-group relations are combined, is more difficult but much more productive. This approach requires, however, assumptions to be specified about the way in which macro- and micro-effects are put together. The present chapter presents some multilevel procedures that are based on only a minimum of such assumptions (e.g., the additive model of equation (3.1)). Later chapters in this book are based on a more elaborate model, the so-called hierarchical linear model, which since about 1990 has been the most widely accepted basis for multilevel analysis.

3.3 The intraclass correlation

The degree of resemblance between micro-units belonging to the same macro-unit can be expressed by the *intraclass correlation coefficient*. The use of the term 'class' is conventional here and refers to the macro-units in the classification system under consideration. There are, however, several definitions of this coefficient, depending on the assumptions about the sampling design. In this section we assume a two-stage sampling design and infinite populations at either level. The macro-units will also be referred to as *groups*.

A relevant model here is the *random effects ANOVA* model.[1] Denoting by Y_{ij} the outcome value observed for micro-unit i within macro-unit j, this model can be expressed as

$$Y_{ij} = \mu + U_j + R_{ij}, \tag{3.1}$$

[1] This model is also known in the statistical literature as the one-way random effects ANOVA model and as Eisenhart's Type II ANOVA model. In multilevel modeling it is known as the empty model, and is treated further in Section 4.4.

where μ is the population grand mean, U_j is the specific effect of macro-unit j, and R_{ij} is the residual effect for micro-unit i within this macro-unit. In other words, macro-unit j has the 'true mean' $\mu + U_j$, and each measurement of a micro-unit within this macro-unit deviates from this true mean by some value R_{ij}. Units differ randomly from one another, which is reflected in the fact that U_j is a random variable and the name 'random effects model'. Some units have a high true mean, corresponding to a high value of U_j, others have a true mean close to average, and still others a low true mean. It is assumed that all variables are independent, the group effects U_j having population mean 0 and population variance τ^2 (the *population between-group variance*), and the residuals having mean 0 and variance σ^2 (the *population within-group variance*). For example, if micro-units are pupils and macro-units are schools, then the within-group variance is the variance within the schools about their true means, while the between-group variance is the variance between the schools' true means. The total variance of Y_{ij} is then equal to the sum of these two variances,

$$\mathrm{var}(Y_{ij}) = \tau^2 + \sigma^2.$$

The number of micro-units within the jth macro-unit is denoted by n_j. The number of macro-units is N, and the total sample size is $M = \sum_j n_j$.

In this situation, the intraclass correlation coefficient ρ_{I} can be defined as

$$\rho_{\mathrm{I}} = \frac{\text{population variance } \textit{between} \text{ macro-units}}{\text{total variance}} = \frac{\tau^2}{\tau^2 + \sigma^2}. \tag{3.2}$$

This is the proportion of variance that is accounted for by the group level. This parameter is called a correlation coefficient because it is equal to the correlation between values of two randomly drawn micro-units in the same, randomly drawn, macro-unit. Hedges and Hedberg (2007) report on a large variety of studies of educational performance in American schools, and find that values often range between 0.10 and 0.25.

It is important to note that the population variance between macro-units is not directly reflected by the observed variance between the means of the macro-units (the observed between-macro-unit variance). The reason is that in a two-stage sample, variation between micro-units will also show up as extra observed variance between macro-units. It is indicated below how the observed variance between cluster means must be adjusted to yield a good estimator for the population variance between macro-units.

Example 3.1 ***Random data.***
Suppose we have a series of 100 observations as in the random digits in Table 3.1. The core part of the table contains the random digits. Now suppose that each row in the table is a macro-unit, so that for each macro-unit we have observations on 10 micro-units. The averages of the scores for each macro-unit are in the last column. There seem to be large differences between the randomly constructed macro-units, if we look at the variance in the macro-unit averages (which is 105.7). The total observed variance between the 100 micro-units is 814.0. Suppose the macro-units were schools, the micro-units pupils, and the random digits test scores. According to these two observed variances we might conclude that the schools differ considerably with respect to their average test scores. We know in this case, however, that in 'reality' the macro-units differ only by chance.

The following subsections show how the intraclass correlation can be estimated and tested. For a review of various inference procedures for the intraclass correlation we refer to Donner (1986). An extensive overview of many methods for estimating and testing the within-group and between-group variances is given by McCulloch and Searle (2001).

Table 3.1: Data grouped into macro-units (random digits from Glass and Stanley, 1970, p. 511).

j	Scores Y_{ij} for micro-units (random digits)										Average $\bar{Y}_{.j}$
01	60	36	59	46	53	35	07	53	39	49	43.7
02	83	79	94	24	02	56	62	33	44	42	51.9
03	32	96	00	74	05	36	40	98	32	32	44.5
04	19	32	25	38	45	57	62	05	26	06	31.5
05	11	22	09	47	47	07	39	93	74	08	35.7
06	31	75	15	72	60	68	98	00	53	39	51.1
07	88	49	29	93	82	14	45	40	45	04	48.9
08	30	93	44	77	44	07	48	18	38	28	42.7
09	22	88	84	88	93	27	49	99	87	48	68.5
10	78	21	21	69	93	35	90	29	12	86	53.4

3.3.1 Within-group and between-group variance

We continue to refer to the macro-units as groups. To disentangle the information contained in the data about the population between-group variance and the population within-group variance, we consider the *observed variance between groups* and the *observed variance within groups*. These are defined as follows. The mean of macro-unit j is denoted by

$$\bar{Y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij},$$

and the overall mean is

$$\bar{Y}_{..} = \frac{1}{M} \sum_{j=1}^{N} \sum_{i=1}^{n_j} Y_{ij} = \frac{1}{M} \sum_{j=1}^{N} n_j \bar{Y}_{.j}.$$

The observed variance within group j is given by

$$S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2 .$$

This number will vary from group to group. To have one parameter that expresses the within-group variability for all groups jointly, one uses the observed within-group variance, or pooled within-group variance. This is a weighted average of the variances within the various macro-units, defined as

$$\begin{aligned} S^2_{\text{within}} &= \frac{1}{M-N} \sum_{j=1}^{N} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2 \qquad (3.3) \\ &= \frac{1}{M-N} \sum_{j=1}^{N} (n_j - 1) S_j^2 . \end{aligned}$$

If model (3.1) holds, the expected value of the observed within-group variance is exactly equal to the population within-group variance:

$$\text{Expected variance } \textit{within} = \mathcal{E}S^2_{\text{within}} = \sigma^2 . \tag{3.4}$$

The situation for the between-group variance is a little more complicated. For equal group sizes n_j, the observed between-group variance is defined as the variance between the group means,

$$S^2_{\text{between}} = \frac{1}{N-1}\sum_{j=1}^{N}(\bar{Y}_{.j} - \bar{Y}_{..})^2 . \tag{3.5}$$

For unequal group sizes, the contributions of the various groups need to be weighted. The following formula uses weights that are useful for estimating the population between-group variance:

$$S^2_{\text{between}} = \frac{1}{\tilde{n}(N-1)}\sum_{j=1}^{N} n_j(\bar{Y}_{.j} - \bar{Y}_{..})^2 . \tag{3.6}$$

In this formula, $\tilde{n}$ is defined by

$$\tilde{n} = \frac{1}{N-1}\left\{M - \frac{\sum_j n_j^2}{M}\right\} = \bar{n} - \frac{s^2(n_j)}{N\,\bar{n}}, \tag{3.7}$$

where $\bar{n} = M/N$ is the mean sample size and

$$s^2(n_j) = \frac{1}{N-1}\sum_{j=1}^{N}(n_j - \bar{n})^2$$

is the variance of the sample sizes. If all n_j have the same value, then $\tilde{n}$ also has this value. In this case, S^2_{between} is just the variance of the group means, given by (3.5).

It can be shown that the total observed variance is a combination of the within-group and the between-group variances, expressed as follows:

$$\begin{aligned} \text{observed } \textit{total} \text{ variance} &= \frac{1}{M-1}\sum_{j=1}^{N}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_{..})^2 \\ &= \frac{M-N}{M-1}S^2_{\text{within}} + \frac{\tilde{n}(N-1)}{M-1}S^2_{\text{between}} . \end{aligned} \tag{3.8}$$

The complications with respect to the between-group variance arise from the fact that the micro-level residuals R_{ij} also contribute, albeit to a minor extent, to the observed between-group variance. Statistical theory tells us that the expected between-group variance is given by

$$\begin{aligned} &\text{Expected observed variance } \textit{between} \\ &\qquad = \text{True variance } \textit{between} + \text{Expected sampling error variance}. \end{aligned}$$

More specifically, the formula is

$$\mathcal{E}S^2_{\text{between}} = \tau^2 + \frac{\sigma^2}{\tilde{n}} \tag{3.9}$$

(cf. Hays (1988, Section 13.3) for the case with constant n_j and Searle et al. (1992, Section 3.6) for the general case), which holds provided that model (3.1) is valid. The second term in this formula becomes small when $\tilde{n}$ becomes large. Thus for large group sizes, the expected observed between variance is practically equal to the true between variance. For small group sizes, however, it tends to be larger than the true between variance due to the random differences that also exist between the group means.

In practice, we do not know the population values of the between and within macro-unit variances; these have to be estimated from the data. One way of estimating these parameters is based on formulas (3.4) and (3.9). From the former it follows that the population within-group variance, σ^2, can be estimated without bias by the observed within-group variance:

$$\hat{\sigma}^2 = S^2_{\text{within}}. \tag{3.10}$$

From the combination of the last two formulas it follows that the population between-group variance, τ^2, can be estimated without bias by taking the observed between-group variance and subtracting the contribution that the true within-group variance makes, on average, according to (3.9), to the observed between-group variance:

$$\hat{\tau}^2 = S^2_{\text{between}} - \frac{S^2_{\text{within}}}{\tilde{n}}. \tag{3.11}$$

(Another expression is given in (3.14).) This expression can take negative values. This happens when the difference between group means is less than would be expected on the basis of the within-group variability, even if the true between-group variance τ^2 were 0. In such a case, it is natural to estimate τ^2 as 0.

It can be concluded that the split between the observed within-group variance and observed between-group variance does not correspond precisely to the split between the within-group and between-group variances in the population: the observed between-group variance reflects the population between-group variance plus a little of the population within-group variance.

The intraclass correlation is estimated according to formula (3.2) by

$$\hat{\rho}_{\text{I}} = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2}. \tag{3.12}$$

(Formula (3.15) gives another, equivalent, expression.) The standard error of this estimator in the case where all group sizes are constant, $n_j = n$, is given by

$$\text{S.E.}(\hat{\rho}_{\text{I}}) = (1 - \rho_{\text{I}})(1 + (n-1)\rho_{\text{I}})\sqrt{\frac{2}{n(n-1)(N-1)}}. \tag{3.13}$$

This formula was given by Fisher (1958, Section 39) and by Donner (1986, equation (6.1)), who also gives the (quite complicated) formula for the standard error for the case of variable group sizes. Donner and Wells (1986) compare various ways to construct confidence intervals for the intraclass correlation coefficient.

The estimators given above are so-called analysis of variance or ANOVA estimators. They have the advantage that they can be represented by explicit formulas. Other much used estimators are those produced by the maximum likelihood (ML) and residual maximum likelihood (REML) methods (cf. Section 4.7). For equal group sizes, the ANOVA estimators are the same as the REML estimators (Searle et al., 1992). For unequal group sizes, the ML and REML estimators are slightly more efficient than the ANOVA estimators. Multilevel software can be used to calculate the ML and REML estimates.

Example 3.2 ***Within- and between-group variability for random data.***
For our random digits table of the earlier example the observed between variance is $S^2_{\text{between}} = 105.7$. The observed variance within the macro-units can be computed from formula (3.8). The observed total variance is known to be 814.0 and the observed between variance is given by 105.7. Solving (3.8) for the observed within variance yields $S^2_{\text{within}} = (99/90) \times (814.0 - (10/11) \times 105.7) = 789.7$. Then the estimated true variance within the macro-units is also $\hat{\sigma}^2 = 789.7$. The estimate for the true between macro-units variance is computed from (3.11) as $\hat{\tau}^2 = 105.7 - (789.7/10) = 26.7$. Finally, the estimate of the intraclass correlation is $\hat{\rho}_{\mathrm{I}} = 26.7/(789.7 + 26.7) = 0.03$. Its standard error, computed from (3.13), is 0.06.

3.3.2 Testing for group differences

The intraclass correlation as defined by (3.2) can be zero or positive.[2] A statistical test can be performed to investigate whether a positive value for this coefficient could be attributed to chance. If it may be assumed that the within-group deviations R_{ij} are normally distributed, one can use an exact test for the hypothesis that the intraclass correlation is 0, which is the same as the null hypothesis that there are no group differences, or the true between-group variance is 0. This is just the F-test for a group effect in the one-way analysis of variance, which can be found in any textbook on ANOVA. The test statistic can be written as

$$F = \frac{\tilde{n} S^2_{\text{between}}}{S^2_{\text{within}}},$$

and it has an F distribution with $N - 1$ and $M - N$ degrees of freedom if the null hypothesis holds.

Example 3.3 ***The F-test for the random data set.***
For the data of Table 3.1, $F = (10 \times 105.7)/789.7 = 1.34$ with 9 and 90 degrees of freedom. This value is far from significant ($p > 0.10$). Thus, there is no evidence of true between-group differences.

Statistical computer packages usually give the F-statistic and the within-group variance, S^2_{within}. From this output, the estimated population between-group variance can be calculated by

$$\hat{\tau}^2 = \frac{S^2_{\text{within}}}{\tilde{n}}(F - 1) \qquad (3.14)$$

[2] In a data set it is possible for the estimated intraclass correlation coefficient to be negative. This is always the case, for example, for group-centered variables. In a population satisfying model (3.1), however, the population intraclass correlation cannot be negative.

and the estimated intraclass correlation coefficient by

$$\hat{\rho}_{\mathrm{I}} = \frac{F - 1}{F + \tilde{n} - 1}, \tag{3.15}$$

where $\tilde{n}$ is given by (3.7). If $F < 1$, it is natural to replace both of these expressions by 0. These formulas show that a high value for the F-statistic will lead to large estimates for the between-group variance as well as the intraclass correlation, but that the group sizes, as expressed by $\tilde{n}$, moderate the relation between the test statistic and the parameter estimates.

If there are covariates, it often is relevant to test whether there are group differences in addition to those accounted for by the effect of the covariates. This is achieved by the usual F-test for the group effect in an analysis of covariance (ANCOVA). Such a test is relevant because it is possible that the ANOVA F-test does not demonstrate any group effects, but that such effects do emerge when controlling for the covariates (or vice versa). Another check on whether the groups make a difference can be carried out by testing the group-by-covariate interaction effect. These tests can be found in textbooks on ANOVA and ANCOVA, and they are contained in the well-known general-purpose statistical computer programs.

So, to test whether a given nesting structure in a data set calls for multilevel analysis, one can use standard ANOVA techniques. In addition to testing for the main group effect, it is also advisable to test for group-by-covariate interactions. If there is neither evidence for a main effect nor for interaction effects involving the group structure, then the researcher may leave aside the nesting structure and analyze the data by single-level methods such as ordinary least squares ('OLS') regression analysis. This approach to testing for group differences can be employed whenever the number of groups is not too large for the computer program being used. If there are too many groups, however, the program will refuse to do the job. In such a case it will still be possible to carry out the tests for group differences that are treated in the following chapters, following the logic of the hierarchical linear model. This will require the use of statistical multilevel software.

3.4 Design effects in two-stage samples

In the design of empirical investigations, the determination of sample sizes is an important decision. For two-stage samples, this is more complicated than for simple ('one-stage') random samples. An elaborate treatment of this question is given in Cochran (1977). This section gives a simple approach to the precision of estimating a population mean, indicating the basic role played by the intraclass correlation. We return to this question in Chapter 11.

Large samples are preferable in order to increase the precision of parameter estimates, that is, to obtain tight confidence intervals around the parameter estimates. In a simple random sample the standard error of the mean is related to the sample size by the formula

$$\text{standard error} = \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}. \tag{3.16}$$

This formula can be used to indicate the required sample size (in a simple random sample) if a given standard error is desired.

When using two-stage samples, however, the clustering of the data should be taken into account when determining the sample size. Let us suppose that all group sizes are equal, $n_j = n$ for all j. Then the (total) sample size is Nn. The *design effect* is a number that indicates how much the sample size in the denominator of (3.16) is to be adjusted because of the sampling design used. It is the ratio of the variance of estimation obtained with the given sampling design to the variance of estimation obtained for a simple random sample from the same population, supposing that the total sample size is the same. A large design effect implies a relatively large variance, which is a disadvantage that may be offset by the cost reductions implied by the design. The design effect of a two-stage sample with equal group sizes is given by

$$\text{design effect} = 1 + (n - 1)\,\rho_I . \tag{3.17}$$

This formula expresses the fact that, from a purely statistical point of view, a two-stage sample becomes less attractive as ρ_I increases (clusters become more homogeneous) and as the group size n increases (the two-stage nature of the sampling design becomes stronger).

Suppose, for example, we were studying the satisfaction of patients with the treatments provided by their doctors. Furthermore, let us assume that some doctors have more satisfied patients than others, leading to a ρ_I of 0.30. The researchers used a two-stage sample, since that is far cheaper than selecting patients simply at random. They first randomly selected 100 doctors, from each chosen doctor selected five patients at random, and then interviewed each of these. In this case the design effect is $1 + (5 - 1) \times 0.30 = 2.20$. When estimating the standard error of the mean, we no longer can treat the observations as independent of each other. The effective sample size, that is, the equivalent total sample size that we should use in estimating the standard error, is equal to

$$N_{\text{effective}} = \frac{Nn}{\text{design effect}}, \tag{3.18}$$

in which N is the number of selected macro-units. For our example we find $N_{\text{effective}} = (100 \times 5)/2.20 = 227$. So the two-stage sample with a total of 500 patients here is equivalent to a simple random sample of 227 patients.

One can also derive the total sample size using a two-stage sampling design on the basis of a desired level of precision, assuming that ρ_I is known, and fixing n because of budgetary or time-related considerations. The general rule is: this required sample size increases as ρ_I increases and it increases with the number of micro-units one wishes to select per macro-unit. Using (3.17) and (3.18), this can be derived numerically from the formula

$$N_{\text{ts}} = N_{\text{srs}} + N_{\text{srs}}\,(n - 1)\,\rho_I .$$

The quantity N_{ts} in this formula refers to the total desired sample size when using a two-stage sample, whereas N_{srs} refers to the desired sample size if one had used a simple random sample.

In practice, ρ_I is unknown. However, it often is possible to make an educated guess about it on the basis of earlier research.

In Figure 3.2, N_{ts} is graphed as a function of n and ρ_I (0.1, 0.2, 0.4, and 0.6, respectively), and taking $N_{\text{srs}} = 100$ as the desired sample size for an equally informative simple random sample.