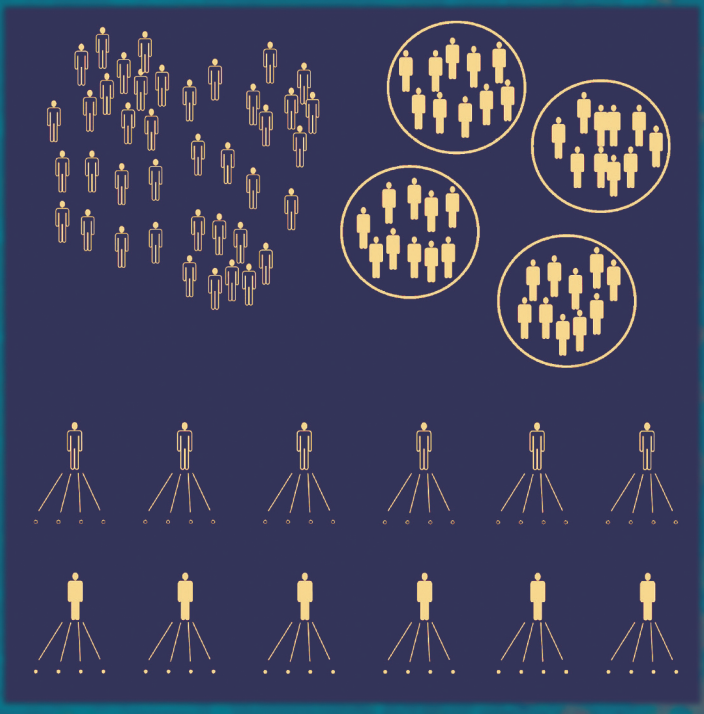
Chapman & Hall/CRC
Interdisciplinary Statistics Series

Power Analysis of Trials with Multilevel Data



Mirjam Moerbeek
Steven Teerenstra

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Power Analysis of Trials with Multilevel Data

CHAPMAN & HALL/CRC

Interdisciplinary Statistics Series

Series editors: N. Keiding, B.J.T. Morgan, C.K. Wikle, P. van der Heijden

Published titles

AGE-PERIOD-COHORT ANALYSIS: NEW MODELS, METHODS, AND EMPIRICAL APPLICATIONS Y. Yang and K. C. Land

ANALYSIS OF CAPTURE-RECAPTURE DATA R. S. McCrea and B. J. T. Morgan

AN INVARIANT APPROACH TO STATISTICAL ANALYSIS OF SHAPES S. Lele and J. Richtsmeier

ASTROSTATISTICS G. Babu and E. Feigelson

BAYESIAN ANALYSIS FOR POPULATION ECOLOGY R. King, B. J. T. Morgan, O. Gimenez, and S. P. Brooks

BAYESIAN DISEASE MAPPING: HIERARCHICAL MODELING IN SPATIAL EPIDEMIOLOGY, SECOND EDITION A. B. Lawson

BIOEQUIVALENCE AND STATISTICS IN CLINICAL PHARMACOLOGY S. Patterson and B. Jones

CLINICAL TRIALS IN ONCOLOGY, THIRD EDITION S. Green, J. Benedetti, A. Smith, and J. Crowley

CLUSTER RANDOMISED TRIALS R.J. Hayes and L.H. Moulton

CORRESPONDENCE ANALYSIS IN PRACTICE, SECOND EDITION M. Greenacre

DESIGN AND ANALYSIS OF QUALITY OF LIFE STUDIES IN CLINICAL TRIALS, SECOND EDITION D.L. Fairclough

DYNAMICAL SEARCH L. Pronzato, H. Wynn, and A. Zhigljavsky

FLEXIBLE IMPUTATION OF MISSING DATA S. van Buuren

GENERALIZED LATENT VARIABLE MODELING: MULTILEVEL, LONGITUDINAL, AND STRUCTURAL EQUATION MODELS A. Skrondal and S. Rabe-Hesketh

GRAPHICAL ANALYSIS OF MULTI-RESPONSE DATA K. Basford and J. Tukey

INTRODUCTION TO COMPUTATIONAL BIOLOGY: MAPS, SEQUENCES, AND GENOMES M. Waterman

MARKOV CHAIN MONTE CARLO IN PRACTICE W. Gilks, S. Richardson, and D. Spiegelhalter

MEASUREMENT ERROR AND MISCLASSIFICATION IN STATISTICS AND EPIDEMIOLOGY: IMPACTS AND BAYESIAN ADJUSTMENTS P. Gustafson

MEASUREMENT ERROR: MODELS, METHODS, AND APPLICATIONS J. P. Buonaccorsi

MEASUREMENT ERROR: MODELS, METHODS, AND APPLICATIONS J. P. Buonaccorsi

Published titles

MENDELIAN RANDOMIZATION: METHODS FOR USING GENETIC VARIANTS IN CAUSAL ESTIMATION S.Burgess and S.G. Thompson

META-ANALYSIS OF BINARY DATA USINGPROFILE LIKELIHOOD D. Böhning, R. Kuhnert, and S. Rattanasiri

POWER ANALYSIS OF TRIALS WITH MULTILEVEL DATA M. Moerbeek and S. Teerenstra

STATISTICAL ANALYSIS OF GENE EXPRESSION MICROARRAY DATA T. Speed

STATISTICAL AND COMPUTATIONAL PHARMACOGENOMICS R. Wu and M. Lin

STATISTICS IN MUSICOLOGY J. Beran

STATISTICS OF MEDICAL IMAGING T. Lei

STATISTICAL CONCEPTS AND APPLICATIONS IN CLINICAL MEDICINE J. Aitchison, J.W. Kay, and I.J. Lauder

STATISTICAL AND PROBABILISTIC METHODS IN ACTUARIAL SCIENCE P.J. Boland

STATISTICAL DETECTION AND SURVEILLANCE OF GEOGRAPHIC CLUSTERS P. Rogerson and I. Yamada

STATISTICS FOR ENVIRONMENTAL BIOLOGY AND TOXICOLOGY A. Bailer and W. Piegorsch

STATISTICS FOR FISSION TRACK ANALYSIS R.F. Galbraith

VISUALIZING DATA PATTERNS WITH MICROMAPS D.B. Carr and L.W. Pickle

Chapman & Hall/CRC
Interdisciplinary Statistics Series

Power Analysis of Trials with Multilevel Data

Mirjam Moerbeek

Utrecht University, The Netherlands

Steven Teerenstra

Radboud University Medical Center, The Netherlands

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20150417

International Standard Book Number-13: 978-1-4987-2990-1 (eBook - PDF)

Contents

List of figures

List of tables

Preface

What is the most effective method to lose body weight: a diet, a diet combined with physical exercise or a diet combined with a peer pressure group? Do special training programs increase unemployed people's chances of finding jobs on the labor market? Does a hormone treatment have an effect on the psychosocial functioning of children with growth retardation? Such questions are asked daily, not only by scientists but also by the general public. The similarity of these questions is that they focus on effective methods to delay or prevent disease or some unwanted and undesirable behavior, opinion or condition.

To compare various alternative methods in an experimental setting, it is important to seek a design that results in the highest statistical power at the lowest costs. It is common knowledge that the power level of a statistical test increases with increasing sample size. A sample size that is too small may result in not being able to detect an effect of a treatment, while an excessive sample size may be a waste of time and money of both the researchers and the trial's participants. It is therefore important to carefully calculate the required sample size before a trial is conducted.

Many textbooks and computer programs are of aid to the researcher who is planning a trial. These books and programs mainly apply to trials where the outcome measurements on disease, behavior or condition of a given subject are uncorrelated to those of other subjects in the trial. This assumption is very likely to be violated in trials where subjects are nested within groups, such as in multicenter clinical trials with patients nested within clinics and school-based smoking prevention interventions with pupils nested within schools. In the latter example, a pupil's smoking behavior will be influenced not only by the treatment condition he or she is assigned to, but also by the smoking behavior of other pupils in the same school, the teachers' smoking behavior and the school's policy toward smoking.

Ignoring the nested data structure may result in a study that is either under- or overpowered. Moreover, while the sample size affects a trial's power level, the allocation of units is also of importance. In the school-based smoking intervention, for instance, one has to decide whether to sample many schools and include just a few pupils per school, or to sample few schools and include many pupils per school.

In the past decades many journal papers that focus on statistical power analysis and optimal design for trials with nested data structures have appeared. It is now time to compile the published findings in these papers in

one single book, and to present a related computer program to perform the power calculations. This is the book that you currently hold in your hands.

The two authors have traveled a long way before we were able to write this book. During this journey various people have been of invaluable support to our careers in the field of applied statistics.

Mirjam Moerbeek. My journey started in spring 1996, when I was appointed PhD researcher at the Department of Methodology and Statistics at Maastricht University, the Netherlands. Under the supervision of Martijn Berger and Gerard van Breukelen I wrote my PhD thesis on the design and analysis of multilevel intervention studies. Part of this research was done in collaboration with Weng Kee Wong and conducted at the Department of Biostatistics at the University of California at Los Angeles. During my PhD research, I frequented meetings of the Netherlands' Multilevel Modeling group, which were organized by Tom Snijders and Cora Maas. Both of them may be considered pioneers of multilevel modeling in the Netherlands.

After receiving my PhD, I continued my research on optimal design and power analysis at the Department of Methodology and Statistics at Utrecht University, the Netherlands. Peter van der Heijden encouraged me to apply for a prestigious research grant from the Netherlands Organization of Scientific Research (NWO). I was awarded a Veni grant in 2003 to extend my research on optimal design for trials with multilevel and longitudinal data, and a Vidi and Aspasia grant in 2008 to study optimal designs for discrete-time survival analysis and to write this book.

The Department of Methodology and Statistics at Utrecht University is the best place to conduct research like this because it has an enthusiastic group of researchers in the field of multilevel analysis, among whom are Joop Hox, Rens van de Schoot and Leoniek Wijngaards-de Meij. Together we organize the biennial International Conference on Multilevel Analysis.

Part of the research in this book was done by former PhD students of mine: Elly Korendijk studied the effects of a misspecification of the intraclass correlation coefficient in the design phase of a cluster randomized trial and Esther Oomen-de Hoop studied the design and analysis of stepped-wedge cluster randomized trials. Charlotte Rietbergen, a former master student of mine, studied crossover in cluster randomized trials. Sander van Schie, also a former master student, studied sample size re-estimation in cluster randomized trials and prepared an overview of papers with estimates of the intraclass correlation coefficient. Katarzyna Jóźwiak wrote the SPA-ML computer program to perform many of the sample size and power calculations that are presented in this book.

Steven Teerenstra. What I like about statisticians is their diversity of origins. My interest in statistics arose when I was looking for a more practically oriented job after finishing my appointment as a PhD student in mathematics in Nijmegen. At that time (2002) I had a freelance job in the Radboud

University Nijmegen Medical Center to build a research database and I was asked to also think about the statistical analysis of the data. I contacted a former colleague of mine, Wiebe Pestman, who had already travelled the road from mathematics to statistics. He gave me not only statistical advice but also suggested I become a statistician. Without formal statistical training or working experience, my first application outside Nijmegen was not successful. However, opportunities were closer by than I thought. After attending a talk at the department of Epidemiology, Biostatistics and HTA in Nijmegen (now the Department for Health Evidence), I had a chat with Gerhard Zielhuis, who told me that the biostatistics group was looking for a candidate. Under guidance of George Borm, my journey in biostatistics began: hands-on and steep. Because much of my consultancy involved cluster randomized trials, George encouraged me to do research in this topic. A choice that I enjoy till today, because the techniques also gave me insight in other fields of statistics such as longitudinal data and meta-analysis.

While becoming acquainted with the literature in the field, I came across Mirjam's papers. I saw that we had common interests and contacted her to collaborate, which led to our joint endeavors. A research grant by NWO allowed me to work with Anouk Spijker to investigate the efficiency of ANCOVA analysis of cluster randomized trials. With a grant from the Radboud University Nijmegen Medical Center, I engaged Esther Oomen-de Hoop as a PhD student to investigate practical solutions for cluster randomized trials with few clusters.

Mirjam Moerbeek, Utrecht, the Netherlands
Steven Teerenstra, Nijmegen, the Netherlands

1

Introduction

The scientific community performs tens of thousands of experiments each year to improve the treatment of diseases and psychological and social conditions with the aim to increase quality of life and life expectancy. Examples include substance use prevention and cessation interventions, trials that compare cognitive behavioral group therapy to pharmacotherapy for the treatment of social phobia, and trials that evaluate the effects of growth hormones on the psychosocial functioning of children with growth retardation.

The process of data collection can take up to several decades in the case of long-term interventions. In addition, efforts should be paid to good data analysis, including the use of an appropriate statistical model and proper treatment of missing data arising from drop-out and non-response. It is therefore of utmost importance to plan an experiment in such a way that the best treatments are selected with highest probability.

In the design phase of a trial, the treatments to be compared are selected, and choices are made on eligibility criteria, the duration of the study and the best training of professionals who are delivering the treatments to their patients or clients. Another important choice in the design phase is the required number of subjects, as the probability to detect a significant difference between treatment conditions depends on the chosen sample size. This probability is called the statistical power and the calculation of the required sample size to achieve a desired power level is called a power analysis.

In many trials in the social and biomedical sciences, the data have a so-called hierarchical structure, meaning that subjects are nested within groups. For such trials, not only the total number of subjects needs to be determined, but also their allocation over the groups. Should few groups with many subjects each be sampled, or many groups with few subjects each?

Similar questions may be asked for the design of longitudinal trials with repeated measurements over time that are nested within subjects. Should few subjects be sampled and should they be measured often, or should many subjects be sampled and they be measured just a few times?

Guidelines for sample sizes for trials with hierarchical data structures have been studied extensively in the last two decades. The aim of this book is to provide formulae to perform power calculations to social and biomedical researchers and to statisticians working in these fields of science. This first chapter serves as an introduction to basic concepts with respect to experimentation, hierarchical data structures, and study design and further elucidates the

need for power calculations. A further description of the aim and contents of the book is given at the end of this chapter.

1.1 Experimentation

This section aims to provide an overview of the basic concepts of experimentation: controlling, randomization, stratification or blocking and matching, replication and blinding. Although these concepts appear obvious nowadays, their formal introduction was initiated less than a hundred years ago by the work of Fisher in the field of agriculture in the 1920s and 1930s (Fisher, 1926, 1935). The first modern randomized controlled trial in health care research is generally considered to be that of the Medical Research Council (1948), that compared the effect of streptomycin and bed rest to that of bed rest only on the treatment of pulmonary tuberculosis. The first randomized controlled trial in social research appears to have been conducted in the late 1920s (Forsetlund, Chalmers, & Bjørndal, 2007). For a more thorough introduction to experimentation the reader is referred to Jadad and Enkin (2007) and Torgerson and Torgerson (2008).

To evaluate the effect of a new type of therapy, surgery, teaching method or any other type of treatment, the measurements on outcome variables of subjects in the new treatment group have to be compared to those receiving a standard treatment, no treatment at all, or those who are assigned to a waiting list. The latter groups of subjects are often called control groups, and the corresponding treatment condition is called the control condition. Including a control is important since it enables the researcher to compare the performance of the new treatment to the performance of a standard treatment or no treatment at all. Without the control group, a researcher is unable to tell whether the performance of the new treatment is better than, worse than or equal to that of the standard or no treatment. In other words, the control serves as a benchmark and enables the estimation of the treatment effect, which is the difference in performance of the two treatment conditions on the outcome variable.

If one wishes to generalize the findings to all subjects within the population, a random draw should be made from this population. To achieve a fair comparison of treatment conditions, the treatment groups should be as similar as possible at baseline with respect to all measured and unmeasured covariates, which are variables that are supposed to have an effect on the outcome. For instance, as parents' smoking behavior has an effect on adolescent smoking behavior, it is important that the baseline percentages of adolescents whose parents smoke are equal over the treatment conditions in a smoking prevention intervention that targets adolescents.

For large sample sizes, this can be achieved by randomly assigning adoles-

cents to treatment conditions and the design is called a completely randomized design. For small sample sizes, random assignment does not guarantee balance with respect to parent smoking behavior. This can be resolved by stratifying adolescents with respect to their parents' smoking behavior: one stratum consisting of adolescents who have one or two parents who smoke, while adolescents in the other stratum do not have a parent who smokes. Randomization to treatment conditions is then done within strata and the design is called a stratified randomized design. Alternative terminology is to use blocking, blocks and randomized block design wording instead of stratifying, strata and stratified randomized design. The difference is that the terminology of blocking, blocks and randomized block design is more often used for factors that are generally determinable and often controllable in the experiment and for which the sample size can be determined upfront, and called blocking factors (e.g., temperature).

The terms stratifying, strata and stratified randomized design apply more often to factors that cannot be determined upfront (e.g., concomitant medication use). Often the distinction is not clear-cut. An option that is more useful when multiple covariates are present is the matched-pair design. With this design, pairs of subjects who are as similar as possible with respect to all covariates are formed, and randomization is done within each pair. Special algorithms are available to balance treatment groups with respect to multiple covariates; see for instance G. F. Borm, Hoogendoorn, Den Heijer, and Zielhuis (2005). If such balancing methods have not performed well, or if these methods were not used at all, then the method of statistical control can be used to correct for between-group differences. This is achieved by including important covariates as predictor variables in a regression or analysis of covariance model.

Whichever strategy is chosen to prevent imbalance or to correct for it, it is important that relevant covariates are defined in the planning of a trial to be able to measure them in the data collection phase. It is therefore necessary to search the literature for relevant covariates prior to data collection.

Several replications should be made to estimate the variability of the error terms in a statistical model. That is, more than one subject should be available within each combination of treatment condition and covariate values. In a trial that randomizes to treatment conditions within gender, for instance, at least two males and two females are required in the control condition, and another two males and two females in the experimental condition. Replication should not be confused with taking repeat measurements on the same subjects. If repeated measurements over time are taken on each subject, then time should be included as a predictor in the statistical model.

In an ideal case, the trial is double-blind, meaning that neither the researchers nor the subjects know to which treatment condition each subject is assigned. Only a third party has access to the key that identifies to which condition each subject is randomized. With single blinding only the study participants are unaware to which treatment condition they are assigned. Double

blinding is a means to lessen observer bias that occurs when the results are influenced by conscious or unconscious bias on the part of the observer. Such bias occurs when observers tend to adjust their scores on a subject's outcome variables to their expectations or judgments of the (known) treatment condition to which a subject is assigned.

Another reason for double blinding is to reduce selection bias which may occur when recruiters know beforehand the treatment to which a subject is assigned. The risk of both biases is illustrated by the finding that the outcomes and characteristics of included participants may be different in blinded trials than in unblinded trials (Veerus, Fischer, Hakama, & Hemminki, 2012) and hence blinding may influence both internal and external validity.

Unfortunately, blinding is not always an option. Double blinding is often used in pharmaceutical trials where different substances are compared. The pills or injections containing these substances should be as similar with respect to size, color, smell, weight and so forth as possible. In social and behavioral trials, treatment often relies on interpersonal interactions, such as risk-reduction sessions, peer pressure groups and training programs and blinding is often not possible.

The degree to which a researcher has control over the environment in which the experiment is conducted determines whether the experiment is a laboratory or field experiment. In a laboratory (pure) experiment, the control over the environment is maximal and all subjects are exposed to the same influences expect for the treatment condition to which they are assigned. Consider as an example a trial that investigates the effect of the opponent's ethnicity on the aggressiveness of players of dual-player computer games. Participants are assigned to an opponent of the same or different ethnicity and all other sources of variation are kept under the experimenter's control. Thus, the participants play the same computer game in rooms with the same background music, temperature, lighting, accessibility to food and drinks and so forth.

Laboratory experiments are often artificial and doubts on generalization of results to the real world exist. In a field experiment, the daily life environment is treated as the laboratory. Randomization to treatments is under experimental control but the experimenter does not have an influence on uncontrolled events. This is not considered a problem as long as both treatment groups are equally exposed to uncontrolled events. However, a problem arises when the one treatment group is exposed to uncontrolled events to a higher degree than the other group and/or when the uncontrolled events interact with treatment condition. In both situations, the external factors influence the one group more than the other, thereby threatening the validity of the study. Consider as an example the effects of exercise on the reduction of body weight in a controlled trial. A difference in body weight between both treatment groups can only be attributed to the effects of exercise if the calorie intake is equal across groups.

1.1.1 Problems with random assignment

Although randomization is a strong tool to achieve treatment groups that are equal with respect to relevant covariates, it also has some drawbacks. In the ideal case, randomization is done by the researcher or an independent statistician, but in practice it is often done by those who control access to potential participants, such as general practitioners, therapists and school principals. They often tend to assign the participants they consider most deserving or most likely to benefit to the new treatment group, a process which is called selection bias.

People who are recruited to participate in an experiment are often only willing to do so if they are assigned to the interesting and promising new treatment. Some of them will even try to force the person performing the randomization to assign them to the new treatment, resulting in a treatment group that consists of more demanding or more assertive subjects than the control condition, a situation that is most undesirable when the outcome variable is related to these qualities of character.

When participants cannot influence the randomization procedure, they can still undermine the study. For instance, those assigned to the control condition can try to benefit as much from the new treatment as possible by contacting participants in the new treatment group. This is especially a threat if the new treatment condition consists of oral or written information to improve lifestyles, and less so when the new condition consists of medication or surgery. This process is called control group contamination since information on the new treatment somehow leaks to participants in the control group. In addition, drop-out rates among those in the control group are often higher than those in the new treatment group and disappointment may be common after allocation to the control condition (Lindström, Sundberg-Petersson, Adami, & Tönnesen, 2010).

Resentment of those in the control condition can be lessened by using a placebo, which is an irrelevant treatment in the control group, such as a sugar pill in pharmaceutical trials. As an alternative, one can inform the participants in the control group on the necessity of random assignment and the use of a control group and assure them of having one of the more desirable treatments at the end of the trial, provided these treatments have a relevant effect and do not show harmful side effects.

A related problem occurs when participants in the new treatment condition do not wish to participate if they find this new treatment undesirable, if they expect its effects are negligible or if they expect unwanted or harmful side effects. For instance, parents who smoke might not be interested in their children participating in a smoking prevention intervention. This problem may be solved to only include those participants who have expressed their willingness to participate. As is obvious, this may induce problems with respect of generalization to the general population.

In some cases randomization is not possible due to ethical aspects. For

instance, a general practitioner might not be willing to let only half of his patients benefit from the new and promising treatment. In other cases, one cannot force subjects to adhere to the treatment assigned. For instance, one cannot force subjects to smoke or not in a trial on the effects of smoking on lung cancer. In such cases, treatments can only be compared on basis of a quasi-experimental design. With this design, the experimenter cannot randomly assign subjects to treatment conditions and must rely on existing differences between people with respect to the variable of interest. The pitfall is that this variable may be correlated with covariates that also have an effect on the outcome and an exhaustive literature review to detect all such covariates is therefore required.

The sample size formulae that are presented in the next chapters are valid for experimental designs where randomization is done in such a way that the treatment groups are comparable with respect to covariates. In Section 5.2 the effects of matching and pre-stratification on power are treated in further detail.

1.2 Hierarchical data structures

In many experiments in the social and biomedical sciences, the data have a so-called hierarchical structure.This means that subjects are nested within groups that may themselves be nested within higher order groups, and so on. The words *nested*, *clustered* and *multilevel* are often used as synonyms to the word *hierarchical*. Examples are clinical trials with patients nested within clinics, school-based smoking and substance abuse prevention interventions with pupils nested within classes within schools, and studies in the field of clinical psychology with clients nested within therapists.

Hierarchical data structures arise from multistage sampling where sampling is done in a number of successive steps. For instance, therapists are sampled from a population of therapists in the first step. In the second step, a sample of clients is taken from therapists selected in the first step. Only those patients whose therapists were selected in the first step have a chance of enrolling into the trial; the changes of the other patients decline to zero once the first step is executed. Thus, selection probabilities are not constant when multistage sampling is used. The sampling scheme for two-stage sampling is illustrated in Figure 1.1. The large circles represent the therapists; the small circles represent their clients. The black large circles are therapists who are not selected in the first step, and neither are any of their patients in the second step, as represented by the small black circles. The white large circles represent therapists selected in the first step. For each of these, some of his or her patients are selected in the second step and some of them are not, as represented by the small white and black circles, respectively.

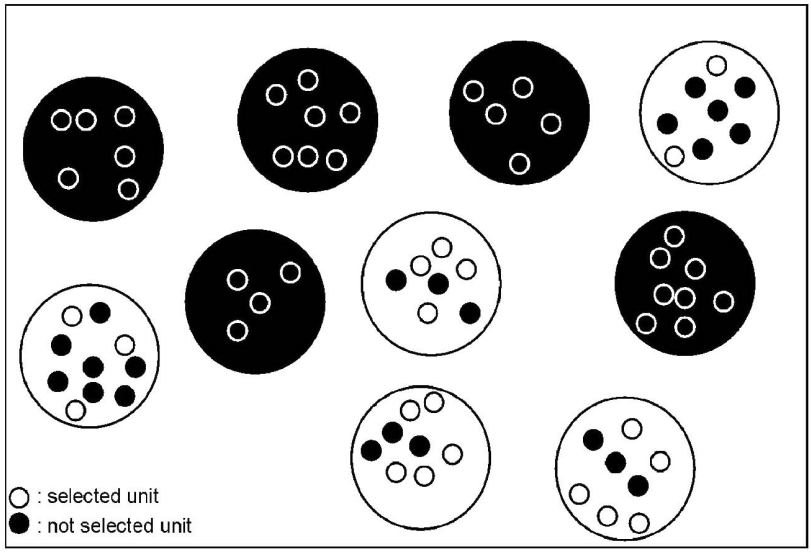


Figure 1.1: Graphical representation of a two-stage sampling scheme.

The advantage of the two-stage sampling scheme is that it is often less expensive than taking a simple random sample where each patient has equal changes of being selected. This is explained by the fact that the number of therapists is often lower if multistage sampling is used, meaning that fewer therapists have to be communicated with and fewer therapists have to be trained to deliver the new type of therapy to their clients. In addition, the patients of the same therapists are geographically organized, which may reduce travel costs.

Similar arguments hold for longitudinal studies where repeated measurements are nested within subjects. With such studies it may be more convenient to follow the same subjects over time than to take a new sample of subjects at each time point. In other words, a cohort design may be more convenient than a cross-sectional design.

The other side of the coin is that data analysis and power calculations for trials with hierarchical data are generally more complicated than those for trials without nesting because measurements of subjects within the same group are likely to be correlated. In other words, a subject's measurements on health, behavior, attitude, opinion and other relevant outcomes cannot be regarded as independent of those from other subjects within the same group. Such dependent data may arise from therapist effects, which occur when the success of treatment depends on the skill and training of the person delivering the treatment, or on the quality of the equipment that is used by this person. This is often the case in trials that rely on surgery, interviewing or physical or