39. Cluster Randomized Trials: Design and Analysis

The first section of this chapter gives an introduction to cluster randomized trials, and the reasons why such trials are often chosen above simple randomized trials. It also argues that more advanced statistical methods for data obtained from such trials are required, since these data are correlated due to the nesting of persons within clusters. Traditional statistical techniques, such as the regression model ignore this dependency, and thereby result in incorrect conclusions with respect to the effect of treatment. In the first section it is also argued that the design of cluster randomized trials is more complicated than that of simple randomized trials; not only the total sample size needs to be determined, but also the number of clusters and the number of persons per cluster.

The second section describes and compares the multilevel regression model and the mixed effects analysis of variance (ANOVA) model. These models explicitly take into account the nesting of persons within clusters, and thereby the dependency of outcomes of persons within the same cluster. It is shown that the traditional regression model leads to an inflated type I error rate for treatment testing.

Optimal sample sizes for cluster randomized trials are given in Sects. 39.3 and 39.4. These sample sizes can be shown to depend on the intra-class correlation coefficient, which measures the amount of variance in the outcome variable at the cluster level. A guess of the true value of this parameter must be available in the design stage in order to calculate the optimal sample sizes. Section 39.5 focuses on the robustness of the optimal sample size against incorrect guesses of this parameter. Section 39.6 focuses on optimal designs when the aim is to estimate the intra-class correlation with the greatest precision.

Cluster randomized trials are experiments in which complete clusters of persons, rather than the persons themselves, are randomized to treatment conditions. Such trials are frequently used in the agricultural, (bio-)medical, social, and behavioral sciences. Examples are school-based smoking prevention and cessation interventions with pupils nested within classes within schools, clinical trials with patients nested within clinics or general practices, and studies on interventions to reduce absences due to sick leave with employers nested within divisions within companies. Cluster randomized trials are very natural in the case of existing clusters, but can also be used when groups are created for the purpose of the trial. An example is a trial with therapy groups to reduce alcohol addiction. Alcoholics are assigned to small therapy groups, which in turn are assigned to treatment conditions. The difference is that, in trials with existing clusters, the persons also meet and interact outside the time slots during which the intervention is delivered, resulting in a larger degree of mutual influence.

39.1 Cluster Randomized Trials

Cluster randomized trials are often chosen above simple randomized trials that randomize persons to treatment conditions, although cluster randomized trials can easily be shown to be less efficient (Sect. 39.2). The reasons why cluster randomized trials are so often adopted must rest on other considerations than statistical efficiency, and these are often of an administrative, financial, political or ethical nature. As an example consider a study on the impact of vitamin A supplementation on childhood mortality [39.1]. In this study complete villages in Indonesia were randomly assigned to either the supplementation or control group because it was not considered politically feasible to randomize children. Another advantage of adopting a cluster randomized trial in this example is that the capsules containing the vitamin A supplements only have to be delivered to those villages in the supplementation group, which results in a reduction of travel costs. A trial that randomizes children would suffer from control-group contamination if the children in the control group were able to get access to the vitamin A from children in the supplementation group in the same village. In some cases there is no alternative to cluster randomized trials, such as in community-based interventions where the intervention will necessarily affect all members in the community. Another reason to adopt a cluster randomized trial is the wish to increase compliance. It may be reasonable to expect that compliance increases in a study where complete families, rather that just a few family members, are randomized to treatment conditions.

Due to the nesting of persons within clusters, the design and analysis of cluster randomized trials is more complicated than for simple randomized trials. The traditional assumption of independence is by definition violated when data have a nested structure. This is obvious, since there is mutual influence among persons within the same cluster. So a person's opinion, behavior, attitude or health is influenced by that of other persons in the same cluster. Furthermore, persons are influenced by cluster policy and cluster leaders. In school-based smoking prevention interventions, for instance, a pupil's smoking behavior is influenced by that of other pupils within the same class and (to a lesser degree) school, that of teachers, the school policy towards smoking and the availability of cigarettes and advertisements on smoking in the school and its neighborhood.

The traditional regression model, which assumes independent outcomes, cannot be used for the analysis of nested data. Naively using this model may lead to incorrect point estimates and standard errors of regression coefficients, and therefore to incorrect conclusions on the effect of treatment conditions and covariates on the outcome [39.2–5]. The appropriate model is the multilevel model [39.6], which is also referred to as the hierarchical (linear) model [39.7], or random coefficient model [39.6]. The multilevel model treats the persons as the unit of analysis, but explicitly takes into account nesting of persons within clusters and the correlation of outcomes of persons within the same cluster. It assumes that the clusters in the study represent a random sample from their population, and treats their effects as random in the regression analysis so that the results can be generalized to this population. Multilevel models are an extension of the variance components models and mixed effects ANOVA (analysis of variance) models [39.8] that have long been used in the biological and agricultural sciences. They are an extension in the sense that they do not only include categorical, but also continuous explanatory variables. They have been developed since the early 1980s, and in the first instance especially gained attention from the educational sciences, where data by nature have a multilevel structure due to the nesting of pupils within classes within schools. Nowadays, multilevel models are used in various fields of science, ranging from political sciences to nursery, and from studies on interviewer effects to studies with longitudinal data. It is to be expected that multilevel analysis will become part of the standard statistical techniques in the near future and that editors of scientific journals will no longer consider contributions that use old-fashioned methods to analyze multilevel data.

The design of cluster randomized trials is more complicated than that of simple randomized trials, since it does not only involve the calculation of the required number of persons, but also the calculation of the optimal allocation of units, that is, the optimal sample sizes at the cluster level and the person level. One may wonder if it is more efficient to sample many small clusters or to sample just a few large clusters. Of course, the number of available clusters is limited and the optimal number of clusters cannot therefore be larger than the available number of clusters. Likewise, the optimal cluster size cannot be larger than the actual cluster size, and such preconditions must be taken into account when calculating the optimal design. Furthermore, it is often less expensive to sample a person within an already sampled cluster than to sample a new cluster. So, the costs of sampling persons and clusters and the available budget

should also be taken into account, and it is worthwhile to calculate the required budget to achieve a pre-specified power level to detect a relevant treatment effect, or vice versa, the maximum power level given a fixed budget.

A concern in the design of cluster randomized trials is the fact that the optimal design depends on the true value of the intra-class correlation coefficient, a parameter which measures the amount of variance of the outcome variable at the cluster level. Of course, the true value is not known at the design stage, and an educated guess of this parameter must be used instead. Such a guess can be obtained from knowledge of the subject matter or from similar studies in the past. There is, however, no guarantee that such an educated guess is correct, and it is therefore worthwhile to study the robustness of optimal designs against an incorrect prior value of the intra-class correlation coefficient, and to development robust optimal design techniques.

The contents of this chapter are as follows. In the next section the multilevel regression model and the mixed effects ANOVA model are described and compared. Section 39.3 gives formulae for the optimal allocation of units for models without covariates. The extension to models with covariates is the topic of Sect. 39.4. In Sect. 39.5 we focus on the robustness properties of optimal designs. In Sect. 39.6 we present designs that are useful when interest lies in the degree of the intra-class correlation. Finally, conclusions and a discussion are given in Sect. 39.7. For the sake of simplicity, we focus on optimal designs for models with two levels of nesting, two treatment conditions, and a continuous outcome.

39.2 Multilevel Regression Model and Mixed Effects ANOVA Model

In the simplest version of a cluster randomized trial we wish to compare the effects of an intervention and a control on a single continuous outcome variable. The multilevel regression model relates outcome y_{ij} for person i in cluster j to treatment condition z_j

$$y_{ij} = \beta_0 + \beta_1 z_j + u_j + e_{ij} \,. \tag{39.1}$$

In this chapter, the treatment condition is coded $z_j = -0.5$ for the control condition and $z_j = +0.5$ for the intervention condition. So, β_0 is the mean outcome, and β_1 is the treatment effect, which is estimated by the difference in mean outcomes in both treatment groups. The null hypothesis of no treatment effect is tested by the statistic $t = \hat{\beta}_1 / \sqrt{\hat{\text{var}}(\hat{\beta}_1)}$, which has a t-distribution with $n_2 - 2$ degrees of freedom under the null hypothesis.

The multilevel model differs from the traditional regression model since it contains two random error terms. The term $u_j \sim N(0, \tau^2)$ at the cluster level is the deviation of cluster j from the mean outcome in its treatment condition, and the term $e_{ij} \sim N(0, \sigma^2)$ at the person level is the deviation of person i from the mean outcome in cluster j. These two error terms are assumed to be independent of each other and of possible covariates in the model. In general, the within-cluster variance σ^2 is much larger than the between-cluster variance τ^2.

The first two terms and the right-hand side of (39.1) are the fixed part of the model, whereas the second two terms are the random part. Since it contains both fixed and random effects, the multilevel model is a mixed effects model. Fixed effects are effects that are attributable to a finite set of levels of a factor, and they occur in the data because interest lies only in them, and not in any other levels of that factor. An example of a fixed effect is a treatment factor in a smoking prevention intervention with two levels: intervention and control. We are only interested in the comparison of these two treatment conditions, and not in any other. Random effects, on the other hand, are attributable to an infinite set of levels of a factor, of which only a random sample is included in the study at hand. An example of a random effect is the school effect in a school-based smoking prevention intervention. Although not all schools of the population under study are included in the study, we wish to generalize its findings to all schools in the population. Therefore, school is included as a random effect rather than a fixed effect.

The variances σ^2 and τ^2 are called the variance components since they sum up to the total variance of the outcome of person i within cluster j:

$$\text{var}(y_{ij}) = \sigma^2 + \tau^2 \,. \tag{39.2}$$

Furthermore, there is correlation between outcomes of two persons within the same cluster j:

$$\text{cov}(y_{ij}, y_{i'j}) = \tau^2 \,. \tag{39.3}$$

The intra-class correlation coefficient ρ measures the proportion of variation in the outcomes at the cluster

level, that is

$$\rho = \frac{\mathrm{var}(y_{ij})}{\mathrm{cov}(y_{ij}, y_{i'j})} = \frac{\tau^2}{\sigma^2 + \tau^2} \,. \tag{39.4}$$

This parameter may be interpreted as the standard Pearson correlation coefficient between any two outcomes in the same cluster. Intra-class correlation coefficients are often considerably larger in small clusters such as households, than in large clusters such as postcode levels. This can be explained by the fact that members in small clusters meet each other more often, which results in a higher level of mutual influence. As we will see in the next section, the intra-class correlation coefficient plays a crucial role in calculating the optimal sample sizes.

In a balanced design, randomization is done such that both treatment conditions have $\frac{1}{2}n_2$ clusters, and each cluster consists of n_1 persons. The variance of the treatment effect estimator is then given by

$$\mathrm{var}(\hat{\beta}_1) = 4\frac{\sigma^2 + n_1\tau^2}{n_1 n_2} = 4\frac{\sigma^2 + \tau^2}{n_1 n_2}\left[1 + (n_1 - 1)\rho\right] \,. \tag{39.5}$$

This variance is larger than that obtained with the traditional regression model due to the inclusion of the factor $[1 + (n_1 - 1)\rho]$. This factor is called the design effect, and it increases with the cluster size n_1 and the intra-class correlation coefficient ρ. Since it is always larger than 1, a cluster randomized trial is less efficient than a trial that randomizes persons to treatment conditions. Even for small values of ρ, the design effect may already be considerable. For example, if $\rho = 0.1$ and $n_1 = 10$ the design effect is equal to 1.9, and so the $\mathrm{var}(\hat{\beta}_1)$ as obtained with the multilevel model is about twice that obtained with ordinary regression analysis. So, incorrectly using the traditional regression model results in a value of $\mathrm{var}(\hat{\beta}_1)$ that is too low, and consequently in an inflated type I error rate [39.2]. This is especially the case when the cluster size n_1 and the intra-class correlation coefficient ρ are large.

When treatment condition is the only predictor variable we can write the multilevel model in terms of a mixed effects ANOVA model. For person $i = 1, \ldots, n_1$ in cluster $j = 1, \ldots, n_2$ in treatment $t = 1, 2$ we have

$$y_{ijt} = \mu + \alpha_t + u_{jt} + e_{ijt} \,. \tag{39.6}$$

Here, μ is the grand mean, α_t is the fixed effect of the t-th treatment, and u_{jt} and e_{ijt} are the random effects at the cluster and person level, which are assumed to be normally distributed with zero mean and variances of τ^2 and σ^2 respectively. Since clusters are nested within treatment conditions, we have a nested ANOVA model.

When $t = 1$ corresponds to the control group and $t = 2$ corresponds to the intervention group the correspondence between the parameters in the multilevel regression model in (39.1) and the mixed effects ANOVA model in (39.6) is given by

$$\mu = \beta_0, \quad \alpha_2 - \alpha_1 = \beta_1, \quad u_{jt} = u_j, \quad e_{ijt} = e_{ij} \,. \tag{39.7}$$

Table 39.1 gives the expected means squares (MS) for the mixed effect ANOVA model. The test statistic for the null hypothesis of no treatment effect is given by $F = MS_{\mathrm{treatment}}/MS_{\mathrm{cluster}}$, which, under the null hypothesis, has an F-distribution with 1 and $n_2 - 2$ degrees of freedom. The value of the F-test statistic for the mixed effects ANOVA model can be shown to be equal to the square of the value of the t-test statistic for the multilevel regression model [39.9]. The two variance components are estimated by

$$\hat{\sigma}^2 = \mathbf{MS}_{\mathrm{person}} \,, \tag{39.8}$$

and

$$\hat{\tau}^2 = \left(\mathbf{MS}_{\mathrm{cluster}} - \mathbf{MS}_{\mathrm{person}}\right)/n \,, \tag{39.9}$$

and the intra-class correlation coefficient is estimated by

$$\hat{\rho}^2 = \frac{\mathbf{MS}_{\mathrm{cluster}} - \mathbf{MS}_{\mathrm{person}}}{\mathbf{MS}_{\mathrm{cluster}} + (n-1)\mathbf{MS}_{\mathrm{person}}} \,. \tag{39.10}$$

For a long time the estimation of mixed models was a difficulty because of the lack of suitable estimation methods and computer programs. Different models

Table 39.1 Values for the mixed effects ANOVA model

Source	Degrees of freedom	Mean squares	Expected MS
Treatment	1	$MS_{\mathrm{treatment}}$	$\sigma^2 + n_1\tau^2 + n_1 n_2 \sum_t \alpha_t^2$
Clusters within treatment	$n_2 - 2$	MS_{cluster}	$\sigma^2 + n_1\tau^2$
Persons within clusters	$n_1 n_2 - n_2$	MS_{person}	σ^2
Total	$\boldsymbol{n_1 n_2 - 1}$		

were used but these can be shown to result in incorrect estimates of regression coefficients and their standard errors [39.2]. One such model is the traditional ordinary regression model, which assumes independent outcomes and thereby ignores nesting of persons within clusters and correlation of outcomes within the same cluster. Another approach is to calculate mean scores of variables at the cluster level and to use these in a regression model. With this approach, clusters are used as the unit of analysis, which results in loss of information. A third approach is to include clusters as fixed effects in the regression model, even if the results have to be generalized to the populations of clusters.

A method for estimation of mixed effects model became available with the development of full-information maximum-likelihood (ML), and restricted maximum-likelihood estimators (REML). The first calculates the regression coefficients and (co-)variance components such that the log likelihood log (L) is maximized, where

$$\log(L) = -\frac{1}{2}\sum_j n_{1j}\log 2\pi - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})\,. \qquad (39.11)$$

The vector $\boldsymbol{y}$ is the vector of outcomes, $\boldsymbol{\beta}$ is the vector of regression coefficients, and $\mathbf{V}$ is the covariance matrix of the outcomes, which is a function of the variance components. The design matrix $\mathbf{X}$ contains the measures on the predictor variables. REML is an adjustment of ML since it takes into account the loss of degrees of freedom resulting from estimating the fixed effects while estimating the variance components. So, the ML estimates of the variance components are downward-biased, while those for REML are not. For a large number of clusters (say $n_2 > 30$), the difference between the two estimates is negligible.

During the 1980s much attention was paid to the development of methods for the computation of ML and REML estimates, such as iterative generalized least squares (IGLS) [39.10], and restricted iterative generalized least squares (RIGLS) [39.11], which in the normal case produce ML and REML estimates, respectively. Furthermore, attention was paid to the application of existing methods, such as the Fisher scoring algorithm [39.12], and the expectation-maximization (EM) algorithm [39.13, 14]. The introduction and widespread use of personal computers have initiated the development of specialized computer programs for multilevel analysis, such as MLwin [39.15] and HLM [39.16]. Nowadays, multilevel analysis is part of general-purpose statistical software, such as SPSS and STATA.

39.3 Optimal Allocation of Units

39.3.1 Minimizing Costs to Achieve a Fixed Power Level

The primary aim of an experiment is to gain insight into the magnitude of the treatment effect, and to test if it is different from zero. Thus, we wish to test the null hypothesis $H_0 : \beta_1 = 0$ against the alternative that its value is different from zero. This hypothesis is tested by the test statistic $t = \hat{\beta}_1/\sqrt{\hat{\text{var}}(\hat{\beta}_1)}$, which has a t-distribution with $n_2 - 2$ degrees of freedom under the null hypothesis. When the number of clusters is large, the standard normal distribution can be used as an approximation, as will be done in the remainder of this chapter. For a two-sided alternative hypothesis $H_1 : \beta_1 \neq 0$, the power $1-\gamma$, type I error rate α, and the true value of β_1 are related to the variance $\text{var}(\hat{\beta}_1)$ as follows:

$$\text{var}(\hat{\beta}_1) = \left(\frac{\beta_1}{z_{1-\alpha/2} + z_{1-\gamma}}\right)^2, \qquad (39.12)$$

where $z_{1-\alpha/2}$ and $z_{1-\gamma}$ are the $100(1-\alpha/2)\%$ and $100(1-\gamma)\%$ standard normal deviates. For a one-sided alternative hypothesis, $1-\alpha/2$ may be replaced by $1-\alpha$. In general, the true value of the treatment effect β_1 is unknown at the design stage, and it is replaced by the minimal relevant deviation of β_1 from zero. If this effect is expressed in terms of units of the standard deviation $\sqrt{\sigma^2+\tau^2}$ of the outcome y_{ij}, then it is a relative treatment effect. Relative treatment effects equal to 0.2, 0.5, and 0.8 can be considered small, medium, and large, respectively, where a medium treatment effect is visible to the naked eye of a careful researcher [39.17].

As follows from (39.12), the power increases with the true value of β_1, which is obvious since large treatment effects are easier to detect than small treatment effects. Also, the power increases with the type I error rate, since null hypotheses are easier rejected if the probability of a type I error is large. Furthermore, the power is inversely related to the $\text{var}(\hat{\beta}_1)$. So, maximizing the power corresponds to minimizing the variance

of the estimated treatment effect. For studies with non-nested data this variance is related to the total sample size, and minimal sample sizes can be found in, for instance, *Cochran* [39.18]. For studies with two levels of nesting, $\text{var}(\hat{\beta}_1)$ does not only depend on the total sample size $n_1 n_2$, but also on the cluster size n_1, as follows from (39.5). Note that we use non-varying cluster sizes since that leads to the most efficient design [39.19]. In reality, cluster sizes generally vary, so that we have to take a sample of size n_1 from each cluster, meaning that not all persons in the sampled clusters are enrolled in the experiment.

The required sample sizes n_1 and n_2 can be calculated by substituting $\text{var}(\hat{\beta}_1)$ from (39.5) into (39.12). For fixed cluster size n_1 the required number of clusters is equal to

$$n_2 = 4\frac{\sigma^2+\tau^2}{n_1}\left[1+(n_1-1)\rho\right]\left(\frac{z_{1-\alpha/2}+z_{1-\gamma}}{\beta_1}\right)^2 . \tag{39.13}$$

For a fixed number of clusters n_2, the required cluster size is equal to

$$n_1 = \frac{4\sigma^2}{\left(\frac{\beta_1}{z_{1-\alpha/2}+z_{1-\gamma}}\right)^2 n_2 - 4\tau^2} \tag{39.14}$$

Figure 39.1 shows the power to detect a small relative treatment effect in a two-sided test with a type I error rate of $\alpha = 0.05$ as a function of the cluster size n_1, number of clusters n_2, and the intra-class correlation coefficient ρ. As is obvious, more clusters, larger cluster sizes and a lower intra-class correlation lead to higher power levels. For instance, 114 clusters are needed to achieve a power of 0.8 when there are 10 persons per cluster and the intra-class correlation coefficient is equal to $\rho = 0.05$. For a cluster size of $n_1 = 30$ only 66 clusters are needed. However, the total sample size for the first scenario ($n_1 n_2 = 1140$) is smaller than that for the second ($n_1 n_2 = 1980$). So, the first scenario is favorable when the aim is to minimize the total sample size, whereas the second should be selected when the aim is to minimize the number of clusters, provided that enough clusters with 30 persons are available.

As follows from the left pane in Fig. 39.1 the power increases to one when the number of clusters increases and the cluster size is fixed. On the other hand the power increases to a value not necessarily equal to one when the cluster size increases, given a fixed number of clusters. This can be explained by the fact that the cluster size n_1 appears in both the numerator and denominator of the $\text{var}(\hat{\beta}_1)$, which is inversely related to power, whereas the number of clusters n_2 appears in both. So

$$\lim_{n_1\to\infty} \text{var}(\hat{\beta}_1) = \lim_{n_1\to\infty} 4\frac{\sigma^2+n_1\tau^2}{n_1 n_2} = 4\frac{\tau^2}{n_2} , \tag{39.15}$$

and

$$\lim_{n_2\to\infty} \text{var}(\hat{\beta}_1) = \lim_{n_2\to\infty} 4\frac{\sigma^2+n_1\tau^2}{n_1 n_2} = 0 , \tag{39.16}$$

which explains why a low number of clusters cannot be compensated by a larger cluster size in order to achieve sufficient power.

When both n_1 and n_2 are free to vary, the optimal sample sizes are calculated such that the costs C for

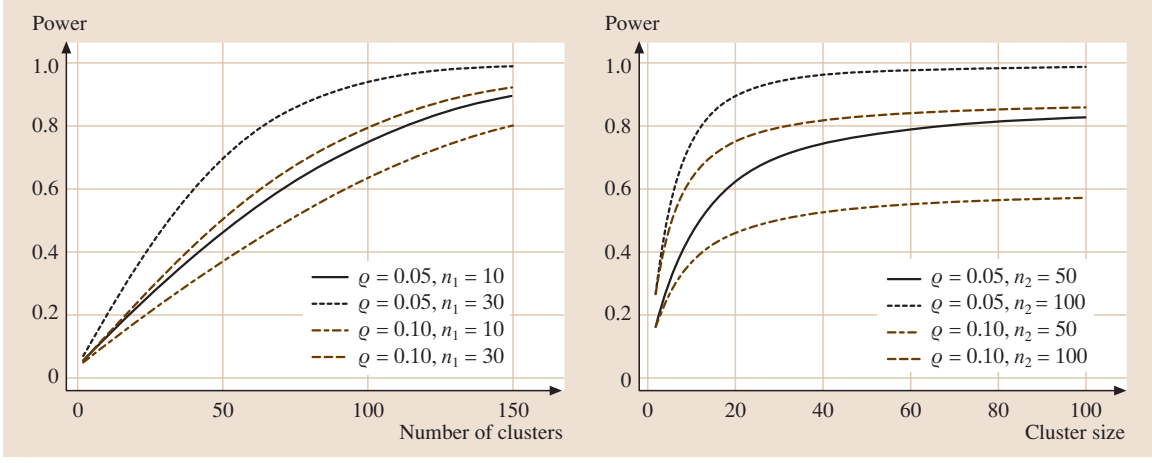

Fig. 39.1 Power as a function of cluster size, number of clusters, and intra-class correlation

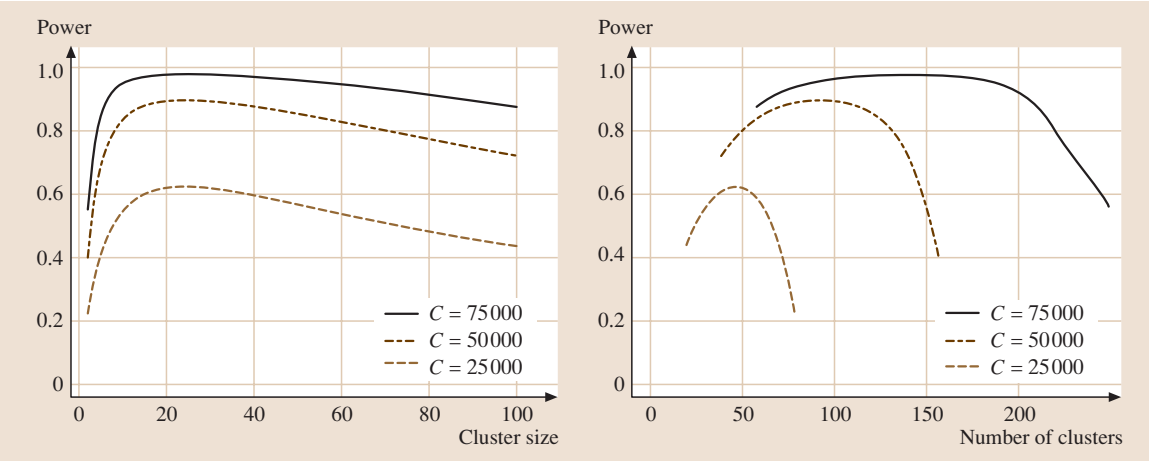

Fig. 39.2 Power as a function of cluster size and number of clusters for various budgets C and costs $c_1 = 300$ and $c_2 = 10$

recruiting and measuring persons and clusters are minimized. These costs are a function of the total number of persons $n_1 n_2$, the number of clusters n_2, the costs per person c_1, and the costs per cluster c_2:

$$C = c_1 n_1 n_2 + c_2 n_2 \,. \tag{39.17}$$

Since the design is balanced, c_1 and c_2 are the costs at the person and cluster level averaged over the two treatment conditions. In general the costs at the cluster level will be much higher than the costs at the person level. The optimal cluster size can be shown to be equal to

$$n_1 = \sqrt{\frac{c_2(1-\rho)}{c_1 \rho}} \,. \tag{39.18}$$

and n_2 follows from (39.13). Equation (39.18) was derived by expressing n_2 in terms of n_1 and C using (39.17), substituting in (39.5) for $\text{var}(\hat{\beta}_1)$, and minimizing with respect to n_1. In some cases the optimal number of persons per cluster is larger than the actual number of persons per cluster. Then, all persons have to be sampled, and additional money should be spent on sampling more clusters.

39.3.2 Maximizing Power Given a Fixed Budget

Equation (39.18) gives the optimal cluster size to achieve a pre-specified power level while minimizing costs C. On the other hand, we can also calculate the optimal cluster size for maximizing the power level when the budget is fixed to a constant C. The optimal cluster size is again equal to that given in (39.18), and the optimal number of clusters is equal to

$$n_2 = \frac{C}{\sqrt{\frac{1-\rho}{\rho} c_1 c_2} + c_2} \,. \tag{39.19}$$

The variance of the treatment effect estimator can be calculated by substituting the optimal n_1 and n_2 from (39.18) and (39.19) into (39.5), which gives

$$\text{var}(\hat{\beta}_1) = (\sigma^2 + \tau^2) \frac{\left[\sqrt{\rho c_2} + \sqrt{(1-\rho) c_1}\right]^2}{C} \,. \tag{39.20}$$

As is obvious, a larger budget C results in a smaller optimal $\text{var}(\hat{\beta}_1)$. Furthermore, a larger budget C results in sampling more clusters, but not in sampling more persons per cluster since the optimal cluster size does not depend on C. The optimal cluster size is an increasing function of the intra-class correlation coefficient ρ, so that larger cluster sizes are required when there is much variation in the outcome at the person level. Furthermore, the optimal cluster size is a function of the costs c_2 for recruiting a cluster relative to the costs c_1 for sampling a person. So, fewer clusters will be sampled in favor of sampling more persons per cluster when it is relatively expensive to sample a cluster.

Figure 39.2 shows the power to detect a small treatment effect as a function of the cluster size, number of clusters and total budget C when $c_1 = 300$ and $c_2 = 10$ and $\rho = 0.05$. The optimal cluster size is $n_1 = 24$ and this value does not depend on the budget. A budget approximately equal to $C = 75\,000$ is required to achieve a power level of 0.9 to detect a small treatment effect. The optimal number of clusters is an increasing function

of the budget C. For a large budget the power curve is rather flat around it optimum, but this is not the case for lower budgets. Of course, these power curves hold when dropout is absent, and a somewhat larger sample size is required when persons and/or clusters are expected to drop out.

39.4 The Effect of Adding Covariates

Until now we have only considered optimal designs for models without covariates. This section focuses on the effects of adding a single covariate x_{ij} that varies at the cluster and/or person level on the optimal sample sizes. The extension to multiple covariates is straightforward and not given here. The between- and within-cluster effect of the covariate on the outcome are not necessarily the same [39.20]. The covariate is therefore split up into a between-cluster component $\bar{x}_{.j}$ and a within-cluster component $(x_{ij} - \bar{x}_{.j})$, and the multilevel model is given by

$$y_{ij} = \beta_0^* + \beta_1^* z_j + \beta_2^* \bar{x}_{.j} + \beta_3^* (x_{ij} - \bar{x}_{.j}) + u_j^* + e_{ij}^* , \tag{39.21}$$

where $\beta_2^* \neq \beta_3^*$. As in the model without covariates, the random terms $u_j^* \sim N(0, \tau^{*2})$ and $e_{ij}^* \sim N(0, \sigma^{*2})$ are assumed to be independent of each other and the covariate.

When the covariate only varies at the cluster level, the term $\beta_3^*(x_{ij} - \bar{x}_{.j})$ is equal to zero and may be removed from model. An example of a cluster-level covariate is the type of school (public versus private) in a school-based smoking prevention intervention. Likewise, when the covariate only varies at the person level, the term $\beta_2^* \bar{x}_{.j}$ is equal to zero and may be removed from the model. An example of such a covariate is gender, given that the percentage of boys per school does not vary across the schools.

Note that the regression coefficients and random terms are superscripted with an asterisk in order to stress that their values may differ from those in the model without covariates (39.1). Given a grand-mean centered covariate and treatment condition coded $z_j = -0.5$ for the control group and $z_j = +0.5$ for the intervention group, the treatment effect is estimated by

$$\hat{\beta}_1^* = \frac{\sum z_j y_{ij} \sum x_j^2 - \sum z_j x_j \sum x_j y_{ij}}{n_1 n_2 \sum x_j^2 (1 - r_{xc}^2)} , \tag{39.22}$$

with variance

$$\mathrm{var}(\hat{\beta}_1^*) = 4 \frac{\sigma^{*2} + n_1 \tau^{*2}}{n_1 n_2} \frac{1}{(1 - r_{zx}^2)} . \tag{39.23}$$

When comparing formulae (39.23) with that for the variance in a model without covariates, we see that an additional factor $1/(1 - r_{zx}^2)$ is introduced. This factor is often called the variance inflation factor (VIF), and $\mathrm{var}(\hat{\beta}_1^*)$ reaches it minimum when the correlation r_{zx}^2 between the treatment condition and covariate is equal to zero. The within-cluster component $(x_{ij} - \bar{x}_{.j})$ and the treatment condition z_j are orthogonal, and therefore r_{zx}^2 is equal to the correlation between the between-cluster component $\bar{x}_{.j}$ and the treatment condition z_j. For normally and binary covariates this correlation is approximately normally distributed with variance $1/n_2$ [39.21], and thus $r_{zx}^2 \in (0, 4/n_2)$ with 95% probability. So, this correlation will be small when the number of clusters is large, and clusters are randomly assigned to treatment conditions. When the cluster randomized trial only has a small number of clusters, a correlation r_{zx}^2 equal to zero may be achieved by pre-stratification on the covariate, which means that for each value of $\bar{x}_{.j}$ half of the clusters are randomized to the control condition while the others are randomized to the intervention condition.

In the remainder of this section we will assume that the correlation between covariate and treatment condition is zero. Then, the estimated treatment effect is equal to that in a model without covariates, and the optimal sample sizes are equal to those in a model without covariates as given in (39.18) and (39.19) with τ^2 and σ^2 replaced by τ^{*2} and σ^{*2}, respectively [39.22]. The relations between the variance components in a model with and without covariates can be established using the method described in [39.23]. During the analysis stage the total variation in the outcome y_{ij} is given by the observed data and the estimated variance components change if covariates are added to or excluded from the model. The chance in the estimated variance components can be derived by assuming that the variance of the observed outcomes and covariance of two outcomes within the same cluster are given by the data and are therefore equal for model (39.21) and (39.1):

$$\begin{aligned} \mathrm{var}(y_{ij}) &= \mathrm{var}(\beta_1 z_j + u_j + e_{ij}) \\ &= \mathrm{var}\left[\beta_1^* z_j + \beta_2^* (x_{ij} - \bar{x}_{.j}) + \beta_3^* \bar{x}_{.j} + u_j^* + e_{ij}^*\right] \end{aligned} \tag{39.24}$$

Table 39.2 Changes in the variance components due to the inclusion of a covariate

Changes due to the inclusion of $\bar{x}_{.j}$	Changes due to the inclusion of $(x_{ij} - \bar{x}_{.j})$
$\tau^2 - \tau^{*2} = \hat{\beta}_2^{*2}\text{var}(\bar{x}_{.j}) > 0$	$\tau^2 - \tau^{*2} = \hat{\beta}_3^{*2}\text{cov}(x_{ij} - \bar{x}_{.j}, x_{i,j} - \bar{x}_{.j}) < 0$ ≈ 0 for large n_1
$\sigma^2 - \sigma^{*2} = 0$	$\sigma^2 - \sigma^{*2} = \hat{\beta}_3^{*2}\left[\text{var}(x_{ij} - \bar{x}_{.j}) - \text{cov}(x_{ij} - \bar{x}_{.j}, x_{i,j} - \bar{x}_{.j})\right] > 0$ $\approx \hat{\beta}_3^{*2}\text{var}(x_{ij} - \bar{x}_{.j}) > 0$ for large n_1

Note: It is assumed that $r_{zx}^2 = 0$

and

$$\begin{aligned}\text{cov}(y_{ij}, y_{i'j}) &= \text{cov}(\beta_1 z_j + u_j, \beta_1 z_j + u_j) \\ &= \text{cov}\Big[\beta_1^* z_j + \beta_2^*(x_{ij} - \bar{x}_{.j}) + \beta_3^* \bar{x}_{.j} \\ &\quad + u_j^*, \beta_1^* z_j + \beta_2^*(x_{i'j} - \bar{x}_{.j}) \\ &\quad + \beta_3^* \bar{x}_{.j} + u_j^*\Big] . \end{aligned} \tag{39.25}$$

Table 39.2 shows the changes in the estimated variance components due to the inclusion of one covariate. The variance component at the person level remains unchanged when a cluster-level covariate is added to the model, and decreases when a person-level covariate is added to the model. The variance component at the cluster level decreases when a cluster-level component is added, but *increases* when a person-level covariate is added. However, for large cluster sizes this increase is negligible, and it may be nullified by the decreasing effect of adding a cluster-level covariate. So, adding covariates will in general lead to a decrease in the variance components, and therefore in a more efficient design, given a zero correlation between the covariate and treatment condition.

Of course, costs are associated with measuring covariates and one may wonder when adding a covariate may be a more cost-efficient strategy to increase the power to detect a treatment effect than sampling more clusters. Both strategies have recently been compared, and it was concluded that adding covariates is more efficient when the costs to measure these covariates are small and the correlation between the covariate and the outcome is large [39.24]. Adding a covariate at the cluster level is recommended when clusters are large (say $n_1 = 100$) and the costs to recruit and measure a cluster are small in relation to the costs to recruit and measure a person. Vice versa, adding a covariate that only varies at the person level is recommended when clusters are small (say $n_1 = 4$) and the relative costs to recruit and measure a cluster are large.

39.5 Robustness Issues

In the Sect. 39.3 it was shown that the optimal sample sizes depend on the value of the intra-class correlation coefficient. The value of this parameter is generally unknown at the design stage and an educated guess must be obtained from subject-matter knowledge or similar studies in the past. Table 1 in [39.25] gives an overview of recent papers that report values of the intra-class correlation coefficient. There is, however, no guarantee that the values of similar studies in the past are the true values for the current study at hand, since the study may be conducted in a different year of country, or may target a different population (e.g. elementary-school children instead of high-school children).

As an example consider a cluster randomized trials that aims at detecting a small relative treatment effect at power level 0.9 in a two-sided test with $\alpha = 0.05$. The cluster size is equal to $n_1 = 30$, and the true but unknown intra-class correlation is $\rho = 0.05$. The required number of clusters at prior value $\rho = 0.05$ is equal to $n_2 = 86$, and this results in a power equal to 0.9, since the prior ρ is equal to the true ρ. However, if the prior estimate is equal to $\rho = 0.10$, then the required number of clusters can be calculated to be equal to $n_2 = 138$. Thus, the number of clusters is overestimated by 60%, and the power level at the true ρ is equal to 0.98. For a prior estimate as small as $\rho = 0.025$, the required number of clusters is equal to $n_2 = 62$, which results in a power level of 0.78 at the true ρ. Hence, cluster randomized trials are not very robust against an incorrect prior estimate of the intra-class correlation coefficient.

Since it is increasingly difficult to obtain adequate financial recourses, and since cluster randomized trials require the willingness of clusters and persons to participate, it is extremely important to design trials such

Table 39.3 Assumptions about the intra-class correlation coefficient, with associated power with 86 groups and required number of groups for a power level of 0.9

Intra-class correlation coefficient Median (95% interval)	Power with 86 groups Median (95% interval)	Number of groups for power = 0.9 Median (95% interval)
0.05–0.051	0.90–0.898	86–88
0.008–0.099	0.734–0.995	44–136

that they are not under- or overpowered. Two procedures to calculate robust optimal designs are Bayesian optimal designs, where a prior distribution on the intra-class correlation is used, and designs with sample-size re-estimation based on data obtained from a pilot.

39.5.1 Bayesian Optimal Designs

Bayesian methods allow us to implicitly take uncertainty about model parameters into account by using a prior distribution on the parameters. Consider the example given above and suppose that we assume the intra-class correlation to be around 0.05, but that there is some change that it is up to 0.10. This uncertainty may be reflected by a normal distribution with mean 0.05 and standard deviation 0.025, but truncated at zero so that we exclude negative values. We can now sample from this prior distribution and calculate the required number of clusters to achieve a power level of 0.9. In addition, we can also calculate the power level that is achieved with 86 clusters.

The results in Fig. 39.3 and Table 39.3 were obtained after 100 000 iterations, which took less that one minute on a desktop computer with a 2.8-GHz CPU and 1 Gb of RAM. The median intra-class correlation coefficient is equal to 0.051, at which there are hardly any values larger than 0.1. The median power achieved with 86 clusters is equal to 0.0898, so there is a change of about 50% that the power is less than the required level of 0.9. In some cases, it can even be as small as 0.7. The median required number of clusters is equal to 88, whereas the boundaries of the 95% interval are 44 and 136. So, on the basis of the results in Fig. 39.3 and Table 39.3 we might decide to use a number of clusters larger than 86 to be reasonably confident that the study has sufficient power.

39.5.2 Designs with Sample-Size Re-Estimation

Designs with sample-size re-estimation have been proposed by *Stein* [39.26] in the context of comparing two treatment conditions with respect to a continuous outcome. His procedure includes two stages. In the first stage (the internal pilot) the variance of the outcome is estimated using the observations collected so far, and the total sample size is re-estimated based on the variance estimate. In the second stage the remainder of the observations is collected such that re-estimated total sample size is achieved. Only the observations of the first stage are used to estimate the variance

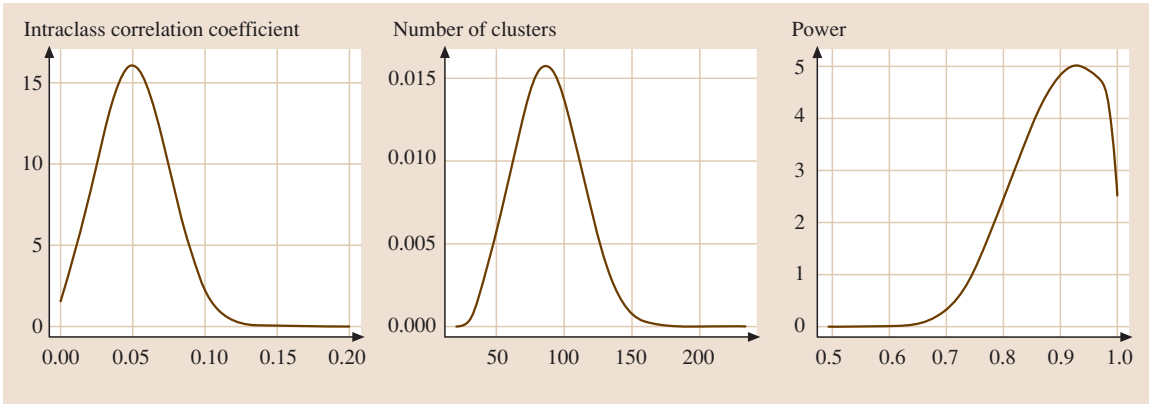

Fig. 39.3 Densities of the prior distribution of the intra-class correlation coefficient, the required number of clusters to achieve a power level 0.9, and the power at 86 clusters

Table 39.4 Empirical type I error rate α and power $1-\beta$ for the standard design and re-estimation design for three values of the prior ρ. The true $\rho = 0.05$

Prior ρ	Standard design		Re-estimation design					
			$\pi = 0.25$		$\pi = 0.5$		$\pi = 0.75$	
	α	$1-\beta$	α	$1-\beta$	α	$1-\beta$	α	$1-\beta$
0.025	0.0538	0.7812	0.0526	0.8690	0.0600	0.9072	0.0586	0.9012
0.05	0.0480	0.9004	0.0576	0.8886	0.0530	0.9094	0.0556	0.8986
0.10	0.0502	0.9832	0.0534	0.8964	0.0588	0.9114	0.0532	0.9474

of the outcomes, while all observations are used in the calculation of the group means. This procedure was modified by *Wittes* and *Britain* [39.27] such that all data are used in the final analysis. In contrast to the Stein procedure, the Wittes and Britain procedure does not preserve the type I error rate since the total sample size depends on the variance estimate in the pilot. Internal pilots have been shown to work well for cluster randomized trials by *Lake* et al. [39.28].

We consider the same example where we wish to detect a small relative treatment effect at power level 0.9. The true $\rho = 0.05$, and we have three prior values $\rho = 0.025$, 0.05, and 0.10. Table 39.4 shows the empirical type I error rates and power levels in a simulation study with 5000 runs. The power levels for the design without sample-size re-estimation (i. e. the standard design) are too small when the prior ρ is underestimated and too large when the prior ρ is overestimated. The values of the type I error rate are close to their nominal value of $\alpha = 0.05$.

For designs with sample-size re-estimation the required number of clusters is calculated on the basis of the prior ρ . Then, a predefined proportion π of this number of clusters is used in the internal pilot. The required number of clusters in the second stage is calculated on the basis of the parameter estimates obtained from data collected in the internal pilot. When the size of the internal pilot is already sufficiently large, a second stage is not needed. Table 39.4 shows that the power levels for studies with incorrect prior values ρ are much closer to the value 0.9 than they are in the standard design. For $\pi = 0.25$ and prior $\rho = 0.025$, the power is somewhat lower than 0.9, which is explained by the fact that the size of the internal pilot is somewhat too small to result in a good estimate of the true ρ. For $\pi = 0.75$ and prior $\rho = 0.10$, the power is larger than 0.9, which is explained by the fact that the size of the internal pilot is already too large. The empirical type I error rates are somewhat, but not dramatically, larger than the nominal value $\alpha = 0.05$.

39.6 Optimal Designs for the Intra-Class Correlation Coefficient

So far we have focussed on optimal designs that maximize the power to detect a treatment effect or, equivalently, minimize the variance of the treatment effect estimator. Another option is to design a study such that it minimizes the variance of the intra-class correlation coefficient estimator, which is equal to

$$\mathrm{var}(\hat{\rho}) = \frac{2(1-\rho)^2(1+(n_1-1)\rho)^2}{(n_1-1)(n_1n_2-n_1)} \,. \qquad (39.26)$$

Such optimal designs are especially useful for pilot studies that aim at an estimate of the intra-class correlation coefficient. Again, we can minimize this variance under the precondition that the costs for recruiting persons and clusters do not exceed the budget, as specified by (39.17). Closed-form equations for the optimal n_1 and n_2 do not exist. Instead, the optimal design may be found by expressing n_2 in terms of n_1, c_1, c_2 and C using (39.17): $n_2 = C/(c_1n_1 + c_2)$. This relation may then be substituted into (39.26), from which the optimal n_1 may be calculated.

For most trials the main focus lies on the treatment effect, but researchers may also be interested in the degree of variability of the outcome that is between clusters. If the amount of between-cluster variability turns out to be high, then one may wish to identify those schools for which the intervention performs worst and try to characterize these schools in terms of their school-level variables. The intervention can then be adjusted for these types of schools. For instance, a smoking prevention intervention that works well for high schools may

have to be adjusted for schools for lower vocational education.

When a researcher has multiple objectives in mind, he or she may design a multiple-objective optimal design. Suppose that we wish to design a trial that aims at estimating both the treatment effect and intra-class correlation with largest precision, that is, it aims at minimizing $\text{var}(\hat{\beta}_1)$ and $\text{var}(\hat{\rho}_1)$. These two variances are the two objectives and the first is the most important since the trial is, in the first instance, designed to gain insight into the value of the treatment effect, whereas the intra-class correlation coefficient is of secondary importance. The two-objective optimal design is the one that does best under the criterion $\text{var}(\hat{\rho}_1)$ subject to the constraint that the value $\text{var}(\hat{\beta}_1)$ is smaller than a user-specified constant c:

$$\min \text{var}(\hat{\rho}) \text{ subject to } \text{var}(\hat{\beta}_1) \leq c \,. \tag{39.27}$$

The design that satisfies this criterion is often called a constrained optimal design. For convenience, this criterion is often rewritten as

$$\min \text{var}(\hat{\rho}) \text{ subject to } \text{eff}(\hat{\beta}_1) \geq e \,, \tag{39.28}$$

where $\text{eff}(\hat{\beta}_1)$ is the efficiency in estimating the treatment effect. So, the least important optimality criterion is minimized subject to the constraint that the efficiency in estimating the treatment effect is larger than a user-selected constraint. The efficiency is calculated as the $\text{var}(\hat{\beta}_1)$ obtained with the optimal sample sizes as given by (39.18) and (39.19) divided by the $\text{var}(\hat{\beta}_1)$ obtained with any other sample sizes n_1 and n_2. The efficiency varies between zero and one. Its interpretation is that, if N observations are used in the optimal design, then $N/\text{eff}(\hat{\beta}_1)$ observations are used in the suboptimal design to obtain the same amount of information.

Constrained optimal designs are often difficult to derive, and one may wish to construct a compound optimal design to minimize

$$\lambda \, \text{var}(\hat{\rho}) + (1-\lambda) \, \text{var}(\hat{\beta}_1) \,. \tag{39.29}$$

Compound optimal designs are generally easier to solve, either numerically or analytically. Under convexity and differentiability constrained and compound optimal designs are equivalent [39.29]. So, in order to derive

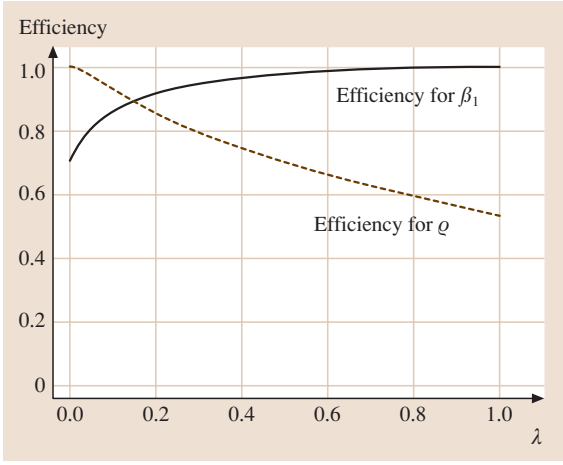

Fig. 39.4 Efficiency plot

the constrained optimal design one may first derive the compound optimal design as a function of the weight λ in (39.29). That is, for each value of λ the sample sizes n_1 and n_2 that minimize (39.29) are derived. Subsequently, an efficiency plot is drawn in which the efficiencies $\text{eff}(\hat{\beta}_1)$ and $\text{eff}(\hat{\rho})$ are plotted as a function of λ. The constrained optimal design is the design for which $\text{eff}(\hat{\beta}_1) \geq e$ and $\text{eff}(\hat{\rho})$ is maximized. In most practical situations the constant e is chosen to be 0.8 or 0.9.

Figure 39.4 shows an efficiency plot for a trial with $C = 50\,000$, $c_1 = 30$, $c_2 = 10$ and $\rho = 0.025$. The optimal design for estimating ρ with the greatest precision is $n_1 = 45.4$ and $n_2 = 103.4$ and is achieved when $\lambda = 1$. The efficiency for ρ is a decreasing function of λ. The optimal for estimating β_1 with largest precision is $n_1 = 10.8$ and $n_2 = 362.8$, and is achieved when $\lambda = 0$. The efficiency for β_1 is a increasing function of λ. Note that the two lines do not necessarily meet at the point. When we wish to estimate ρ with the greatest precision, given the condition that $\text{eff}(\hat{\beta}_1) \geq 0.9$, then we draw a horizontal line at $e = 0.9$ to intersect the graph of $\text{eff}(\hat{\beta}_1)$. Then a vertical line is drawn from this point of intersection to meet the λ-axis. This results in $\lambda = 0.17$, which corresponds to $n_1 = 23.42$ and $n_2 = 189.3$, and $\text{eff}(\hat{\rho}) = 0.876$. Of course, these sample sizes have to rounded off to integer values. Large efficiencies are possible for both criteria, which are therefore called compatible.

39.7 Conclusions and Discussion

Cluster randomized trials randomize complete groups of persons, rather than the persons themselves, to treatment conditions. They are often used in situations where the intervention is delivered to groups of persons, such as in school-based smoking prevention interventions with class teaching on smoking and health. Since the outcomes of persons in a group cannot be considered to be independent, a larger sample size is required to achieve a pre-specified power level than in a simple randomized trial, especially when the intra-class correlation coefficient and/or the cluster size are large.

Multisite trials are an alternative to cluster randomized trials. Multisite trials randomize persons within clusters to treatment conditions, such that all treatments are available within each cluster. So, for multisite trials cluster and treatment condition are crossed, whereas for cluster randomized trails clusters are nested within treatment conditions. Multisite trials have two advantages above cluster randomized trials: they are more powerful, and they allow for the estimation of the cluster by treatment interaction [39.30]. A main drawback of multisite trials is that they do not protect from control-group contamination, which occurs when information on the intervention leaks from the individuals in the intervention group to those in the control group [39.31]. In some cases blinding may be an option to prevent control-group contamination, such as in double-blind placebo-controlled multicentre clinical trial with patients nested within clinics. This is an option when the experimental treatment is a new pill, which only differs from the pills in the control group by the amount of active substance. When patients are randomly assigned to treatment conditions and neither the patient nor the researchers know who belongs to which treatment, a multisite study may be an alternative to a cluster randomized trial. Blinding is of course no option when the intervention consists of interpersonal relationships, such as in peer-pressure groups. Control-group contamination may also be due to the person delivering the intervention, such as in guideline trials with patients nested within family practices. If both a control and intervention group are available in each practice, it will be extremely difficult for the physician not to let patients in the control group benefit from the intervention. Of course, the choice for a cluster randomized trial does not guarantee the absence of control-group contamination. An example is a trial in which general practices are randomized to treatment conditions and the intervention consists of leaflets to promote healthy lifestyles. Control-group contamination can occur when staff members work between practices and distribute leaflets in the control practices. Another example is a school-based smoking prevention intervention where children from different families attend different schools, and thereby encounter different treatment conditions.

This chapter has given an introduction to the design and analysis of cluster randomized trails. It focused on models with two levels of nesting, two treatment conditions, and continuous outcomes. The extension to three or more levels of nesting is straightforward and can be found elsewhere [39.30, 32]. The optimal sample sizes were shown to depend on the value of the intra-class correlation coefficient and it was shown that an incorrect prior may lead to an under- or overpowered study. This may be overcome by using a robust optimal design, such as a Bayesian optimal design or a design using sample-size re-estimation. Such designs are also very useful for the planning of cluster randomized trials with binary outcomes, since then the optimal sample size can be shown not only to depend on the intra-class correlation coefficient, but also on the probabilities of a positive response in each treatment condition [39.33, 34].

References

39.1 A. Sommer, I. Tarwotjo, E. Djunaedi, K. P. West, A. A. Loeden, R. Tilden, L. Mele: Impact of vitamin A supplementation on childhood mortality. A randomised controlled community trial, Lancet **1986**, 1169–1173 (1986)

39.2 M. Moerbeek, G. J. P. van Breukelen, M. P. F. Berger: A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies, J. Clin. Epidemiol. **56**, 341–350 (2003)

39.3 H. Goldstein: *Multilevel Statistical Models*, 3rd edn. (Edward Arnold, London 2003) p. 3

39.4 J. Hox: *Multilevel Analysis. Techniques and Applications* (Erlbaum, New Jersey 2002)

39.5 T. A. B. Snijders, R. J. Bosker: *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (Sage, London 1999)

39.6 N. T. Longford: *Random Coefficient Models* (Clarendon, Oxford 1993)

39.7 S. W. Raudenbush, A. S. Bryk: *Hierarchical Linear Models. Applications and Data Analysis Methods* (Sage, Thousand Oaks 2002)

39.8 S. R. Searle, G. Casella, C. E. McCulloch: *Variance Components* (Wiley, New York 1992)

39.9 S. W. Raudenbush: Hierarchical linear models and experimental design. In: *Applied Analysis of Variance in Behavioral Science*, ed. by L. K. Edwards (Wiley, New York 1993) pp. 459–496

39.10 H. Goldstein: Multilevel mixed linear model analysis using iterative generalized least squares, Biometrika **73**(1), 43–56 (1986)

39.11 H. Goldstein: Restricted unbiased iterative generalized least squares estimation, Biometrika **76**, 622–623 (1989)

39.12 N. T. Longford: A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects, Biometrika **74**, 817–827 (1987)

39.13 A. P. Dempster, D. B. Rubin, R. K. Tsutakawa: Estimation in covariance components models, J. Am. Stat. Assoc. **76**(374), 341–353 (1981)

39.14 W. M. Mason, G. Y. Wong, B. Entwisle: Contextual analysis through the multilevel linear model. In: *Sociological Methodology 1983–1984*, ed. by S. Leinhardt (Jossey-Bass, San Francisco 1983) pp. 72–103

39.15 J. Rasbash, F. Steele, W. Browne: *A User's Guide to MLwiN Version 2.0* (Institute of Education, London 2004)

39.16 S. W. Raudenbush: *HLM 6. Hierarchical Linear and Nonlinear Modeling* (Scientific Software International, Chicago 2004)

39.17 J. Cohen: A power primer, Psychol. Bull. **112**(1), 155–159 (1992)

39.18 W. G. Cochran: *Planning and Analysis of Observational Studies* (Wiley, New York 1983)

39.19 A. K. Manatunga, M. G. Hudges, S. Chen: Sample size estimation in cluster randomized studies with varying cluster size, Biom. J **43**(1), 75–86 (2001)

39.20 J. M. Neuhaus, J. D. Kalbfleisch: Between- and within-cluster covariate effects in the analysis of clustered data, Biometrics **54**, 638–645 (1998)

39.21 M. Kendall, A. Stuart: *The Advanced Theory of Statistics. Vol. 2: Inference and Relationship* (Griffin, London 1979)

39.22 M. Moerbeek, G. J. P. van Breukelen, M. P. F. Berger: Optimal experimental designs for multilevel models with covariates, Commun. Statist. Theory Methods **30**(12), 2683–2697 (2001)

39.23 T. A. B. Snijders, R. J. Bosker: Modeled variance in two-level models, Sociol. Methods Res. **22**(3), 342–363 (1994)

39.24 M. Moerbeek: Power and money in cluster randomized trials: when is it worth measuring a covariate?, Stat. Med. in press

39.25 D. M. Murray, S. P. Varnell, J. L. Blitstein: Design and analysis of group-randomized trials: A review of recent methodological developments, Am. J. Public Health **94**(3), 423–432 (2004)

39.26 A. C. Stein: A two-sample test for a linear hypothesis whose power is independent of the variance, Ann. Math. Stat. **29**, 1271–1275 (1945)

39.27 J. Wittes, E. Brittain: The role of internal pilot studies in increasing the efficiency of clinical trials, Stat. Med. **9**(1), 65–72 (1990)

39.28 S. Lake et al.: Sample size re-estimation in cluster randomization trials, Stat. Med. **21**(10), 1337–1350 (2002)

39.29 D. Cook, W. K. Wong: On the equivalence of constrained and compound optimal designs, J. Am. Stat. Assoc. **89**(426), 687–692 (1994)

39.30 M. Moerbeek, G. J. P. van Breukelen, M. P. F. Berger: Design issues for experiments in multilevel populations, J. Educ. Behav. Stat. **25**(3), 271–284 (2000)

39.31 M. Moerbeek: Randomization of clusters versus randomization of persons within clusters: Which is preferable?, Am. Stat. **59**(1), 72–78 (2005)

39.32 T. C. Headrick, B. D. Zumbo: On optimizing multilevel designs: Power under budget constraints, Austr. New Zealand J. Stat. **47**(2), 219–229 (2005)

39.33 M. Moerbeek, G. J. P. van Breukelen, M. P. F. Berger: Optimal experimental design for multilevel logistic models, Statistician **50**(1), 17–30 (2001)

39.34 M. Moerbeek, C. J. M. Maas: Optimal experimental designs for multilevel logistic models with two binary predictors, Commun. Stat. Theory Methods **34**(5), 1151–1167 (2005)