CLUSTERED TRIALS

Clustered trials are trials with a multiple-level design. For example, we assign hospitals at random to either of two conditions, and then all patients within each hospital are assigned to the same condition. Or, we assign schools at random to either of two conditions, and then all students within each school are assigned to the same condition.

How the use of Cluster Randomized Trials affects power

The logic of power analysis is fundamentally the same for studies that use simple random samples and for studies that use cluster random samples. For both, power is a function of the effect size, of the criterion for significance (alpha), and of the sampling distribution of the treatment effect. For purposes of power, the key difference between the two designs is in the last of these, the sampling distribution of the treatment effect. In the simple randomized trial there is only one source of error, the within-groups variance. By contrast, in the cluster randomized trial, there are two sources of error, within-clusters and between-clusters.

Consider a study where the outcome is the mean level of pain reported by patients subsequent to a surgical procedure. Standard care (Control) calls for patients to take pain medication according to one regimen, and the experimental condition (Treatment) calls for the patients to take this medication according to a new regimen.

Four hospitals will be assigned to the Control condition and four will be assigned to the Treatment condition. Power depends in part on the precision with which we can assess the infection rate in each condition (and the difference, which is the treatment effect). There will be two sources of error in our estimate.

One source of error is that the mean level of pain that we observe in a specific hospital is not the true mean in that hospital. If the true mean in Hospital A is 4 (on a 10-point scale) then we may observe a mean of 3.5 or 4.5 because of sampling error. The primary mechanism for reducing this error is to increase the sample size within hospitals.

The second source of error is that the true mean in these four hospitals is not the same as the true mean of all hospitals. The true mean varies from one hospital to the next, and so the mean in any sample of four hospitals (even if we could eliminate the first source of error) would not be the same as the mean across all possible hospitals. The primary mechanism for reducing this error is to increase the number of hospitals in the sample.

Note. The second source of error (between-studies variance) is a problem for cluster-randomized trials but not for multi-center trials using simple random sampling. This is because in a simple multi-center trial every hospital includes patients assigned to both the Treated and the Control conditions. If a hospital happens to have a low risk or a high risk, this affects both conditions equally and therefore has no impact on the effect size (which is based on the difference between them). Therefore, under the usual assumptions (such as homogeneity of treatment effects across clusters) the between-clusters variance has little or no impact on the error (or the power). By contrast, in a cluster randomized sample, each hospital is assigned entirely to one condition or the other. If a hospital happened to have a low mean or a high mean, this would affect one condition only, and therefore would have a direct impact on the effect size. More

generally, this additional source of variance results in a larger error term and decreases the power of the test.

Implications for study planning

In planning a simple randomized trial we needed to address only one source of error, the dispersion of subjects from the population mean, and we do this by increasing the sample size. The question of allocating resources is straightforward, in the sense that there is only one option available (to increase the sample size).

In planning a cluster randomized trial, by contrast, we need to address the two sources of error. To reduce the error within hospitals we need to increase the *N* within hospitals. To reduce the error between hospitals we need to increase the number of clusters. Since there are now two mechanisms for reducing error (and increasing power) we need to consider how to best allocate resources between these options.

Decisions about allocation will depend largely on the extent to which the outcome varies between hospitals (clusters). If the risk of infection is consistent from one hospital to the next, then we might need only a few hospitals in our sample to obtain an accurate estimate of the infection rate. By contrast, if the risk varies substantially from one hospital to the next, then we might need to sample many hospitals to obtain the same level of accuracy.

For purposes of power it is useful to talk about the between-cluster variance as a proportion of the total variance. This proportion is called the intraclass correlation (*ICC*), denoted by ρ and defined as

$$\rho = \frac{\sigma_B^2}{\sigma_W^2 + \sigma_B^2}$$

where σ^2_B is the variance between clusters, σ^2_W is the variance within a cluster and ρ is the *ICC*. Power depends on

$$Power = 1 - F\left(c_\alpha, df, \lambda\right)$$

where F(*x,v,l*) is the cumulative distribution function of the test statistic at value *x*, with *v* degrees of freedom and non-centrality parameter *l*. In this expression, power is an increasing function of

$$\lambda = \left(\sqrt{\frac{mn}{2}}\right)\left(\delta\sqrt{\frac{1}{1+(n-1)\rho}}\right),$$

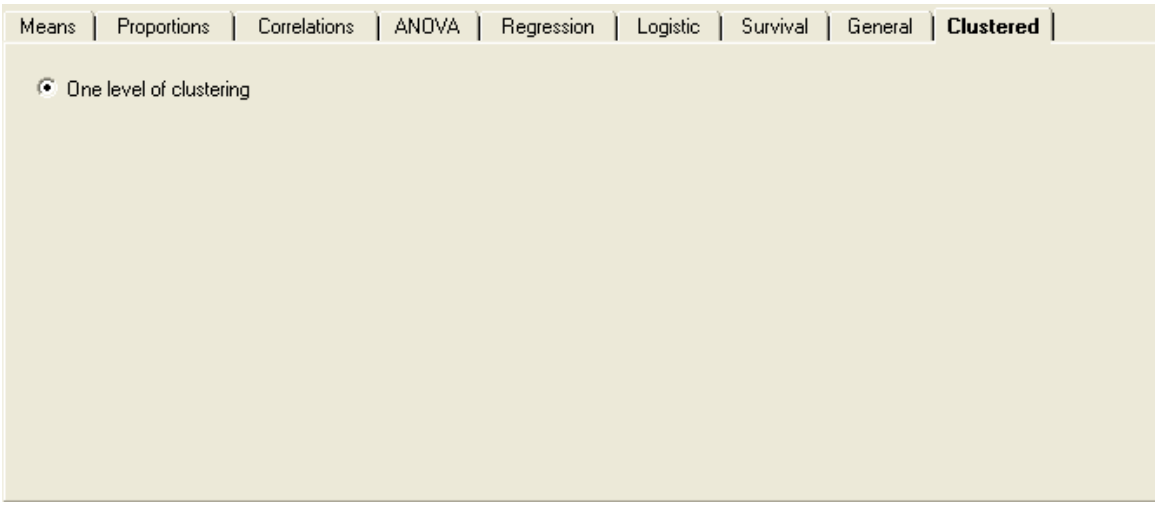
where *m* is the number of clusters in each condition, *n* is the number of individuals in each cluster, δ is the effect size, and ρ is the *ICC*.

As outlined above, the researcher planning a cluster randomized trial must choose to allocate resources in various ways, for example by increasing the sample size within clusters, or by increasing the number of clusters. Since these mechanisms compete with

each other for resources (a finite amount of time and money can be used either to increase the *n* within each cluster or the number of clusters), we must compare the impact of these different approaches and identify the most appropriate design for the study in question.

SELECTING THE PROCEDURE

- Choose New analysis from the File menu.
- Click the Clustered tab.
- Click OK to proceed to the module.

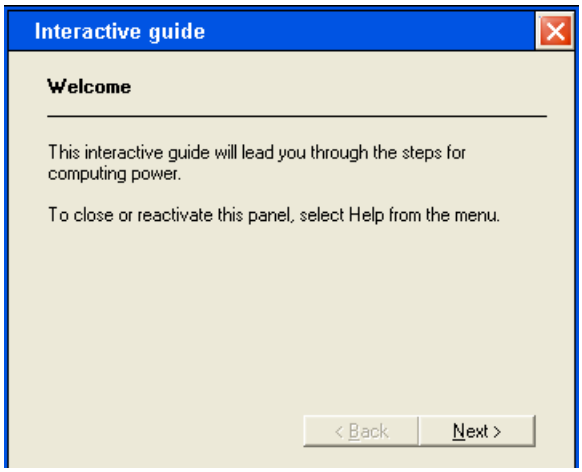


INTERACTIVE SCREEN

Interactive guide

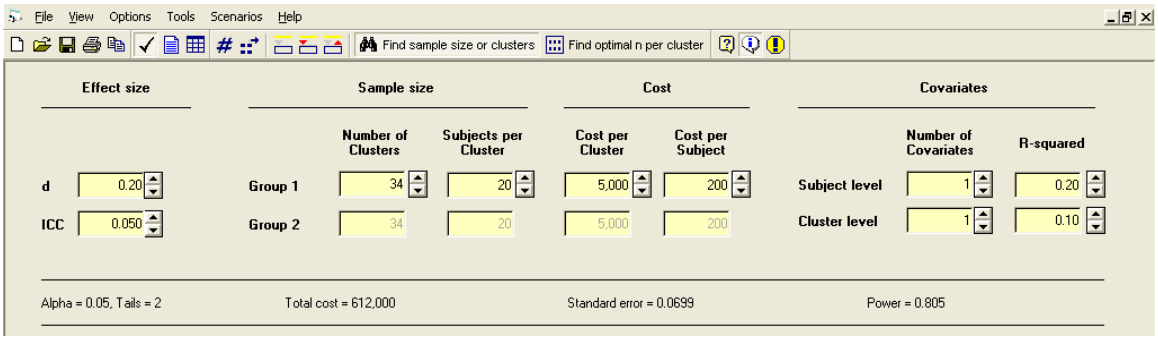
Click Help > Interactive guide

The program displays a guide that will walk you through the full process of power analysis, as explained in the following text.



SETTING VALUES ON THE INTERACTIVE SCREEN

The interactive screen is shown here.



Name the groups

Initially, the program refers to the conditions as “Group 1” and “Group 2”. Click on either name (or click on Tools > Assign labels) and enter labels such as “Treated” and “Control”.

Name the clusters and subjects

Initially, the program refers to the two levels of sampling as “Cluster” and “Subject”. Click on either name (or click on Tools > Assign labels) and enter labels such as “Hospital” and “Patient” or “School” and “Student” (using the singular form rather than the plural).

Effect size.

d is the standardized mean difference, defined as the raw difference between conditions divided by the standard deviation. The standard deviation in this case is computed within conditions and across clusters.

ICC is the intraclass correlation

For purposes of power it is useful to talk about the between-cluster variance as a proportion of the total variance. This proportion is called the intraclass correlation (*ICC*), denoted by ρ and defined as

$$\rho = \frac{\sigma_B^2}{\sigma_W^2 + \sigma_B^2}.$$

The *ICC* reflects how the cluster means differ from each other within a condition. If they differ only a little (if the variance of the cluster means is only slightly more than the variance within clusters) then the *ICC* is relatively close to zero. If they differ by a lot (if the variance of the cluster means is substantially more than the variance within clusters) then the *ICC* is relatively far from zero.

The possible range of the *ICC* is 0.00 to 0.99, but in practice the *ICC* in any given field of research tends to fall within a more narrow range, and can be estimated based on prior studies. In some fields, the *ICC* might be expected to fall in the range of 0.01 to 0.05. In others, it would be expected to fall in the range of 0.10 to 0.30.

Sample size

There are two elements to the sample size – the number of clusters, and the number of subjects per cluster.

- Enter the number of clusters (for example, the number of hospitals or schools).
- Enter the number of subjects per cluster (for example, the number of patients per hospital, or the number of students per school).

Cost

The information in this section is optional. It is not needed to compute power. However, it is needed to find the optimal (most cost-effective) number of subjects per cluster (see below).

- Enter the cost of enrolling a new cluster (for example, the cost of enrolling a new hospital or school).
- Enter the cost of enrolling, treating, and following a new subject (for example, a new patient or student).

Covariates

- Subject-level. The study may have covariates at the subject level, For example, the patient's age or the student's pre-score may serve as a covariate to explain some of the outcome (and thus reduce the error term). If there are subject-level covariates, enter the number of covariates and the expected value of R^2 (the proportion of variance explained by the covariates).

- Cluster-level. The study may have covariates at the cluster level, For example, the hospital's mean success rate, or the student's mean SAT score may serve as a covariate to explain some of the outcome (and thus reduce the error term). If there are cluster-level covariates, enter the number of covariates and the expected value of R^2 (the proportion of variance explained by the covariates).

Alpha and tails.

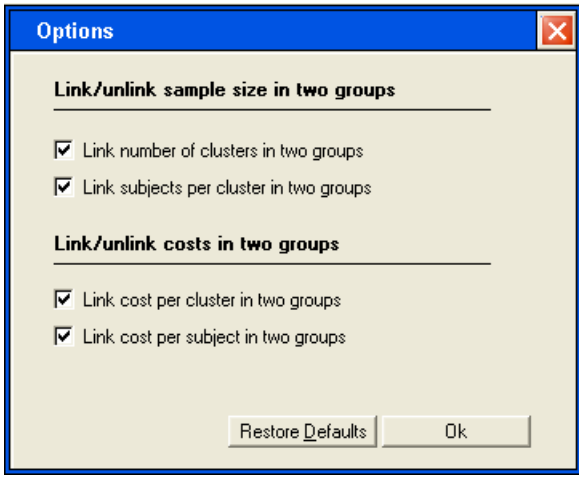
To modify these values, click on the values displayed, and the program will open a dialog box.

Linking/Unlinking the two conditions

By default, the program assumes that the number of clusters is the same in both conditions. If this is true, enter this value for the first condition, and the program will apply it to both conditions.

If the number is different for the two conditions, select Options > Link/Unlink groups and un-check “Link subjects per cluster in the two groups”.

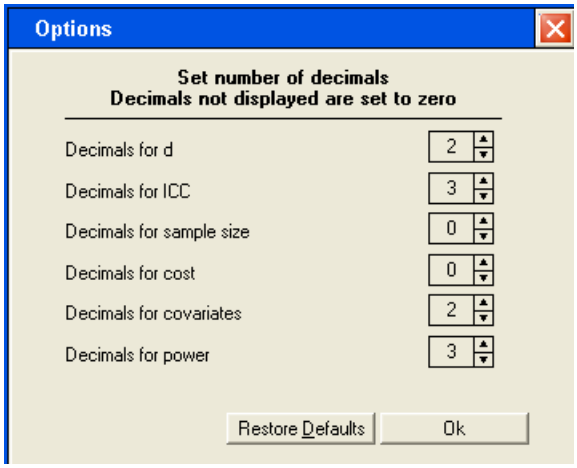
The same applies to the number of subjects within a cluster, to the cost of enrolling a new cluster, and to the cost of enrolling a new subject within a cluster. For each, the program allows you to link the values in the two groups, or to enter the value for each group independently of the other.



SETTING THE NUMBER OF DECIMALS

By default the program displays two decimal places for *d*, three for the *ICC*, and so on for the other parameters. To set the number of decimals displayed –

- Click Options > Decimals displayed



Most values can be adjusted using a spin button. This button will always adjust the least significant decimal place. If the value displayed is 0.005, then each click will increase or decrease the value by 0.001. If the value displayed is 0.05, then each click will increase or decrease the value by 0.01.

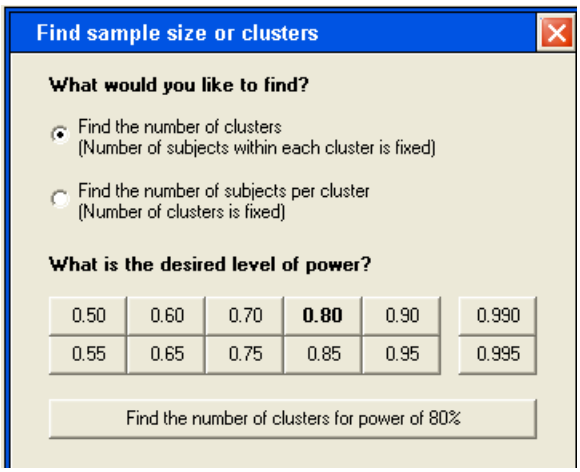
If you increase the number of decimals displayed (from 0.05 to 0.050) the value is not changed. If you decrease the number of decimals displayed (from 0.052 to 0.05) the value will be changed, and the new value will be displayed. Thus, the number displayed will always match the value that is actually being used in the computations.

FINDING POWER AND SAMPLE SIZE

Once you have entered the effect size (*d*), the *ICC*, the number of clusters and the number of subjects, the number of covariates and R^2 for each level, alpha and tails, the program displays the power.

You can modify one (or more) of these values to see the impact on power. For example, you may modify the number of clusters until power reaches the desired level (such as 80% or 90%).

Or, the program can find the required sample size automatically as explained here.



To find the number of clusters

- Enter a value for sample size within clusters.
- Click "Find Sample size or clusters" on the toolbar.
- ➔ Select "Find the number of clusters".
- Click a value such as 0.80.
- The program will display the number of clusters needed to yield power of 80%.

To find the number of subjects per cluster

- Enter a value for the number of clusters.
- Click "Find Sample size or clusters" on the toolbar.
- Select "Find the number of subjects per cluster".
- Click a value such as 0.80.
- The program will display the number of subjects per cluster needed to yield power of 80%.

Note. In a standard (non-clustered) trial, as long as the effect size is not zero, power will always approach 1.0 as the sample size approaches infinity. By contrast, in a cluster randomized trial, with *ICC* greater than zero, the maximum power that can be reached is limited by the *ICC*, effect size, and covariates. For a given set of these values, it will be impossible for power to exceed some value, regardless of the number of subjects per cluster.

If (for example) the maximum power is 0.60 and you try to find the sample size that will yield power of 0.80, the program will issue a warning and explain that you need to increase the number of clusters (or other parameters).

OPTIMAL DESIGN

For any set of parameters (the *ICC*, costs per cluster, cost per subject, cluster-level covariates, and subject-level covariates) there is a specific number of subjects per cluster that will yield the most cost-effective design.

- When the *ICC* is high the balance will shift toward a low number per cluster (since power is dominated by the number of clusters rather than the number of subjects). When the *ICC* is low, the balance will shift toward a high *n* per cluster (since power is controlled by both).

- At the same time, the balance is affected by the relative costs of adding additional clusters vs. adding additional subjects within a cluster. If the ratio is large (say, $10,000 vs. $100) it will be cost effective to add subjects. If it is small (say $200 vs. $100) it will be cost effective to add clusters.

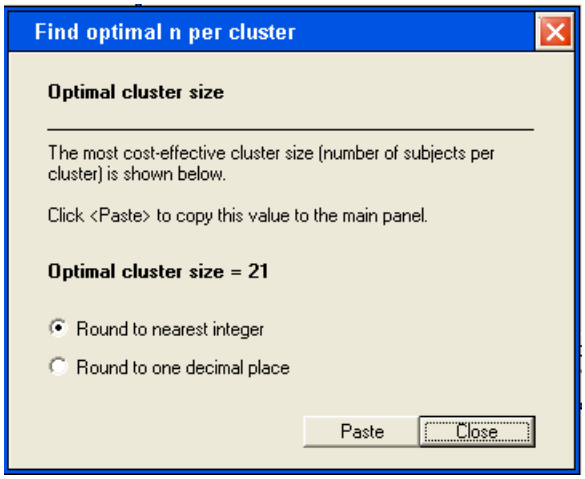
- The covariates also have an impact on the optimal sample size. Subject-level covariates serve to reduce the within-cluster error, and therefore reduce the need for a large *n* within clusters. As such, they shift the balance toward a lower *n* per cluster. By contrast, cluster-level covariates serve to reduce the between-cluster error, and therefore reduce the need for a large number of clusters. As such, they shift the balance toward a higher *n* per cluster.

Of course, these factors interact with each other. We need to balance the cost of a new cluster *and its impact on power* against the cost of a new subject *and its impact on power.*

Importantly, the relationship between subjects-per-cluster and cost is not monotonic. For example, suppose we start with a sample size of one per cluster, find the number of clusters needed for power of 90%, and compute the cost. Then we move to a sample size of two per cluster, find the number of clusters needed for power of 90%, and compute the cost, and so on. The cost will initially be high and will decline as the sample size is increased. At some point, however, the cost will begin to increase. The number of subjects per cluster when the cost curve reaches its lowest point is the optimal sample size.

The program can find the optimal sample size automatically, as follows.

- On the toolbar click “Find optimal N per cluster”.
- The program will open this dialog box and show the optimal *n*.



- Select one of the options (Round to nearest integer or Round to one decimal place) and click Paste.
- The program will paste this value into the main screen, and show the corresponding power.

Tip.

Now that you have the optimal *n* per cluster, click "Find sample size or clusters" and select the first option (Find the number of clusters).

- The program finds the number of clusters needed for the desired power.
- This design is the most cost-effective design that will yield the required power, given the values of the other parameters.

UNDERSTANDING HOW THE PARAMETERS AFFECT POWER

The program computes power for a set of values provided by the user. Since these values reflect a set of decisions and assumptions, it is helpful to understand how each affects power. In particular (as explained below) it is often useful to see how power would be affected if some of these values were modified.

The effect size, *d*

The standardized mean difference (*d*) is the effect size. As *d* increases in absolute value (the further it gets from zero), the power increases.

The *ICC*, number of clusters, and *n* within clusters

In a non-clustered design the standard error is determined largely by the sample size, n. In a clustered design the standard error is determined by the *ICC*, the number of clusters, the *n* within clusters *and the interaction among them*. If the *ICC* is low, increasing either the number of clusters or the *n* within clusters will increase power. If the *ICC* is high, increasing the number of clusters will increase power, but increasing sample size within clusters will have a more limited effect.

The *ICC*

A higher value for the *ICC* will tend to reduce power. Specifically, it will tend to increase the importance of the number of clusters and diminish the importance of the sample size within clusters.

Number of clusters

Increasing the number of clusters will always increase the power. The number of clusters sets an upper limit on the potential power, and increasing the *n* per cluster will not increase power beyond this point.

Number of subjects within a cluster

When the *ICC* is low, the *n* per cluster can have a large impact on power. When the *ICC* is relatively high, the *n* per cluster has a relatively modest impact on power.

Subject-level covariates

The –subject-level covariates reduce the within-cluster error. Increasing this R^2 will tend to increase power. The impact is similar to the impact of increasing the *n* within clusters. For example, the impact of this factor will be more important when the *ICC* is relatively low.

Cluster-level covariates

The cluster-level covariates reduce the between-cluster error. Increasing this R^2 will tend to increase power. The impact is similar to the impact of increasing the number of clusters. For example, the impact of this factor will be more important when the *ICC* is relatively high.

Alpha and tails

The role of alpha and tails in power for a clustered trial is the same as the role these play in a non-clustered trial. To wit –

As alpha is moved from 0.05 to 0.10, power will increase. As alpha is moved from 0.05 to 0.01 or 0.001, power will decrease.

The decision to use a one-tailed vs. a two-tailed test should be based on the nature of the research question. In the overwhelming majority of cases the two-tailed test is appropriate. This is true because even if we expect the effect to take a particular direction (we usually expect the treatment to improve the outcome) we would still interpret an effect that was statistically significant in the other direction.

That said, in the rare case where a one-tailed test is appropriate, it will yield higher power than a two-tailed test (provided, of course,that the true effect is in the expected direction).

PRECISION

In addition to displaying power, the program also displays the standard error of the estimate. In some cases, rather than focus only on the power of the test, we want to know how precisely we will be able to estimate the difference in means between the two treatments.

We can compute a confidence interval for the estimate by using the standard error. The 95% confidence interval will be given by the observed mean plus/minus *t* times the standard error, where *t* is based on the *t*-distribution with df equal to the number of clusters minus 2. With a large enough number of clusters, *t* will approach 1.96.

Note that the standard error is the expected value for the standard error. In half the samples it will be smaller than the expected value, and in half it will be larger.

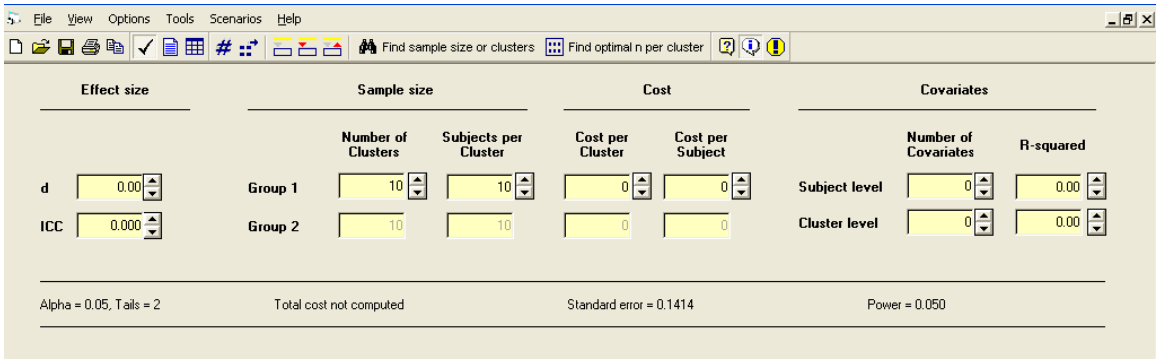
COST

The program automatically shows the cost for the planned study (provided the user has entered costs for each cluster and each subject). This can be helpful when using the interactive screen to consider alternate versions of the study plans.

EXAMPLE 1 – PATIENTS WITHIN HOSPITALS

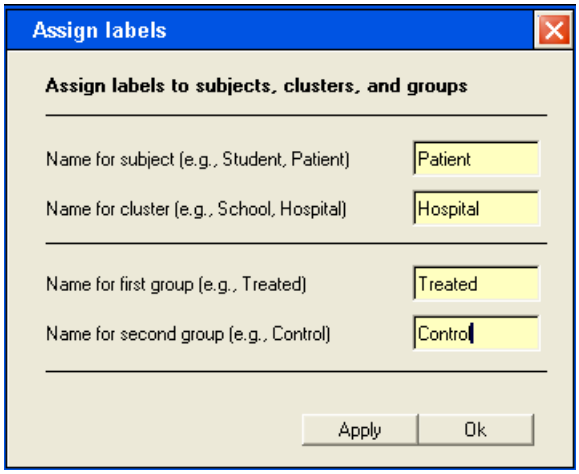
Suppose a researcher is planning a study to test the impact of an intervention on the pain reported by patients following surgery for hernia repair.

Patients in some hospitals (Control) will be given the standard set of instructions while patients in other hospitals (Treated) will be given a new set of instructions. Patients will be asked to record their pain (on a 10-point scale) for two weeks following surgery, and the mean pain reported by each patient will serve as that patient's score.

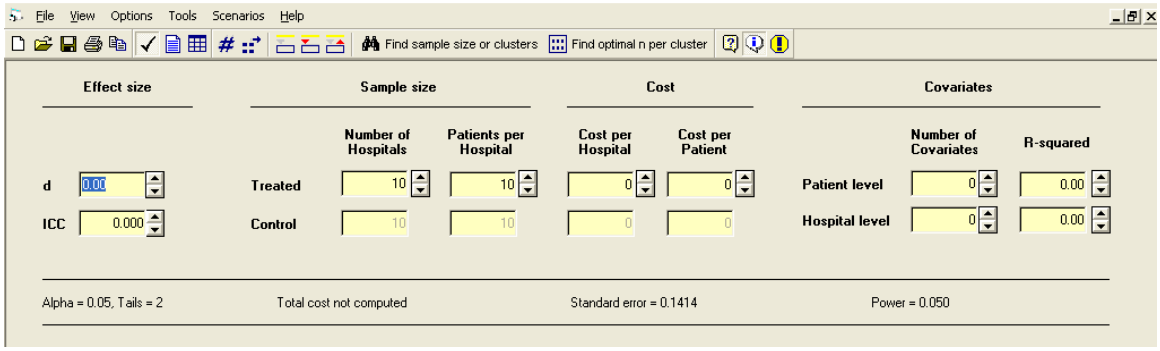


Name the groups

Click Tools > Assign labels and enter labels as follows



Click Ok and these labels are applied to the main screen



Effect size *d*

A pilot study showed that the mean pain level is 6.0, with a standard deviation of 3.0.

We decide that a clinically important effect would be to reduce the mean to 4.0. This yields an effect size (*d*) of 6 minus 4 (that is, 2.0) divided by 3, or 0.67.

ICC

The *ICC* is expected to fall near 0.10.

Number of hospitals and patients

As a starting point, we set the number of hospitals at 10 and the number of patients per hospital at 10.

Costs

The cost of enrolling each hospital is set at $1,000, and the cost of enrolling (and following) each patient is set at $50.

Covariates

Hospital-level covariates

Each hospital has a protocol for preparing the patients to deal with recovery from surgery. We expect that more time will be helpful in itself, and will also serve as an indicator of the general care level provided. The amount of time (in minutes) spent with each patient will serve as a hospital-level covariate.

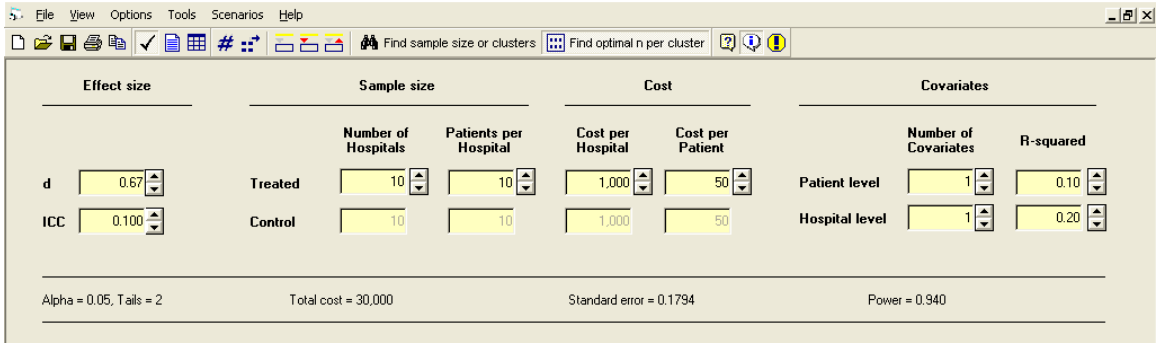
- For number of covariates enter 1,
- For R^2 enter 0.20.

Patient-level covariates

Experience has shown that older patients tend to report more pain following this procedure. Therefore, we plan to use each patient's age as a covariate, and we expect that this will explain some 10% of the variance in pain scores.

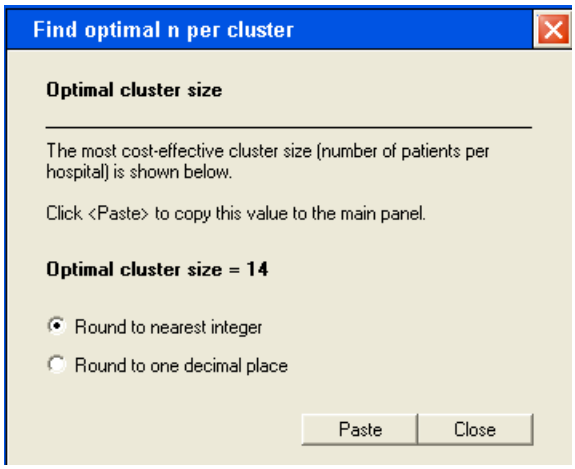
- For number of covariates enter 1.
- For R^2 enter 0.10.

At this point the main screen looks like this



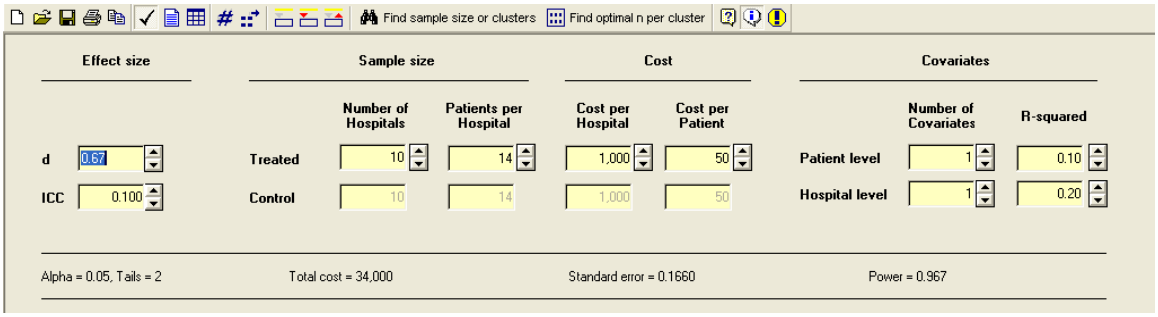
Find the optimal number of patients per hospital

- Click 'Find optimal sample size'



- The program shows that the optimal *n* is 14.
- Click Paste to copy this number into the main screen.

The main screen now looks like this.

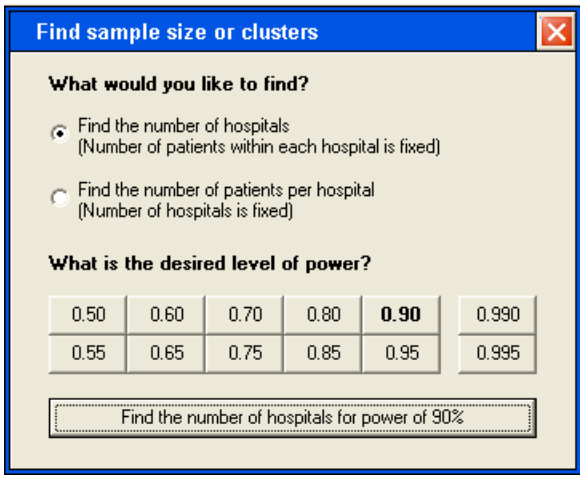


The number of clusters is still 10 (the arbitrary value we had set initially), the number of patients per hospital is 14 (the most cost-effective number). Power is shown as 0.967.

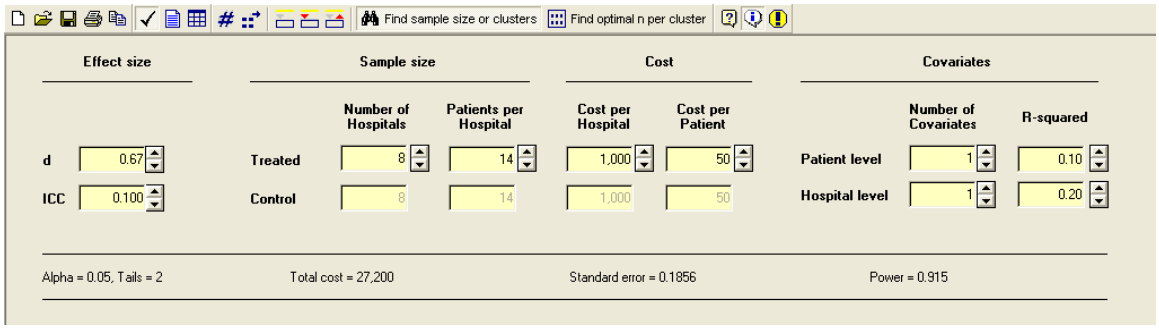
Find the number of clusters

Keeping the number of patients per hospital at 14, we need to find the number of hospitals that will yield power of 90%.

- Click “Find number of clusters”.
- Select the first option (“Find the number of hospitals”).
- Click 0.90.



- The program sets the number of clusters at 8.
- Power is shown as 91.5%.
- The study cost is shown as 27,200.
- The standard error is shown as 0.186.


Find sample size or clusters
Find optimal n per cluster
Effect size
Sample size
Cost
Covariates
Number of Hospitals
Patients per Hospital
Cost per Hospital
Cost per Patient
Number of Covariates
R-squared
d
0.67
Treated
8
14
1,000
50
Patient level
1
0.10
ICC
0.100
Control
8
14
1,000
50
Hospital level
1
0.20
Alpha = 0.05, Tails = 2
Total cost = 27,200
Standard error = 0.1856
Power = 0.915

Generate a report

To generate a report, click Report on the toolbar. The program generates the report shown here, which can be copied to Word™ or any Windows™ program.

Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the population means in the two groups (treated and control) are identical, or (equivalently) that the true effect size (d) is zero.

Study design. This hypothesis will be tested in a study that enrolls patients within hospitals.

Effect size. Power is computed for an effect size (d) of 0.67. The computations assume an intraclass correlation (ICC) of 0.100.

The standard deviation is assumed to be the same in the two groups.

Sample size. For each group we will enroll 8 hospitals with 14 patients per hospital for a total of 112 patients per group.

Patient-level covariates. There are 1 patient-level covariates. The R-squared between these covariates and outcome is assumed to be 0.10

Hospital-level covariates. There are 1 hospital-level covariates. The R-squared between these covariates and outcome is assumed to be 0.20

Alpha and Tails. The criterion for significance (alpha) has been set at 0.05. The test is 2-tailed, which means that an effect in either direction will be interpreted.

Power. Given these assumptions (for the effect size, ICC, and covariates), criteria (for alpha and tails), and plans (for the number of clusters and sample size within cluster), the study will have power of 91.5% to yield a statistically significant result.

Precison for estimating the effect size (d)

Precision. Given these same assumptions (for the ICC and covariates), and plans (for the number of clusters and sample size within cluster), the study will allow us to report the effect size (d)with a standard error of approximately 0.19.

Note that this is an expected (average) value for the standard error. The actual value in any given study will be lower or higher than this.

Disclaimer

This report is intended to help researchers use the program, and not to take the place of consultation with an expert statistician.

Cost

The projected costs for the treated group are as follows.

8 hospitals at 1,000 = 8,000.
14 patients per hospital (112 patients total) at 50 = 5,600.
Total cost for the treated group = 13,600.

The projected costs for the control group are as follows.

8 hospitals at 1,000 = 8,000.
14 patients per hospital (112 patients total) at 50 = 5,600.
Total cost for the control group = 13,600.

Total cost = 27,200.

Consider alternate assumptions

The power analysis is based on a series of decisions and assumptions. For example, we decide to "power" the study for an effect size of 0.67, to set alpha (two-tailed) at 0.05, and to require power of 90%. We assume that the *ICC* is 0.10 and that the proportion of variance explained by the hospital-level and patient-level covariates are 10% and 20%, respectively.

It is important to consider how power would be affected if some of these assumptions or decisions were changed. Or (from another perspective) it would be important to see what number of clusters would be needed to maintain power at 90% even if some of the assumptions or decisions were changed.

It is possible to do this working with the interactive screen. For example, if you change the *ICC* from 0.10 to 0.15, power moves from 91.5 to 84.2. Then, click "Find sample size" and the program shows that the number of clusters needed to maintain power of 90% increases from 8 to 10. The cost increases from 27,200 to 34,000.

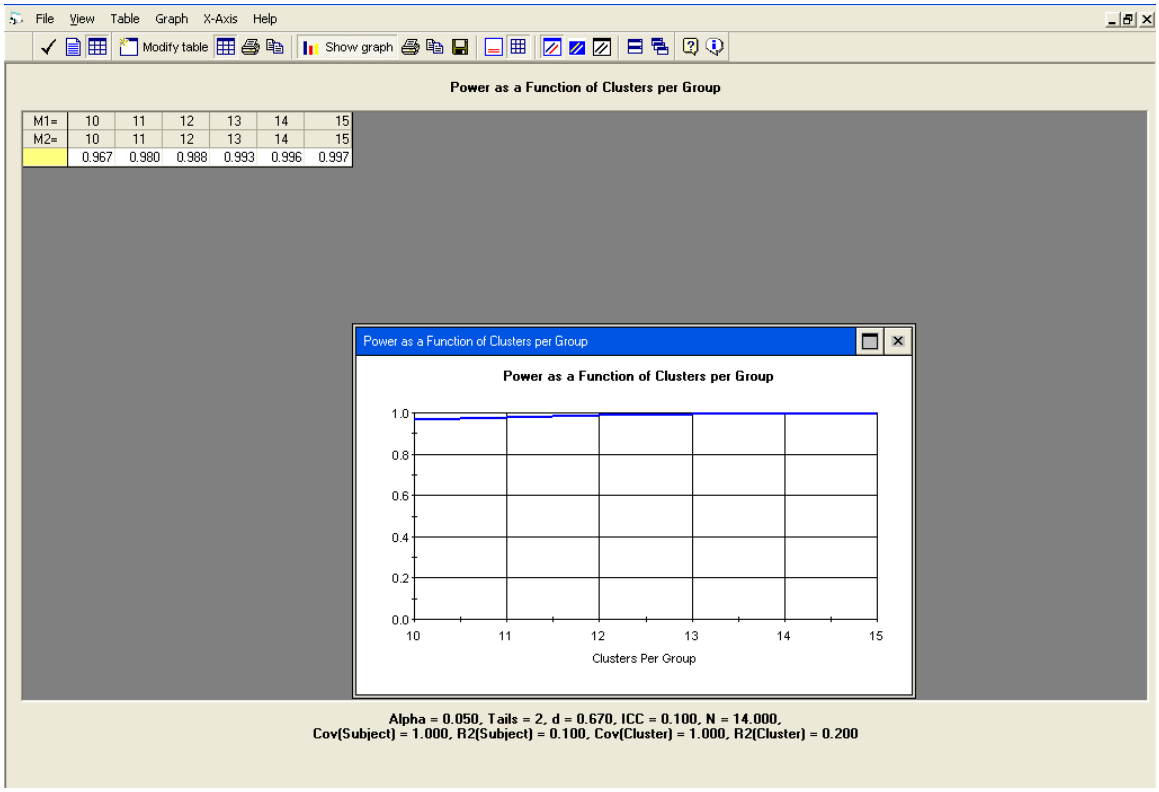
Create a table

The program also allows you to look at these issues systematically by using tables and graphs. First, enter all the values for effect size, *ICC*, and so on, as above.

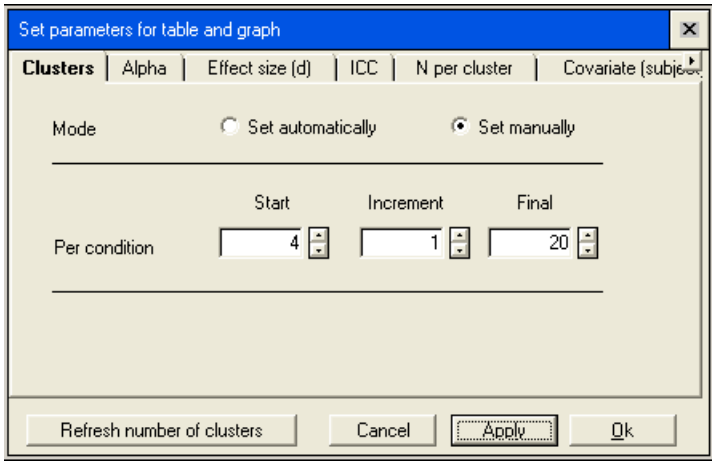
- Reset the *ICC* to 0.10.

- Then, click Tables on the toolbar.

The program immediately creates a table as shown here.

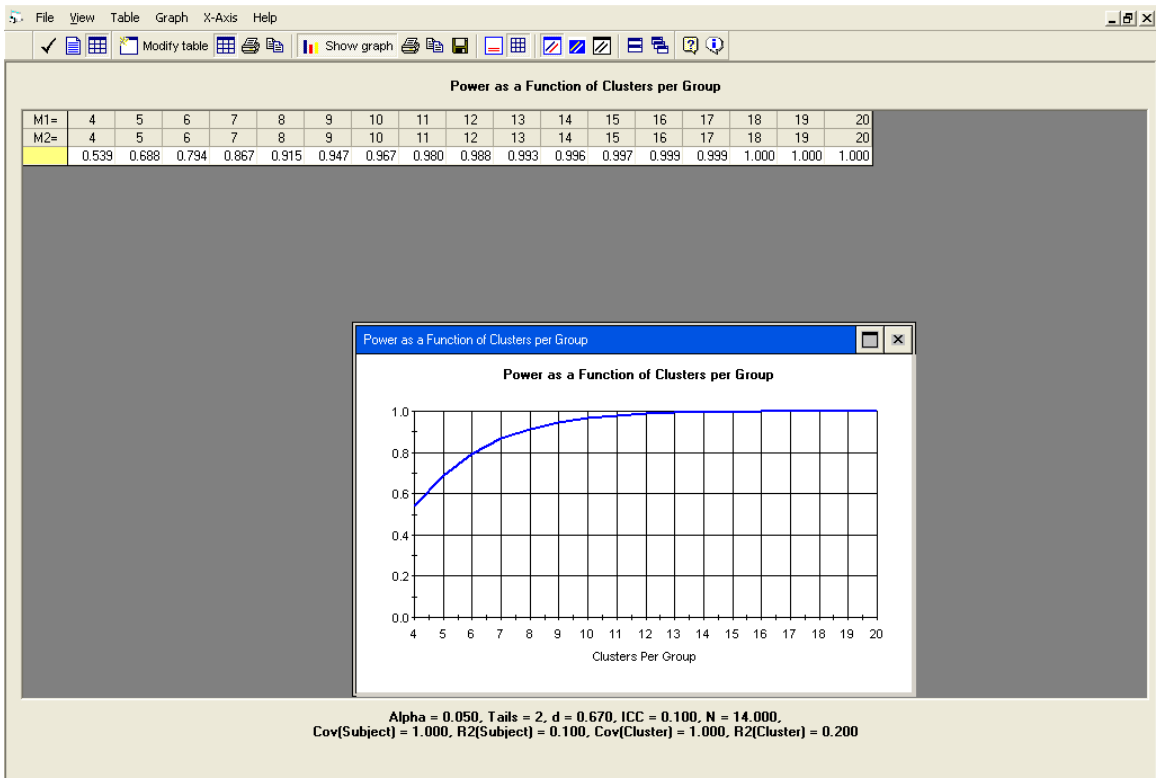


All parameters (the effect size, *ICC*, patient-level and hospital-level covariates, alpha, and tails) are taken from the interactive screen and displayed at the bottom of the table. The number of patients per hospital is taken from the interactive screen. The number of clusters varies from 10 to 15.



- Click Modify table.
- Select the tab for Clusters.

- Set the number of clusters to range from 4 to 20.
- Click Ok.

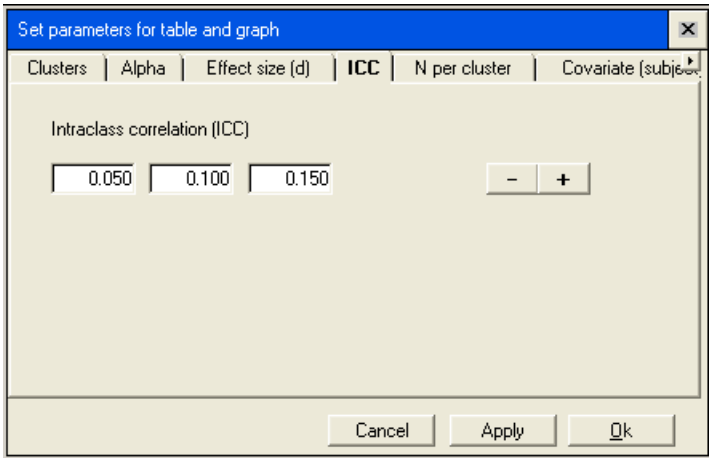


On the main screen, we had seen that we needed 8 hospitals to yield power of approximately 90%. Here, we see that we would need 6 hospitals to yield power of 80%, 8 hospitals to yield power of 90% (as before) and 9 hospitals to yield power of 95%. This provides a general sense of what our options would be if we wanted to think about lower or higher values of power.

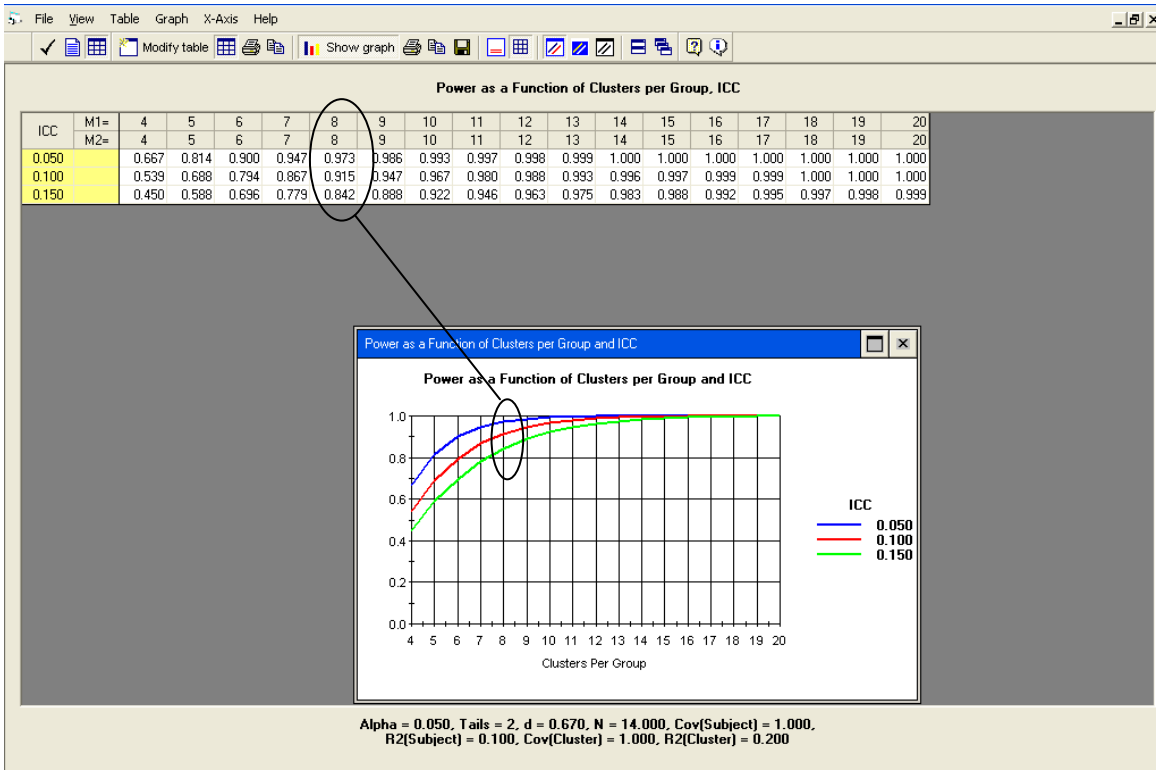
These computations all assume that the *ICC* is 0.10. What would happen if the *ICC* was actually somewhat lower or higher than this? The program can vary the *ICC* systematically and show the results.

Click Modify table

- Select the tab for *ICC.*
- The value is shown as 0.10, which was taken from the interactive screen.
- Click "+" two times, to add two more values for the *ICC.*
  Enter values of 0.05, 0.10, and 0.15.
- Click OK.



Now, the graph shows three lines, one for each value of the *ICC*.



This table offers an overview of our options.

We can “power” the study based on the original *ICC* of 0.10, and set the number of hospitals at 8.

Then, assuming all the other parameters are correct –

- If the *ICC* actually is 0.05, power will be 97%.
- If the *ICC* actually is 0.10, power will be 90% (as before).

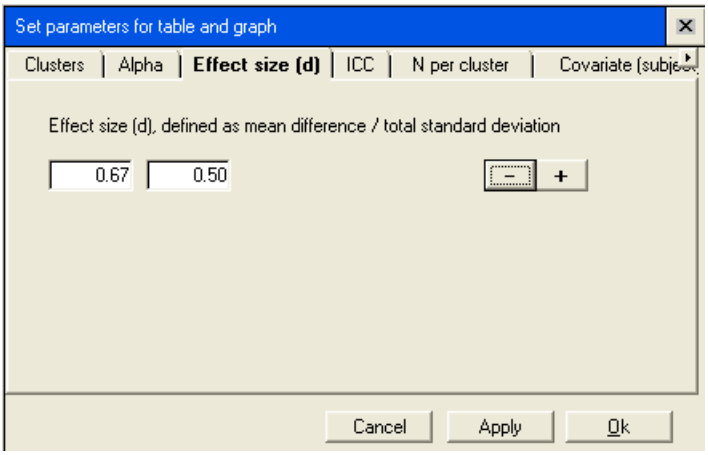
- If the *ICC* actually is 0.15, power will be 84%..

Or, we may want to power the study based on the *ICC* of 0.15 (that is, the worst case among the values being considered). We would set the number of hospitals at 10, to yield power of 90% even for this *ICC*. Then –

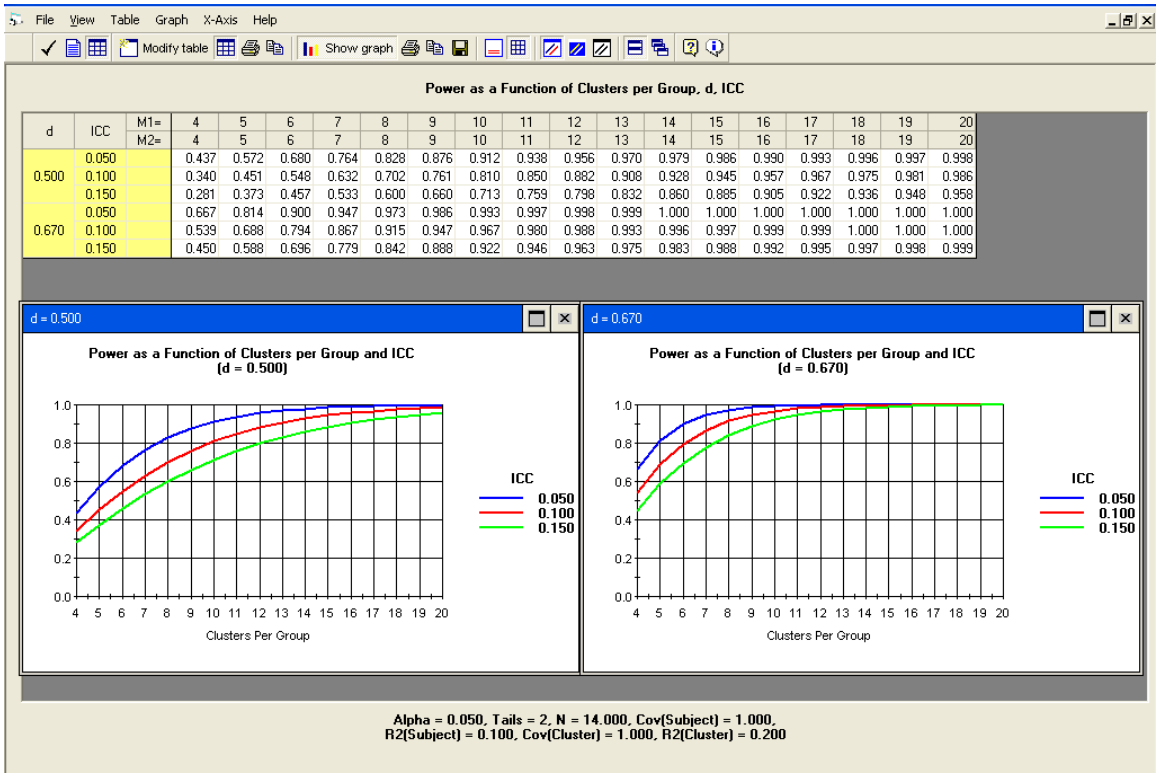
- If the *ICC* actually is 0.05, power will be 99%
- If the *ICC* actually is 0.10, power will be 97%.
- If the *ICC* actually is 0.15, power will be 92%.

The program also allows us to take account of several factors simultaneously. For example, we might want to use these three values of the *ICC*, and also two values for the effect size,

- Click Modify table.
- Select the tab for effect size.
- The value is shown as 0.67, which was taken from the interactive screen.
- Click "+" one time, to add one more value for *d.*
  Enter values of 0.67 and 0.50.
- Click OK.



The screen now looks like this (after clicking on Graph/Tile graphs).



d	ICC	M1= / M2=	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.500	0.050		0.437	0.572	0.680	0.764	0.828	0.876	0.912	0.938	0.956	0.970	0.979	0.986	0.990	0.993	0.996	0.997	0.998
	0.100		0.340	0.451	0.548	0.632	0.702	0.761	0.810	0.850	0.882	0.908	0.928	0.945	0.957	0.967	0.975	0.981	0.986
	0.150		0.281	0.373	0.457	0.533	0.600	0.660	0.713	0.759	0.798	0.832	0.860	0.885	0.905	0.922	0.936	0.948	0.958
0.670	0.050		0.667	0.814	0.900	0.947	0.973	0.986	0.993	0.997	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.100		0.539	0.688	0.794	0.867	0.915	0.947	0.967	0.980	0.988	0.993	0.996	0.997	0.999	0.999	1.000	1.000	1.000
	0.150		0.450	0.588	0.696	0.779	0.842	0.888	0.922	0.946	0.963	0.975	0.983	0.988	0.992	0.995	0.997	0.998	0.999

The graph at left is based on an effect size (*d*) of 0.50, and shows power for three values of the *ICC*. The graph at right is based on an effect size (*d*) of 0.67 (as before), and shows power for three values of the *ICC*.

If we want to power the study to ensure good power for an effect size of 0.50, we would use the graph at the left. To power the study for an effect size of 0.67, we would use the graph at right. In either case, we can see what happens if we want to plan for an *ICC* of 0.05, 0.10, or 0.15.

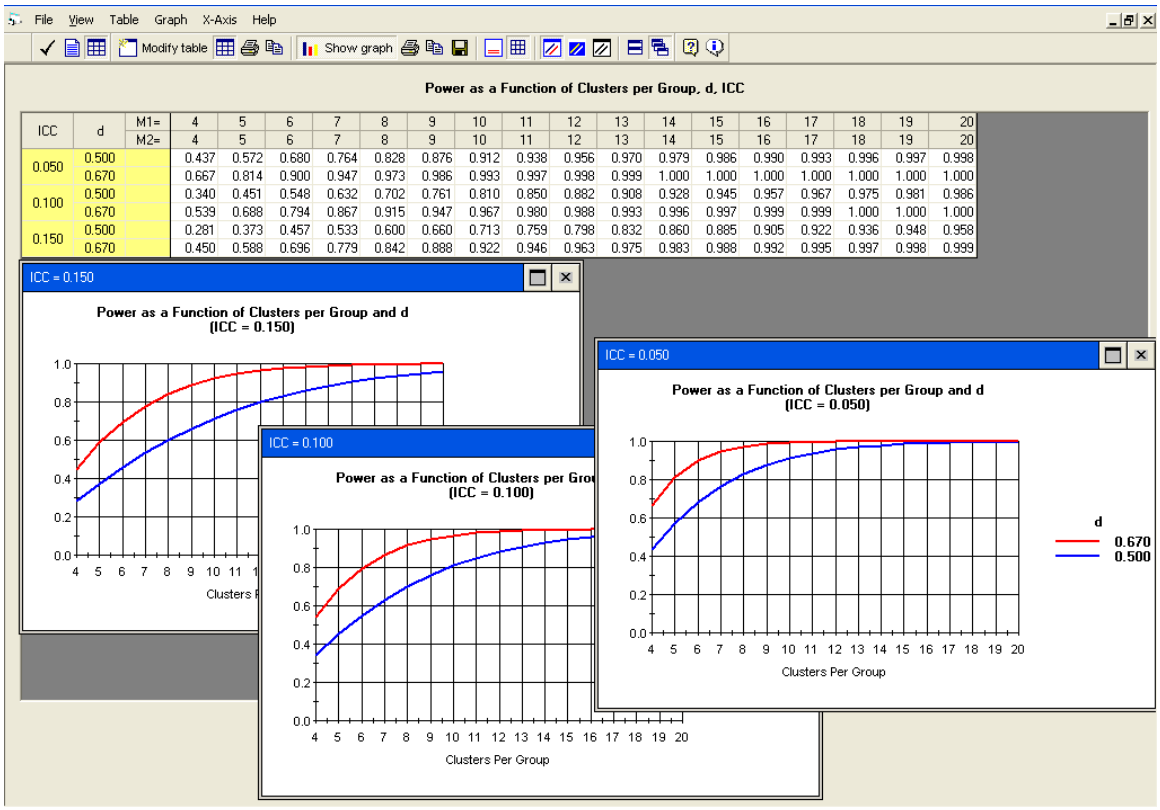
CUSTOMIZE THE GRAPHS

In this case each graph is based on one effect size (0.50 or 0.67), and the lines within the graph show the impact of the *ICC*. In some cases it would be helpful to have each graph reflect one *ICC*, and the lines within the graph show the impact of the effect size.

To make this change, proceed as follows.

The format of the graphs follows the sequence of columns in the table. In this table the sequence of columns is *d* followed by *ICC*, so each graph is based on one value of *d*, and the lines within a graph reflect the values of the *ICC*.

- Move one of the columns (grab the column heading that says *d* and move it to the right.



ICC	d	M1=	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
		M2=	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.050	0.500		0.437	0.572	0.680	0.764	0.828	0.876	0.912	0.938	0.956	0.970	0.979	0.986	0.990	0.993	0.996	0.997	0.998
	0.670		0.667	0.814	0.900	0.947	0.973	0.986	0.993	0.997	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.100	0.500		0.340	0.451	0.548	0.632	0.702	0.761	0.810	0.850	0.882	0.908	0.928	0.945	0.957	0.967	0.975	0.981	0.986
	0.670		0.539	0.688	0.794	0.867	0.915	0.947	0.967	0.980	0.988	0.993	0.996	0.997	0.999	0.999	1.000	1.000	1.000
0.150	0.500		0.281	0.373	0.457	0.533	0.600	0.660	0.713	0.759	0.798	0.832	0.860	0.885	0.905	0.922	0.936	0.948	0.958
	0.670		0.450	0.588	0.696	0.779	0.842	0.888	0.922	0.946	0.963	0.975	0.983	0.988	0.992	0.995	0.997	0.998	0.999

- Now, the table looks like this.
- There is one graph for each *ICC*, and two lines within each graph, reflecting the two values of *d*.

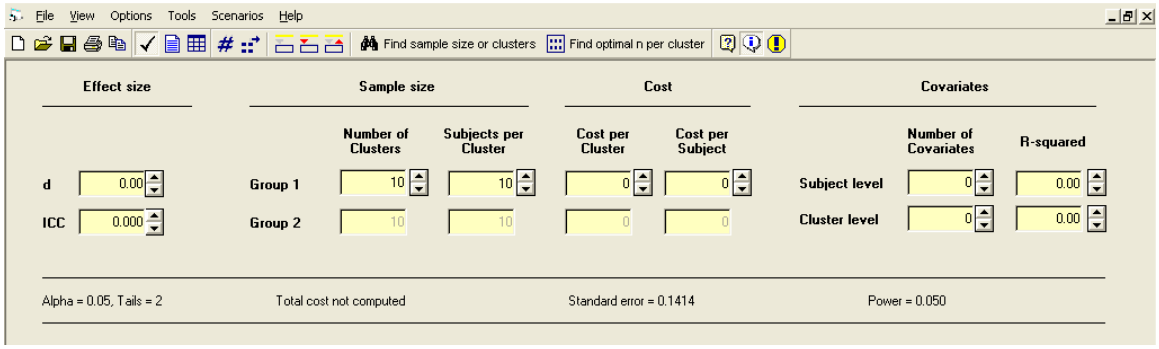
These graphs show that, for any value of the *ICC*, power drops some 10-15 points if we assume an effect size of 0.50 rather than 0.67. Put another way, to power the study for an effect size of 0.50, we would need to add about five hospitals. Using an *ICC* of 0.10 as an example, for power of 90%, with *d* = 0.67 we need 8 hospitals but with *d* = 0.50 we need 13.

Similarly, click Modify table to add any other factor(s) to the table and graphs.

EXAMPLE 2 – STUDENTS WITHIN SCHOOLS

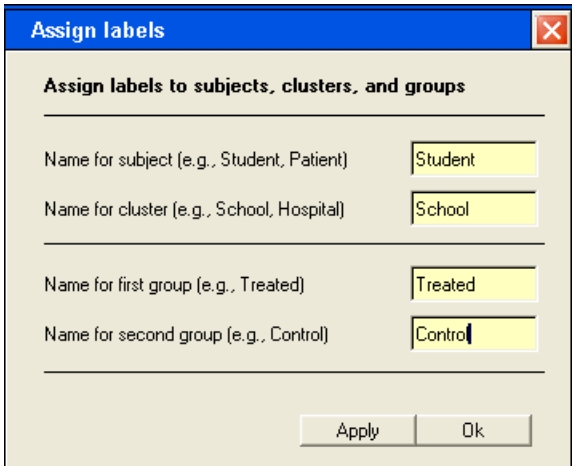
Suppose a researcher is planning a study to test the impact of an intervention on the reading scores of students in the fifth grade.

Students in some schools (Control) will be given the standard curriculum while students in other schools will be given a revised curriculum (Treated). At the end of the school year, reading scores will be assessed using a standardized test.

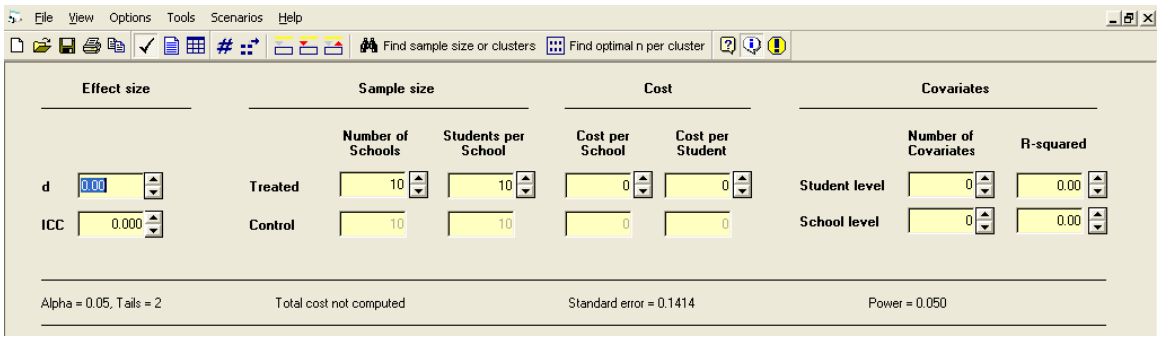


Name the groups

Click Tools > Assign labels and enter labels as follows



Click Ok and these labels are applied to the main screen



Effect size *d*

A pilot study showed that the mean reading score is 70, with a standard deviation of 20.

We decide that a clinically important effect would be to increase the mean to 75. This yields an effect size (*d*) of 75 minus 70 (that is, 5) divided by 20, or 0.25.

ICC

The *ICC* is expected to fall near of 0.30.

Number of schools and students

As a starting point, we set the number of schools at 10 and the number of students per school at 10.

Costs

The cost of enrolling each school is set at $2,500, and the cost of enrolling (and following) each student is set at $20.

Covariates

School-level covariates

The class takes the same standardized test every year. The mean score for the fifth grade class at the end of the prior year will serve as a school-level covariate.

- For number of covariates enter 1.
- For R^2 enter 0.20.

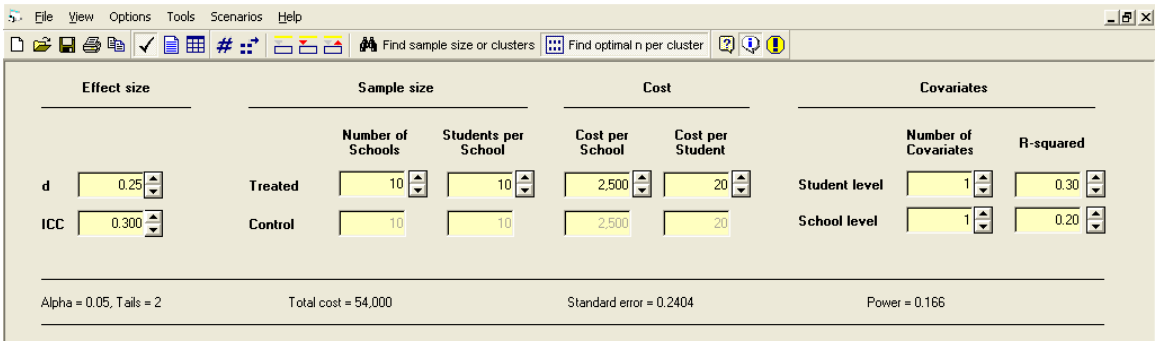
Student-level covariates

For each student entering the fifth grade (and included in the study), the student's reading score from the end of the prior year (fourth grade) will serve as a student-level covariate.

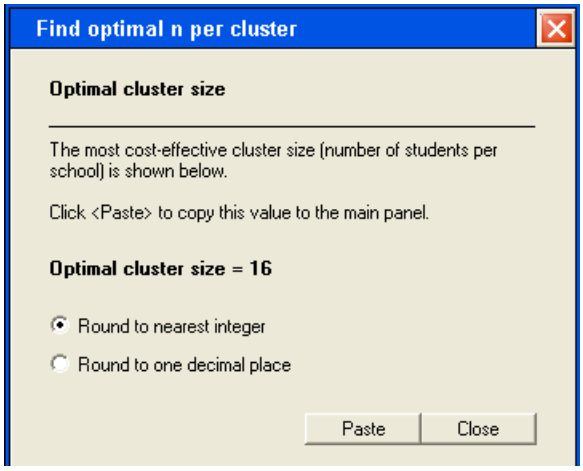
- For number of covariates enter 1.

- For R^2 enter 0.30.

At this point the main screen looks like this

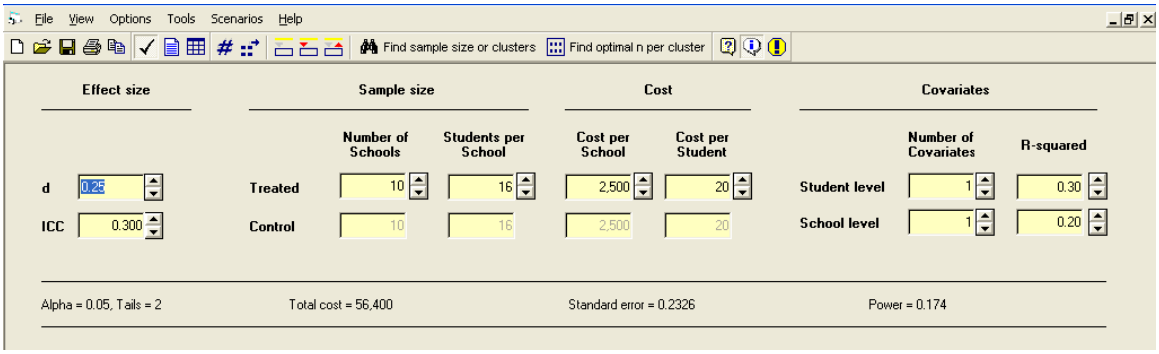


Find the optimal number of students per school



- Click “Find optimal sample size”.
- The program shows that the optimal *n* is 16.
- Click Paste to copy this number into the main screen.

The main screen now looks like this.

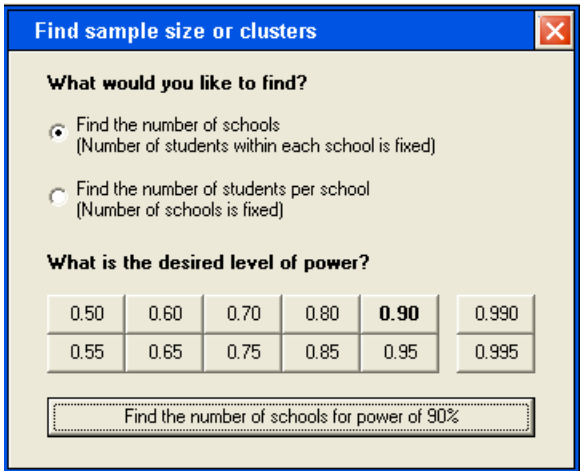


The number of schools is still 10 (the arbitrary value we had set initially), the number of students per school is 16 (the most cost-effective number). Power is shown as 0.174

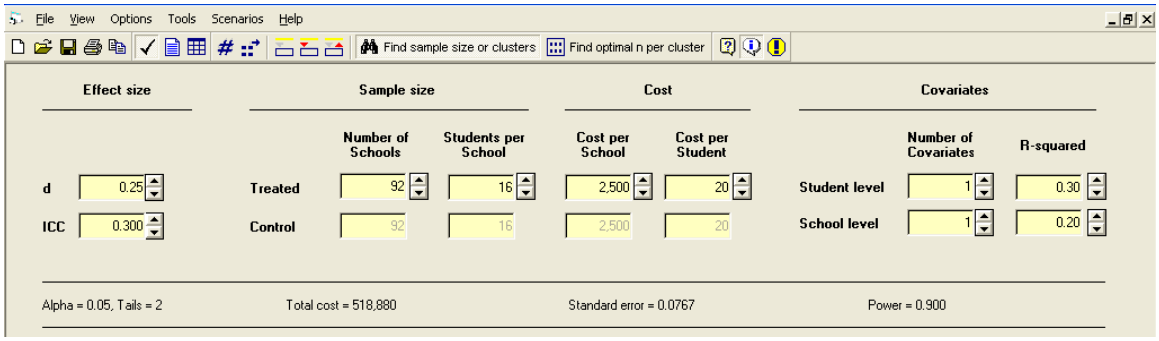
Find the number of schools

Keeping the number of students per school at 16, we need to find the number of schools that will yield power of 90%.

- Click “Find number of schools”.
- Select the first option.
- Click 0.90.



- The program sets the number of schools at 92.
- Power is shown as 90%.
- The study cost is shown as 518,880.
- The standard error is shown as 0.0767.



Generate a report

To generate a report, click Report on the toolbar. The program generates the report shown here, which can be copied to Word™ or any Windows™ program.

Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the population means in the two groups (treated and control) are identical, or (equivalently) that the true effect size (d) is zero.

Study design. This hypothesis will be tested in a study that enrolls students within schools.

Effect size. Power is computed for an effect size (d) of 0.25. The computations assume an intraclass correlation (ICC) of 0.300.

The standard deviation is assumed to be the same in the two groups.

Sample size. For each group we will enroll 92 schools with 16 students per school for a total of 1,472 students per group.

Student-level covariates. There are 1 student-level covariates. The R-squared between these covariates and outcome is assumed to be 0.30

School-level covariates. There are 1 school-level covariates. The R-squared between these covariates and outcome is assumed to be 0.20

Alpha and Tails. The criterion for significance (alpha) has been set at 0.05. The test is 2-tailed, which means that an effect in either direction will be interpreted.

Power. Given these assumptions (for the effect size, ICC, and covariates), criteria (for alpha and tails), and plans (for the number of clusters and sample size within cluster), the study will have power of 90.0% to yield a statistically significant result.

Precison for estimating the effect size (d)

Precision. Given these same assumptions (for the ICC and covariates), and plans (for the number of clusters and sample size within cluster), the study will allow us to report the effect size (d)with a standard error of approximately 0.08.

Note that this is an expected (average) value for the standard error. The actual value in any given study will be lower or higher than this.

Disclaimer

This report is intended to help researchers use the program, and not to take the place of consultation with an expert statistician.

Cost

The projected costs for the treated group are as follows.

92 schools at 2,500 = 230,000.
16 students per school (1472 students total) at 20 = 29,440.
Total cost for the treated group = 259,440.

The projected costs for the control group are as follows.

92 schools at 2,500 = 230,000.
16 students per school (1472 students total) at 20 = 29,440.
Total cost for the control group = 259,440.

Total cost = 518,880.

Consider alternate assumptions

The power analysis is based on a series of decisions and assumptions. For example, we decide to "power" the study for an effect size of 0.25, to set alpha (two-tailed) at 0.05, and to require power of 90%. We assume that the *ICC* is 0.30 and that the proportions of variance explained by the school-level and student-level covariates are 20% and 30%, respectively.

It is important to consider how power would be affected if some of these assumptions or decisions were changed. Or (from another perspective) it would be important to see what number of schools would be needed to maintain power at 90% even if some of the assumptions or decisions were changed.

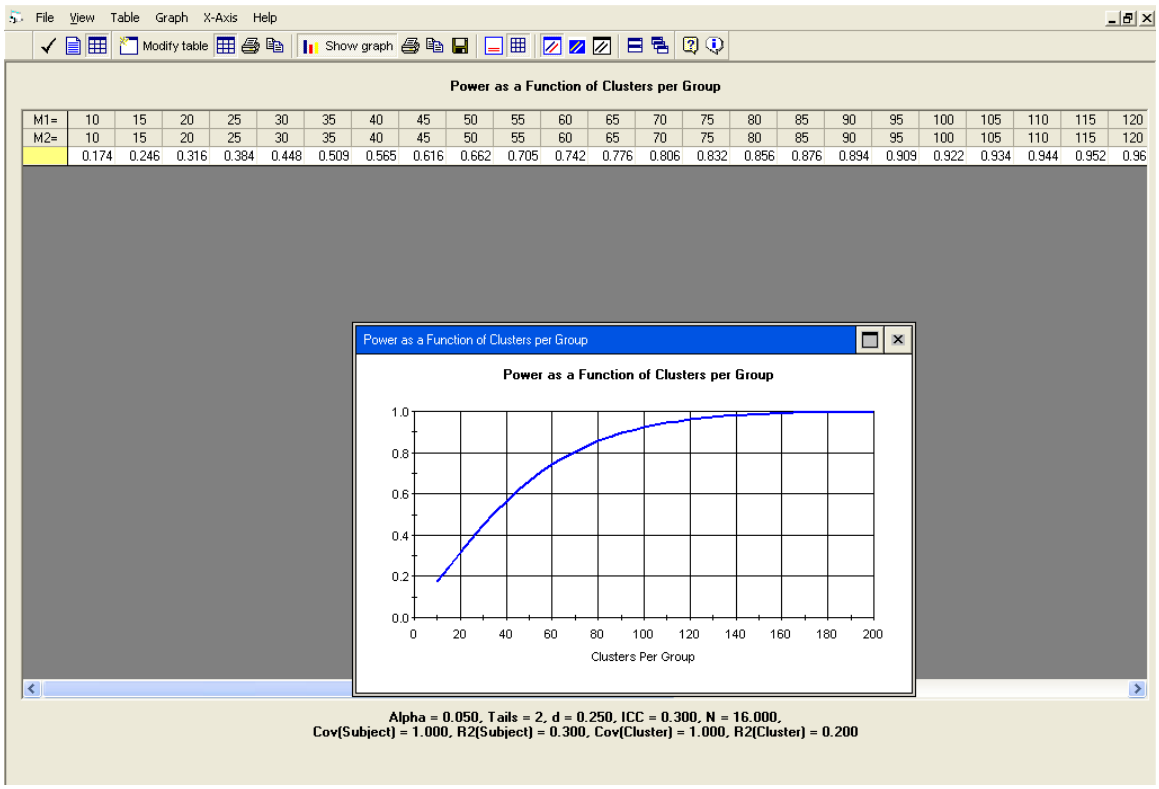
It is possible to do this working with the interactive screen. For example, if you change the *ICC* from 0.30 to 0.35, power moves from 90 to 86. Then, click "Find sample size" and the program shows that the number of schools needed to maintain power of 90% increases from 92 to 105. The cost increases from 518,880 to 592,200.

Create a table

The program also allows you to look at these issues systematically by using tables and graphs. First, enter all the values for effect size, *ICC*, and so on, as above.

- Reset the *ICC* to 0.30.
- Then, click “Tables” on the toolbar.

The program immediately creates a table as shown here.



All parameters (the effect size, *ICC*, student-level and school-level covariates, alpha, and tails) are taken from the interactive screen and displayed at the bottom of the table. The number of students per school is taken from the interactive screen. The number of schools varies from 10 to 200.

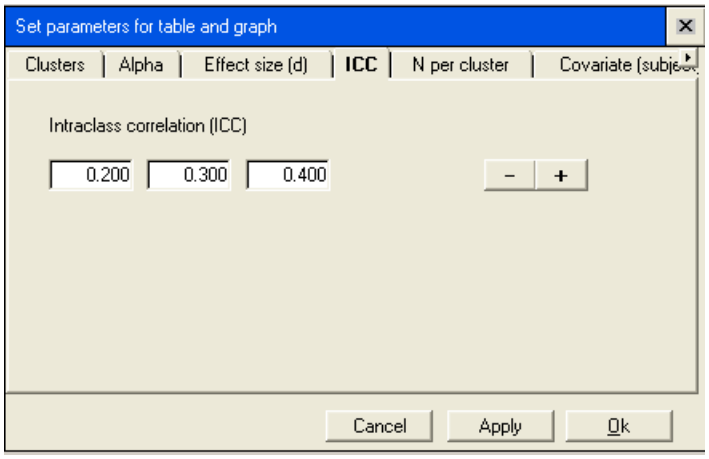
On the main screen, we had seen that we needed 92 schools to yield power of approximately 90%. Here, we see that we would need 70 schools to yield power of 80%, 92 schools to yield power of 90% (as before) and 115 schools to yield power of 95%. This provides a general sense of what our options would be if we wanted to think about lower or higher values of power.

These computations all assume that the *ICC* is 0.30. What would happen if the *ICC* was actually somewhat lower or higher than this? The program can vary the *ICC* systematically and show the results.

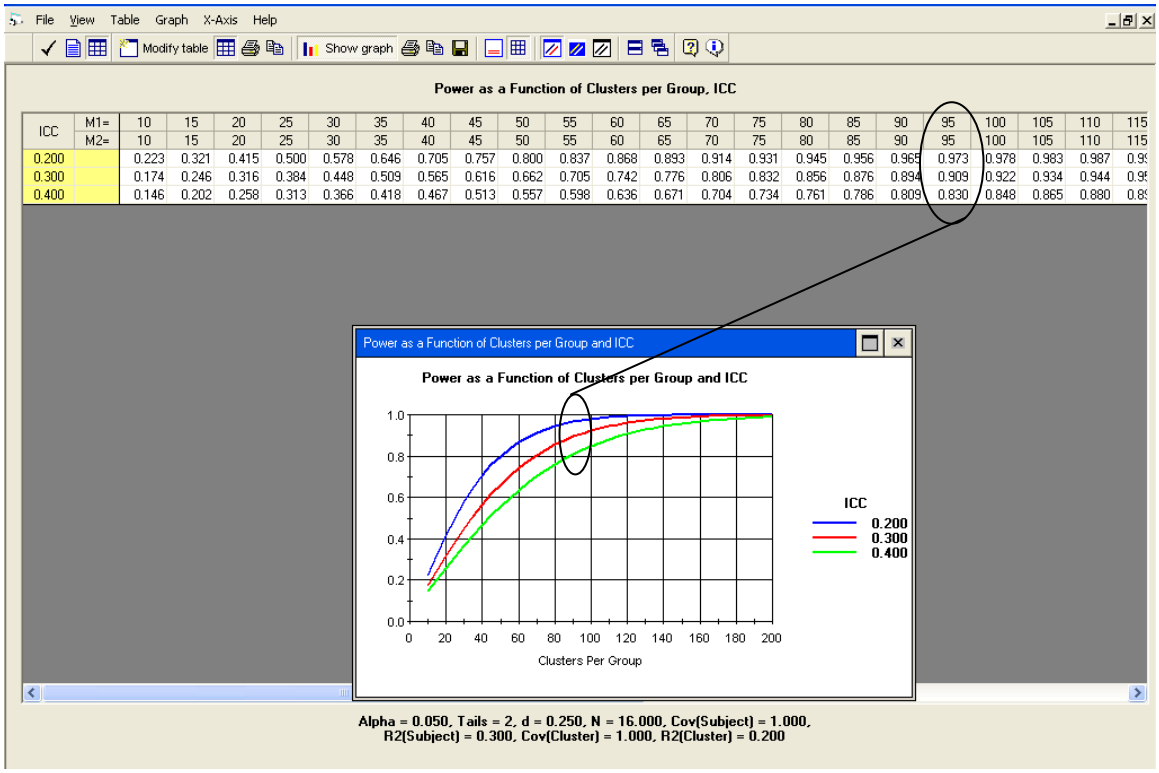
Click Modify table

- Select the tab for *ICC.*
- The value is shown as 0.30, which was taken from the interactive screen.

- Click “+” two times, to add two more values for the *ICC.*
  Enter values of 0.20, 0.30, and 0.40.
- Click OK.



Now, the graph shows three lines, one for each value of the *ICC*. This graph provides the following information.



This table offers an overview of our options.

We can “power” the study based on the original *ICC* of 0.30, and set the number of schools at 92.

Then, assuming all the other parameters are correct –

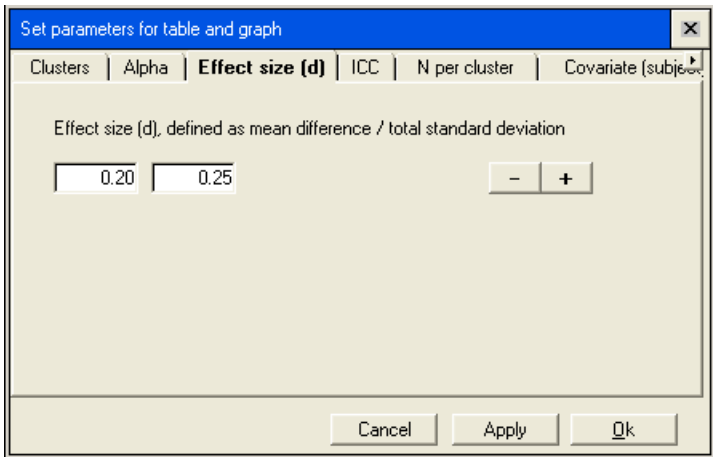
- If the *ICC* actually is 0.20, power will be 97%.
- If the *ICC* actually is 0.30, power will be 90% (as before).
- If the *ICC* actually is 0.40, power will be 81%..

Or, we may want to power the study based on the *ICC* of 0.40 (that is, the worst case among the values being considered). We would set the number of schools at 120, to yield power of 90% even for this *ICC*. Then –

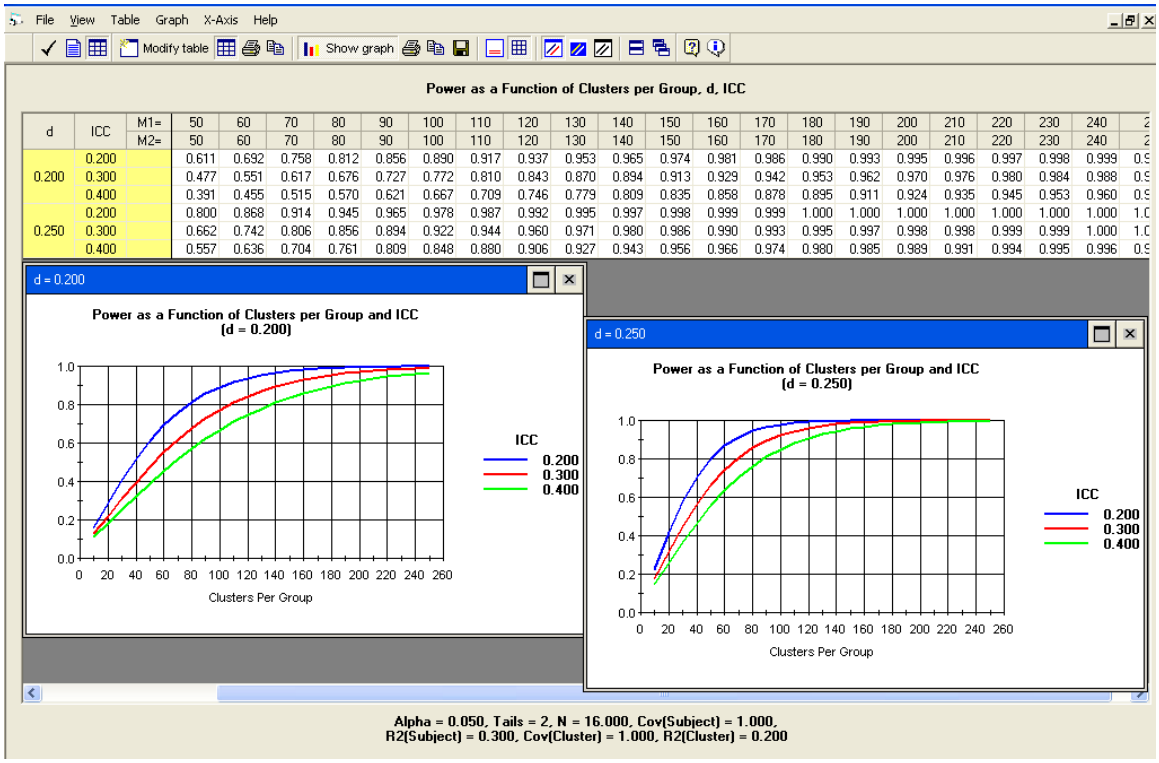
- If the *ICC* actually is 0.20, power will be 99%
- If the *ICC* actually is 0.30, power will be 96%.
- If the *ICC* actually is 0.40, power will be 90%.

The program also allows us to take account of several factors simultaneously. For example, we might want to use these three values of the *ICC*, and also two values for the effect size,

- Click Modify table.
- Select the tab for effect size.
- The value is shown as 0.25, which was taken from the interactive screen.
- Click "+" one time, to add one more value for *d.*
  Enter values of 0.25 and 0.20.
- Click OK.



The screen now looks like this (after arranging the position of the graphs).



d	ICC	M1=	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240
		M2=	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240
0.200	0.200		0.611	0.692	0.758	0.812	0.856	0.890	0.917	0.937	0.953	0.965	0.974	0.981	0.986	0.990	0.993	0.995	0.996	0.997	0.998	0.999
	0.300		0.477	0.551	0.617	0.676	0.727	0.772	0.810	0.843	0.870	0.894	0.913	0.929	0.942	0.953	0.962	0.970	0.976	0.980	0.984	0.988
	0.400		0.391	0.455	0.515	0.570	0.621	0.667	0.709	0.746	0.779	0.809	0.835	0.858	0.878	0.895	0.911	0.924	0.935	0.945	0.953	0.960
0.250	0.200		0.800	0.868	0.914	0.945	0.965	0.978	0.987	0.992	0.995	0.997	0.998	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.300		0.662	0.742	0.806	0.856	0.894	0.922	0.944	0.960	0.971	0.980	0.986	0.990	0.993	0.995	0.997	0.998	0.998	0.999	0.999	1.000
	0.400		0.557	0.636	0.704	0.761	0.809	0.848	0.880	0.906	0.927	0.943	0.956	0.966	0.974	0.980	0.985	0.989	0.991	0.994	0.995	0.996

The graph at left is based on an effect size (*d*) of 0.20, and shows power for three values of the *ICC*. The graph at right is based on an effect size (*d*) of 0.25 (as before), and shows power for three values of the *ICC*.

If we want to power the study to ensure good power for an effect size of 0.20, we would use the graph at the left. To power the study for an effect size of 0.25, we would use the graph at right. In either case, we can see what happens if we want to plan for an *ICC* of 0.20, 0.30, or 0.40.

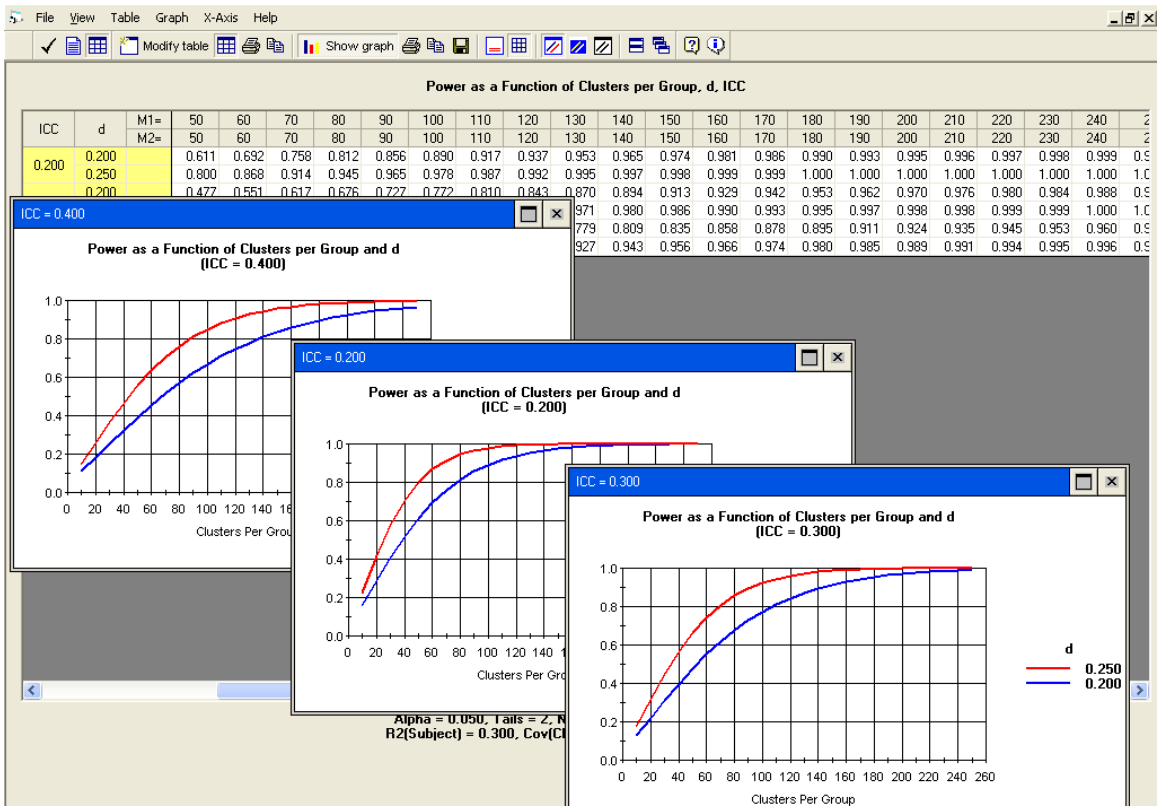
CUSTOMIZE THE GRAPHS

In this case each graph is based on one effect size (0.20 or 0.25), and the lines within the graph show the impact of the *ICC*. In some cases it would be helpful to have each graph reflect one *ICC*, and the lines within the graph show the impact of the effect size.

To make this change, proceed as follows.

The format of the graphs follows the sequence of columns in the table. In this table the sequence of columns is *d* followed by *ICC*, so each graph is based on one value of *d*, and the lines within a graph reflect the values of the *ICC*.

- Move one of the columns (grab the column heading that says *d* and move it to the right.



- Now, the table looks like this.
- There is one graph for each *ICC*, and two lines within each graph, reflecting the two values of *d*.

These graphs show that, for any value of the *ICC*, power drops some 20 points if we assume an effect size of 0.20 rather than 0.25. Put another way, to power the study for an effect size of 0.20 rather than 0.25, we would need to add about 50 schools. Using an *ICC* of 0.30 as an example, for power of 90%, with *d* = 0.25 we need 92 schools but with *d* = 0.20 we need 145.

Similarly, click Modify table to add any other factor(s) to the table and graphs.

APPENDIX – COMPUTATIONAL FORMULAS

Two-level Hierarchical Designs With Covariates

Suppose we have a two-level design (individuals within clusters such as schools) in which treatments are assigned at the highest level, the cluster level (level 2). Suppose that there are m clusters per treatment and n individuals per cluster. Suppose that there are covariates at both levels of the design. In particular suppose that there are q_2 covariates at the cluster level (level 2) that explain the proportion R_2^2 of the variance at the cluster level (so the cluster level multiple correlation is R_2), and q_1 covariates at the individual level (level 1) that explain the proportion R_1^2 of the variance at the individual-within-cluster level (so the individual level multiple correlation is R_1).

Suppose that, before adjustment for any covariates, the between-cluster and between-individual-within-cluster variance components are σ_2^2 and σ_1^2, respectively. Denote the covariate adjusted the between-cluster and between-individual-within-cluster variance components by σ_{A2}^2 and σ_{A1}^2, respectively. Thus we can define R_2^2 and R_1^2 as

$$R_2^2 = 1 - \sigma_{A2}^2 / \sigma_2^2$$

and

$$R_1^2 = 1 - \sigma_{A1}^2 / \sigma_1^2.$$

Define the intraclass correlation ρ (at the cluster level) by

$$\rho = \frac{\sigma_2^2}{\sigma_2^2 + \sigma_1^2}. \tag{1.1}$$

Note that ρ is the proportion of total variance at the cluster level. The quantity $\bar{\rho} = 1 - \rho$ is analogous to it in representing the proportion of the total variance that is at the individual level.

When cluster sizes n are equal, the test for treatment effects in two-level cluster randomized experiments is an exact t-test. The test statistic has a non-central t-distribution with $2m - 2 - q_2$ degrees of freedom and covariate adjusted noncentrality parameter

$$\lambda = d_T \sqrt{\frac{nm}{2}} \sqrt{\frac{1}{1 + (n-1)\rho - \left[R_1^2 + \left(nR_2^2 - R_1^2 \right) \rho \right]}} \tag{1.2}$$

where there are m clusters in each of the control and treatment groups, q_2 is the number of covariates at the cluster level (level 2), d_T is the effect size (the difference between the treatment and control group means divided by the total within-group standard deviation),

and ρ is the (unadjusted) intraclass correlation. Note that the fact that we sample from a finite population of clusters has no impact on the power of the test for treatment effects. The reason is that, although the variance of the school-effects components in the treatment and control groups means is smaller when a finite population of schools is assigned to treatments, these school effects components are negatively correlated. Because the variance of the mean difference is the sum of the variance minus twice the covariance, the negative covariance term increases the variance of the difference and exactly cancels the reduction in variance due to finite population sampling.

Note that the maximum value of the noncentrality parameter as the cluster size $n \to \infty$ with fixed m is

$$\lambda_{Max} = \delta\sqrt{\frac{m}{2}}\sqrt{\frac{1}{\left(1-R_2^2\right)\rho}}. \tag{1.3}$$

Of course, the maximum value of the noncentrality parameter tends to infinity as $m \to \infty$. These maxima are useful for computing the maximum power that can be obtained by increasing n with other design parameters fixed (for example in determining whether *any* value of n can attain a desired power).

The power of a level α one-tailed test for the treatment effect is therefore

$$p_1 = 1 - \mathrm{f}(c_\alpha, 2m - 2 - q_2\ , \lambda),$$

where f(x, v, λ) is the cumulative distribution function of the noncentral t-distribution with v degrees of freedom, and noncentrality parameter λ and c_α is the 100(1 – α) percentile of the central t-distribution with $2m - 2 - q_2$ degrees of freedom. The power of a level α two-tailed test for the treatment effect is therefore

$$p_2 = 1 - \mathrm{f}(c_{\alpha/2}, 2m - 2 - q_2\ , \lambda) + \mathrm{f}(-c_{\alpha/2}, 2m - 2 - q_2\ , \lambda).$$

Unequal Sample Sizes

When the numbers of observations in each cluster are not equal within each treatment group, the test is no longer exact. However if there are m^T clusters in the treatment group and m^C clusters in the control group, the test statistic has approximately a noncentral t-distribution, with $m^C + m^T - 2 - q_2$ degrees of freedom and covariate adjusted noncentrality parameter

$$\lambda = d_T\sqrt{\frac{N^C N^T}{N^C + N^T}}\sqrt{\frac{1}{1+\left(\tilde{n}_U - 1\right)\rho - \left[R_1^2 + \left(\tilde{n}_U R_2^2 - R_1^2\right)\rho\right]}} \tag{1.4}$$

where N^C and N^T are the total number of observations in the control and treatment groups, respectively, and

$$\tilde{n}_U = \frac{N^C \sum_{i=1}^{m^T} \left(n_i^T\right)^2}{N^T N} + \frac{N^T \sum_{i=1}^{m^C} \left(n_i^C\right)^2}{N^C N} \tag{1.5}$$

where n_i^T is the number of observations in the i^{th} cluster of the treatment group and n_i^C is the number of observations in the i^{th} cluster of the control group. Note that, when cluster sizes are equal within treatment groups so that $n_i^C = n^C$, $i = 1, \ldots, m^C$ and $n_i^T = n^T$, $i = 1, \ldots, m^T$, $\tilde{n}_U$ in () reduces to

$$\tilde{n}_U = \frac{n^C n^T \left(m^C + m^T\right)}{m^C n^C + m^T n^T} . \tag{1.6}$$

Note that, when $n^C = n^T = n$, $\tilde{n}_U$ in () reduces to n and () reduces to (1).

The power of a level α one-tailed test for the treatment effect is therefore

$$p_1 = 1 - \mathrm{f}(c_\alpha, m^C + m^T - 2 - q_2, \lambda),$$

where $\mathrm{f}(x, \mathrm{v}, \lambda)$ is the cumulative distribution function of the noncentral t-distribution with v degrees of freedom and noncentrality parameter λ, and c_α is the $100(1 - \alpha)$ percentile of the central t-distribution with $m^C + m^T - 2 - q_2$ degrees of freedom. The power of a level α two-tailed test for the treatment effect is therefore

$$p_2 = 1 - \mathrm{f}(c_{\alpha/2}, m^C + m^T - 2 - q_2 , \lambda) + \mathrm{f}(-c_{\alpha/2}, m^C + m^T - 2 - q_2 , \lambda).$$