Chapter 356

Multi-Arm Non-Inferiority Tests for the Difference of Treatment and Control Proportions in a Cluster-Randomized Design

Introduction

This module computes power and sample size for multi-arm non-inferiority tests of the difference between treatment and control proportions when the data are gathered from a cluster-randomized design. The formulas are based on results in Donner and Klar (2000) and Machin, Campbell, Tan, and Tan (2018).

A *cluster (group) randomized design* is one in which whole units, or clusters, of subjects are randomized to the groups rather than the individual subjects in those clusters. The conclusions of the study concern individual subjects rather than the clusters. Examples of clusters are families, school classes, neighborhoods, hospitals, and doctor's practices.

Cluster-randomized designs are often adopted when there is a high risk of contamination if cluster members were randomized individually. For example, it may be difficult for doctors to use two treatment methods in their practice. The price of randomizing by clusters is a loss of efficiency--the number of subjects needed to obtain a certain level of precision in a cluster-randomized trial is usually much larger than the number needed when the subjects are randomized individually. Hence, standard methods of sample size estimation cannot be used.

In this multi-arm design, there are *G* treatment groups and one control group. A proportion is measured in each group. A total of *G* hypothesis tests are anticipated, each comparing a treatment group with the common control group using a z-test of the difference between two proportions.

The Bonferroni adjustment of the type I error rate may be optionally made because several comparisons are being tested using the same data. Making a multiplicity adjustment is usually recommended, but not always. In fact, Saville (1990) advocates not applying it and Machin, Campbell, Tan, and Tan (2018) include omitting it as a possibility.

Example

Suppose that the current treatment for a disease works 70% of the time. Unfortunately, this treatment is expensive and occasionally exhibits serious side-effects. Two promising new treatments have been developed and are now ready to be tested. Two non-inferiority hypotheses need to be tested in this study: whether each new treatment is as good as the current treatment. Hence, at least three groups are needed to complete this study of the two new treatments.

Because of the many benefits of the new treatment, clinicians are willing to adopt a new treatment even if it is slightly less effective than the current treatment. They must determine, however, how much less effective the new treatment can be and still be adopted. Should it be adopted if 69% respond? 68%? 65%? 60%? There is a percentage below 70% at which the difference between the two treatments is no longer considered ignorable. After thoughtful discussion with several clinicians, it was decided that if a response of at least 63% is achieved, the new treatment will be adopted. The difference between these two percentages is called the *margin of non-inferiority*. The margin of non-inferiority in this example is 7%.

The developers must design an experiment to test the hypothesis that the response rate of the new treatment is at least 0.63. The statistical hypotheses to be tested are

$$H_0: P_A \leq P_C + \delta_0 \quad \text{vs.} \quad H_1: P_A > P_C + \delta_0$$

$$H_0: P_B \leq P_C + \delta_0 \quad \text{vs.} \quad H_1: P_B > P_C + \delta_0$$

where $\delta_0 = -0.07$.

Notice that when the null hypothesis is rejected, the conclusion is that the response rate of the treatment group is at least 0.63. Note that even though the response rate of the current treatment is 0.70, the hypothesis test is about a response rate of 0.63. Also notice that a rejection of the null hypothesis results in the conclusion of interest.

Multiple Treatments Versus a Single Control

Whether you want to test several doses of a single treatment or several types of treatments, good research practice requires that each treatment be compared with a control. For example, a popular three-arm design consists of three groups: control, treatment A, and treatment B. Two tests are run: treatment A versus control and treatment B versus the same control. This design avoids having to obtain a second control group for treatment B. Besides the obvious efficiency in subjects, it may be easier to recruit subjects if their chances of receiving the new treatment are better than 50-50.

Technical Details

Our formulation for cluster-randomized designs comes from Donner and Klar (2000). Suppose you have *G* treatment groups with response probabilities P_i that have samples of size N_i and one control group with response probability P_C that has a sample of size N_C. The total sample size is $N = N_1 + N_2 + \ldots + N_G + N_C$.

The statistical hypotheses to be tested are

$$H_0: P_A \leq P_C + \delta_0 \quad \text{vs.} \quad H_1: P_A > P_C + \delta_0$$

$$H_0: P_B \leq P_C + \delta_0 \quad \text{vs.} \quad H_1: P_B > P_C + \delta_0$$

where δ_0 is the non-inferiority margin.

Denote a binary (0, 1) observation by Y_{ikj} where *i* is the group, *k* = 1, 2, ..., K_i is a cluster within group *i*, and *j* = 1, 2, ..., M_i is an item (often a subject) in cluster *k* of group *i.* The results that follow assume an equal number of items per cluster per group. When the number of items from cluster to cluster are about the same, the power and sample size values should be fairly accurate. In these cases, the average number of items per cluster can be used.

The statistical hypothesis that is tested concerns the difference between a treatment group proportion and the control group proportion: P_i and P_C. With a simple modification, the large-sample sample size formulas that are listed in the module for testing two proportions can be used here.

When the items are randomly assigned to one of the $G + 1$ groups, the variance of the sample proportion is

$$\sigma_{S,i}^2 = \frac{P_i(1 - P_i)}{N_i}$$

When the randomization is by clusters of items, the variance of the sample proportion is

$$\sigma_{C,i}^2 = \frac{P_i(1 - P_i)(1 + (M_i - 1)\rho)}{K_i M_i}$$

$$= \sigma_{S,i}^2[1 + (M_i - 1)\rho]$$

$$= F_{i,\rho}\sigma_{S,i}^2$$

The factor $[1 + (M_i - 1)\rho]$ is called the *inflation factor*. The Greek letter ρ is used to represent the *intracluster correlation coefficient (ICC)*. This correlation may be thought of as the simple correlation between any two subjects within the same cluster. If we stipulate that ρ is positive, it may also be interpreted as the proportion of total variability that is attributable to differences between clusters. This value is critical to the sample size calculation.

The asymptotic formulas that were used in comparing two proportions (see Chapter 200, "Tests for Two Proportions") may be used with cluster-randomized designs as well, as long as an adjustment is made for the inflation factor. The basic form of the z-test becomes

$$z = \frac{\left|\widehat{D} - \delta_0\right|}{\hat{\sigma}_{\widehat{D}}(\delta_0)}$$

where

$$\widehat{D} = \hat{p}_i - \hat{p}_C$$

$$\delta_0 = p_i - p_C|H_0$$

$$\hat{\sigma}_{\widehat{D}}(\delta_0) = \sqrt{\frac{\tilde{p}_i(1 - \tilde{p}_i)F_{i,\rho}}{N_i} + \frac{\tilde{p}_C(1 - \tilde{p}_C)F_{C,\rho}}{N_C}}$$

The quantities $\tilde{p}_i$ and $\tilde{p}_C$ are the maximum likelihood estimates constrained by $\tilde{p}_i - \tilde{p}_C = \delta_0$.

Test Statistics

Three test statistics are available in this routine. These are

Z-Test (or Chi-Square Test) (Pooled and Unpooled)

This test statistic was first proposed by Karl Pearson in 1900. Although this test is usually expressed directly as a Chi-Square statistic, it is expressed here as a *z*-statistic so that it can be more easily used for one-sided hypothesis testing.

Both *pooled* and *unpooled* versions of this test have been discussed in the statistical literature. The pooling refers to the way in which the standard error is estimated. In the pooled version, the two proportions are averaged, and only one proportion is used to estimate the standard error. In the unpooled version, the two proportions are used separately.

The formula for the test statistic is

$$z_i = \frac{\hat{p}_i - \hat{p}_C}{\hat{\sigma}_D}$$

Pooled Version

$$\hat{\sigma}_D = \sqrt{\hat{p}(1-\hat{p})\left(\frac{F_{i,\rho}}{N_i} + \frac{F_{c,\rho}}{N_C}\right)}$$

$$\hat{p} = \frac{N_i\hat{p}_i + N_C\hat{p}_C}{N_i + N_C}$$

Unpooled Version

$$\hat{\sigma}_D = \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)F_{i,\rho}}{N_i} + \frac{\hat{p}_C(1-\hat{p}_C)F_{c,\rho}}{N_C}}$$

Power

The following normal approximation to the binomial is used as presented in Chow et al. (2008).

1. Find the critical value (or values in the case of a two-sided test) using the standard normal distribution. The critical value is that value of *z* that leaves exactly the target value of alpha in the tail.
2. Use the normal approximation to binomial distribution to compute binomial probabilities, compute the power for the pooled and unpooled tests, respectively, using

$$\textbf{Pooled:}\ 1-\beta = \Pr\left(Z < \frac{z_\alpha \sigma_{D,p} + (p_i - p_C)}{\sigma_{D,u}}\right) \qquad \textbf{Unpooled:}\ 1-\beta = \Pr\left(Z < \frac{z_\alpha \sigma_{D,u} + (p_i - p_C)}{\sigma_{D,u}}\right)$$

where

$$\sigma_{D,u} = \sqrt{\frac{p_i q_i}{N_i/F_{i,\rho}} + \frac{p_C q_C}{N_C F_{c,\rho}}} \qquad \text{(unpooled standard error)}$$

$$\sigma_{D,p} = \sqrt{\bar{p}\bar{q}\left(\frac{F_{i,\rho}}{N_i} + \frac{F_{c,\rho}}{N_C}\right)} \qquad \text{(pooled standard error)}$$

with $\bar{p} = \dfrac{N_i p_i + N_C p_C}{N_i + N_C}$ and $\bar{q} = 1 - \bar{p}$

Farrington and Manning's Likelihood Score Test

Farrington and Manning (1990) proposed a test statistic for testing whether the difference is equal to a specified value δ_0. The regular MLE's, $\hat{p}_i$ and $\hat{p}_C$, are used in the numerator of the score statistic while MLE's $\tilde{p}_i$ and $\tilde{p}_C$, constrained so that $\tilde{p}_i - \tilde{p}_C = \delta_0$, are used in the denominator. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{FMD} = \frac{\hat{p}_i - \hat{p}_C - \delta_0}{\sqrt{\left(\dfrac{\tilde{p}_i \tilde{q}_i}{N_i/F_{i,\rho}} + \dfrac{\tilde{p}_C \tilde{q}_C}{N_C/F_{C,\rho}}\right)}}$$

where the estimates $\tilde{p}_i$ and $\tilde{p}_C$ are computed as in the corresponding test of Miettinen and Nurminen (1985) given above.

Multiplicity Adjustment

Because *G* z-tests between treatment groups and the control group are run when analyzing the results of this study, many statisticians recommend that a Bonferroni adjustment be applied. This adjustment is easy to apply: the value of alpha that is used in the test is found by dividing the original alpha by the number of tests. For example, if the original alpha is set at 0.05 and the number of treatment (not including the control) groups is five, the individual tests should be conducted using an alpha of 0.01.

The main criticism of this procedure is that if there are many tests, the value of alpha becomes very small. To mitigate against this complaint, some statisticians recommend separating the treatment groups into those that are of primary interest and those that are of secondary interest. The Bonferroni adjustment is made by the using the number of primary treatments rather than the total number of treatments.

There are some who advocate ignoring the adjustment entirely in the case of randomized clinical trials. See for example Saville (1990) and the discussion in chapter 14 of Machin, Campbell, Tan, and Tan (2018).

Size of the Control Group

Because the control group is used over and over, some advocate increasing the number of clusters in this group. The standard adjustment is to include $\sqrt{G}$ clusters in the control group for each cluster in one of the treatment groups. See Machin, Campbell, Tan, and Tan (2018, pages 231-232). Note that often, the treatment groups all have the same sample size.

Example 1 – Finding the Sample Size

A cluster-randomized, multi-arm trial is being designed to compare two treatments against the standard drug in patients with a specific type of disease. They plan to use the Farrington and Manning likelihood score test to analyze the data.

Historically, the standard treatment has enjoyed a 70% cure rate. The new treatments both reduce the seriousness of certain side effects of the standard treatment. Thus, the new treatments will be adopted even if they are slightly less effective than the standard treatment. The researchers will recommend adoption of the either of the new treatments that exhibit a cure rate of at least 60%. That is, the margin of inferiority is -10%. They want a study that can detect a cure rate of 65%,

The researchers will recruit patients from various hospitals. All patients at a particular hospital will receive the same treatment. They anticipate an average of 20 patients per hospital. They want to see the impact on cluster count of having cluster sizes ranging for 10 to 30.

The investigators would like a sample size large enough to find statistical significance at the 0.025 level and the power is 0.80 in each test. Based on similar studies, they estimate the intracluster correlation to be 0.01.

Since the control group will be used twice, they set the control group allocation ratio to $\sqrt{G} = \sqrt{2} = 1.414$ since the control group is used for two comparisons in this design. The two treatment allocation ratios are set to 1.0.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For	**Sample Size**
Higher Proportions Are	**Better (H1: δ > δ0)**
Test Type	**Likelihood Score (Farr. & Mann.)**
Power of Each Test	**0.90**
Overall Alpha	**0.025**
Bonferroni Adjustment	**Standard Bonferroni**
Group Allocation	**Enter Group Allocation Pattern, solve for group numbers of clusters**
M (Items Per Cluster)	**10 20 30**
δ0 (Non-Inferiority Difference)	**-0.1**
Control Proportion	**0.7**
Control Items Per Cluster	**M**
Control Cluster Allocation	**1.414**
Set A Number of Groups	**2**
Set A Proportion	**0.65**
Set A Items Per Cluster	**M**
Set A Cluster Allocation	**1**
Set B Number of Groups	**0**

```
Set C Number of Groups ...............................0
Set D Number of Groups ...............................0
More..............................................................Unchecked
ρ (Intracluster Correlation).............................0.01
```

Output

Click the Calculate button to perform the calculations and generate the following output.

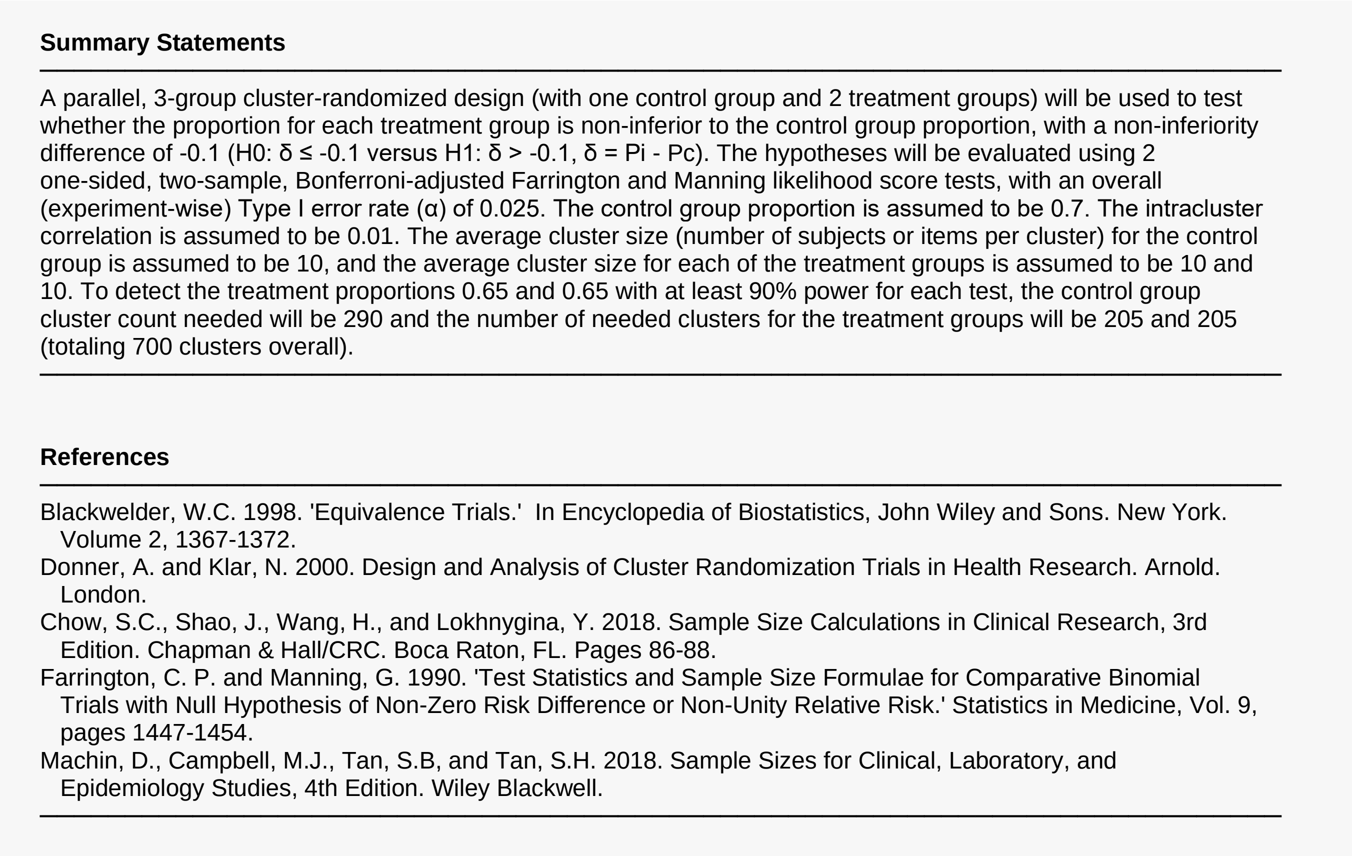
Numeric Reports

Numeric Results

```
Solve For:                Sample Size
Group Allocation:         Enter Group Allocation Pattern, solve for group numbers of clusters
Test Type:                Farrington and Manning Likelihood Score Test
Hypotheses:               H0: δ ≤ δ0   vs.   H1: δ > δ0
Number of Groups:         3
Bonferroni Adjustment:    Standard Bonferroni (Divisor = 2)
```

	Power							Difference			Alpha	
Comparison	Target	Actual	Number of Clusters Ki	Cluster Allocation	Items Per Cluster Mi	Sample Size Ni	Proportion Pi	Non-Inferiority δ0	Actual δi	ICC ρ	Overall	Bonferroni-Adjusted
Control			290	1.414	10	2900	0.70	-0.1		0.01		
vs A1	0.9	0.90091	205	1.000	10	2050	0.65	-0.1	-0.05	0.01	0.025	0.0125
vs A2	0.9	0.90091	205	1.000	10	2050	0.65	-0.1	-0.05	0.01	0.025	0.0125
Total			700			7000						
Control			158	1.414	20	3160	0.70	-0.1		0.01		
vs A1	0.9	0.90086	112	1.000	20	2240	0.65	-0.1	-0.05	0.01	0.025	0.0125
vs A2	0.9	0.90086	112	1.000	20	2240	0.65	-0.1	-0.05	0.01	0.025	0.0125
Total			382			7640						
Control			115	1.414	30	3450	0.70	-0.1		0.01		
vs A1	0.9	0.90181	81	1.000	30	2430	0.65	-0.1	-0.05	0.01	0.025	0.0125
vs A2	0.9	0.90181	81	1.000	30	2430	0.65	-0.1	-0.05	0.01	0.025	0.0125
Total			277			8310						

Comparison The group that is involved in the comparison between the treatment and control displayed on this report line. The comparison is made using the difference.
Target Power The power desired. Power is probability of rejecting a false null hypothesis for this comparison. This power is of the comparison shown on this line only.
Actual Power The power actually achieved.
Ki The number of clusters in the ith group. The total number of clusters is reported in the last row of the column.
Allocation The cluster allocation ratio of the ith group. The value on each row represents the relative number of clusters assigned to the group.
Mi The average number of items per cluster (or average cluster size) in the ith group.
Ni The number of items in the ith group. The total sample size is shown as the last row of the column.
Pi The response proportion in the ith group at which the power is calculated.
δ0 The non-inferiority difference. It is often called the non-inferiority margin (NIM). It is usually set during the planning phase and does not depend on the study data.
δi The difference between the ith group proportion (Pi) and the control group proportion (Pc) at which the power is calculated. The formula is δi = Pi - Pc.
ρ The intracluster correlation (ICC). The correlation between subjects within a cluster.
Overall Alpha The probability of rejecting at least one of the comparisons in this experiment when each null hypothesis is true.
Bonferroni Alpha The adjusted significance level at which each individual comparison is made.



Summary Statements

A parallel, 3-group cluster-randomized design (with one control group and 2 treatment groups) will be used to test whether the proportion for each treatment group is non-inferior to the control group proportion, with a non-inferiority difference of -0.1 (H0: δ ≤ -0.1 versus H1: δ > -0.1, δ = Pi - Pc). The hypotheses will be evaluated using 2 one-sided, two-sample, Bonferroni-adjusted Farrington and Manning likelihood score tests, with an overall (experiment-wise) Type I error rate (α) of 0.025. The control group proportion is assumed to be 0.7. The intracluster correlation is assumed to be 0.01. The average cluster size (number of subjects or items per cluster) for the control group is assumed to be 10, and the average cluster size for each of the treatment groups is assumed to be 10 and 10. To detect the treatment proportions 0.65 and 0.65 with at least 90% power for each test, the control group cluster count needed will be 290 and the number of needed clusters for the treatment groups will be 205 and 205 (totaling 700 clusters overall).

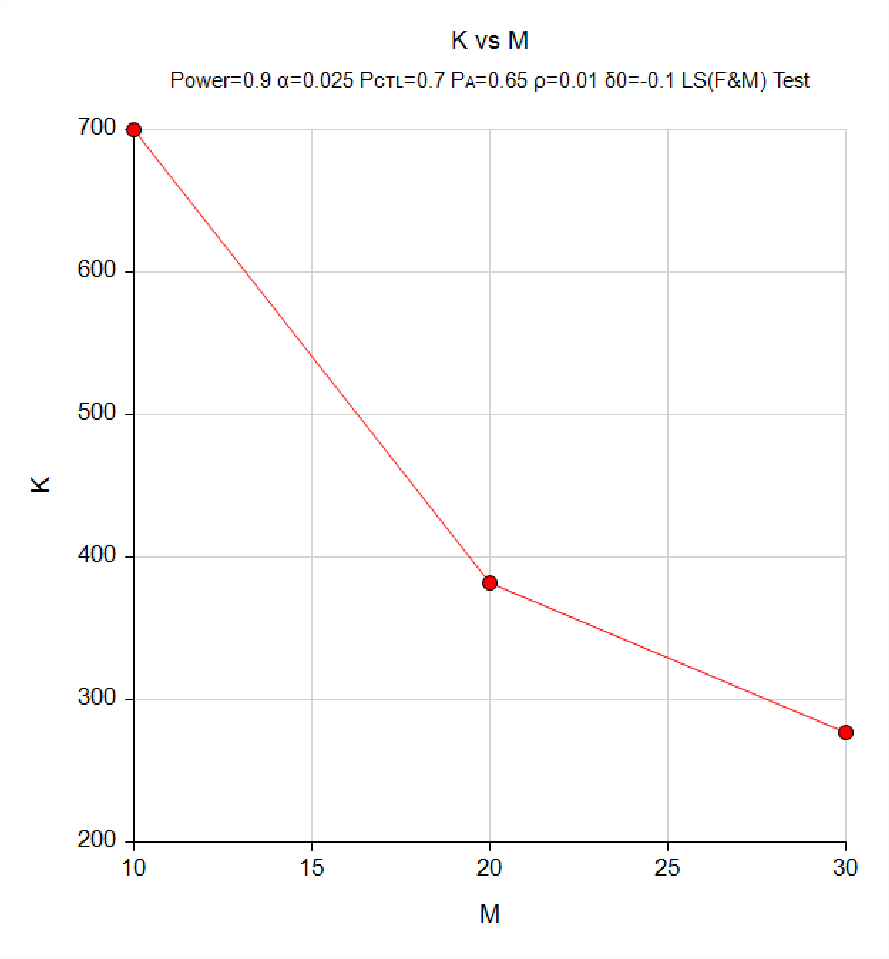
References

Blackwelder, W.C. 1998. 'Equivalence Trials.' In Encyclopedia of Biostatistics, John Wiley and Sons. New York. Volume 2, 1367-1372.
Donner, A. and Klar, N. 2000. Design and Analysis of Cluster Randomization Trials in Health Research. Arnold. London.
Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. 2018. Sample Size Calculations in Clinical Research, 3rd Edition. Chapman & Hall/CRC. Boca Raton, FL. Pages 86-88.
Farrington, C. P. and Manning, G. 1990. 'Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-Zero Risk Difference or Non-Unity Relative Risk.' Statistics in Medicine, Vol. 9, pages 1447-1454.
Machin, D., Campbell, M.J., Tan, S.B, and Tan, S.H. 2018. Sample Sizes for Clinical, Laboratory, and Epidemiology Studies, 4th Edition. Wiley Blackwell.

This report shows the numeric results of this power study. Notice that the results are shown in blocks of three rows at a time. Each block represents a single design.

Plots Section

Plots



This plot gives a visual presentation to the results in the Numeric Report. We can quickly see the impact on the total cluster count, K, of increasing the cluster size, M.

Example 2 – Validation using a Previously Validated Procedure

We could not find a validation result in the statistical literature, so we will use a previously validated **PASS** procedure (**Non-Inferiority Tests for the Difference of Two Proportions in a Cluster-Randomized Design**) to produce the results for the following example.

We will use the settings from Example 1 specifically when M = 30. In this case, the number of clusters allocated to the control group was 115 and to the treatment groups was 81 in each. The Bonferroni adjustment changes the significance level from 0.025 to 0.0125.

The **Non-Inferiority Tests for the Difference of Two Proportions in a Cluster-Randomized Design** procedure is set up as follows.

Design Tab

Solve For .... **Power**
Higher Proportions Are .... **Better (H1: P1 - P2 > D0)**
Test Type... **Likelihood Score (Farr. & Mann.)**
Alpha... **0.0125**
K1 (Clusters in Group 1) **81**
M1 (Average Cluster Size)............................. **30**
K2 (Clusters in Group 2) **115**
M2 (Average Cluster Size)............................. **30**
Input Type... **Proportions**
P1.0 (Non-Inferiority Proportion) **0.6**
P1.1 (Actual Proportion)................................. **0.65**
P2 (Group 2 Proportion)................................. **0.7**
ICC (Intracluster Correlation)......................... **0.01**

This set of options generates the following report.

Numeric Results

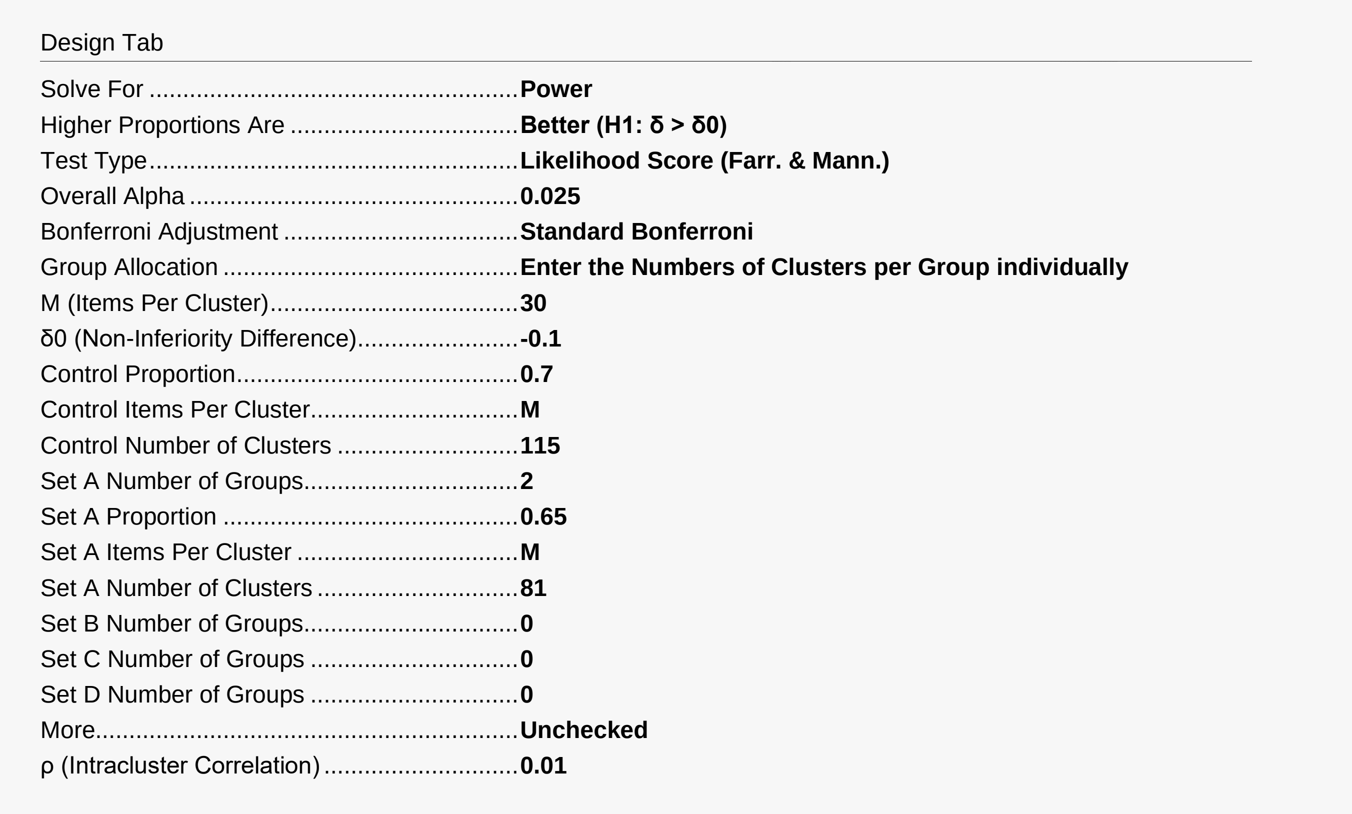
Solve For: Power
Groups: 1 = Treatment, 2 = Reference
Test Statistic: Likelihood Score Test (Farrington & Manning)
Hypotheses: H0: P1 - P2 ≤ D0 vs. H1: P1 - P2 > D0

	Number of Clusters			Cluster Size		Total Sample Size	Proportions			Difference		Intracluster Correlation	
							Non-Inferiority	Actual	Reference	Non-Inferiority	Actual		
Power	K1	K2	K	M1	M2	N	P1.0	P1.1	P2	D0	D1	ICC	Alpha
0.90181	81	115	196	30	30	5880	0.6	0.65	0.7	-0.1	-0.05	0.01	0.0125

The power is computed to be 0.90181.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

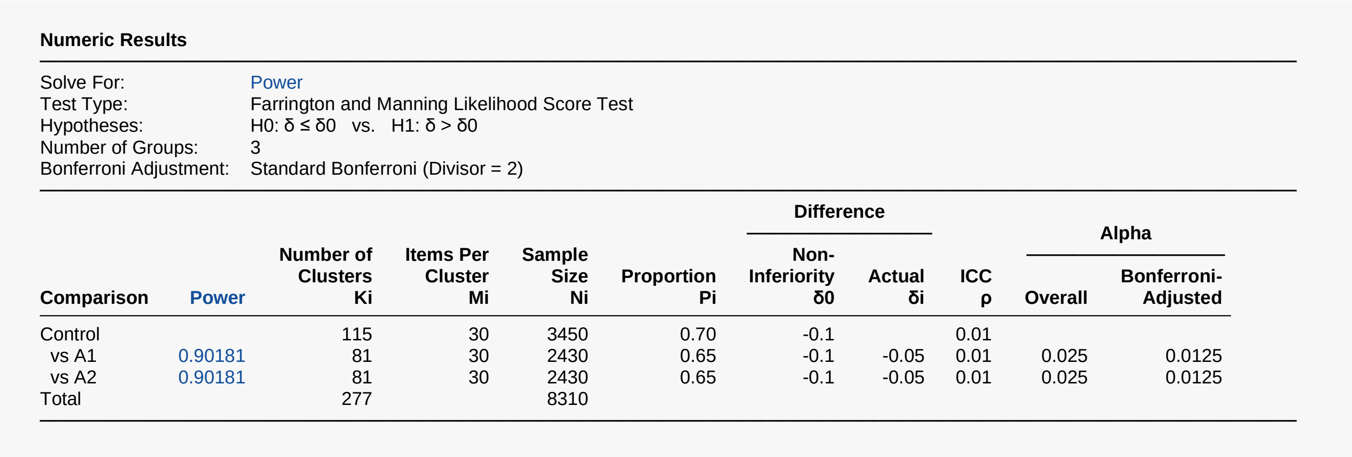


Design Tab

Solve For .... **Power**
Higher Proportions Are .... **Better (H1: δ > δ0)**
Test Type.. **Likelihood Score (Farr. & Mann.)**
Overall Alpha **0.025**
Bonferroni Adjustment .... **Standard Bonferroni**
Group Allocation .... **Enter the Numbers of Clusters per Group individually**
M (Items Per Cluster).................................... **30**
δ0 (Non-Inferiority Difference)....................... **-0.1**
Control Proportion.. **0.7**
Control Items Per Cluster............................... **M**
Control Number of Clusters **115**
Set A Number of Groups................................ **2**
Set A Proportion **0.65**
Set A Items Per Cluster .... **M**
Set A Number of Clusters **81**
Set B Number of Groups................................ **0**
Set C Number of Groups **0**
Set D Number of Groups **0**
More... **Unchecked**
ρ (Intracluster Correlation) **0.01**

Output

Click the Calculate button to perform the calculations and generate the following output.



```
Numeric Results
──────────────────────────────────────────────────────────────────────────────────────────────────────
Solve For:                Power
Test Type:                Farrington and Manning Likelihood Score Test
Hypotheses:               H0: δ ≤ δ0   vs.   H1: δ > δ0
Number of Groups:         3
Bonferroni Adjustment:    Standard Bonferroni (Divisor = 2)
──────────────────────────────────────────────────────────────────────────────────────────────────────
                                                                        Difference
                                                                   ────────────────                  Alpha
                             Number of  Items Per   Sample                  Non-                   ─────────────────────
                              Clusters    Cluster     Size  Proportion  Inferiority  Actual   ICC              Bonferroni-
Comparison     Power               Ki         Mi       Ni          Pi           δ0      δi      ρ    Overall      Adjusted
──────────────────────────────────────────────────────────────────────────────────────────────────────
Control                           115         30     3450        0.70         -0.1            0.01
  vs A1        0.90181             81         30     2430        0.65         -0.1   -0.05   0.01      0.025        0.0125
  vs A2        0.90181             81         30     2430        0.65         -0.1   -0.05   0.01      0.025        0.0125
Total                             277                8310
──────────────────────────────────────────────────────────────────────────────────────────────────────
```

As you can see, the power is 0.90181 for both treatment groups which match the power found in the validation run above. The procedure is validated.