Chapter 602

Multi-Arm Superiority by a Margin Tests for Treatment and Control Means in a Cluster-Randomized Design

Introduction

This module computes power and sample size for multiple superiority by a margin tests for treatment means versus a control mean when the data are gathered from a cluster-randomized design. We could not find any published results about superiority testing with cluster-randomized designs. What we could find were Schuirmann's TOST procedure and a discussion of how to adjust the t-test sample size results given by Campbell and Walters (2014). So, we applied the Campbell and Walters adjustment to Schuirmann's test.

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 mean 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 t-test of the difference between two means.

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.

Background

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 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.

© NCSS, LLC. All Rights Reserved.

Technical Details

Our formulation cluster-randomized designs comes from Campbell and Walters (2014) and Ahn, Heo, and Zhang (2015). Suppose you have *G* treatment groups with means μ_i that have samples of size N_i and one control group with response probability μ_C that has a sample of size N_C. The total sample size is $N = N_1 + N_2 + \ldots + N_G + N_C$.

Superiority by a Margin Test Hypotheses

A *superiority by a margin test* tests that the treatment mean is not worse than the control mean by more than the superiority margin (*SM*). The actual direction of the hypothesis depends on the response variable being studied.

In the following sections, define $\delta_i = \mu_i - \mu_C$.

Case 1: High Values Better

In this case, higher response values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is at least a small amount (*SM*) above the control mean. The null and alternative hypotheses with are

$H_{0i}: \mu_i - \mu_C \leq SM$ vs. $H_{1i}: \mu_i - \mu_C > SM$

$H_{0i}: \mu_i \leq \mu_C + SM$ vs. $H_{1i}: \mu_i > \mu_C + SM$

$H_{0i}: \delta_i \leq SM$ vs. $H_{1i}: \delta_i > SM$

where $SM < 0$.

Case 2: High Values Worse

In this case, lower values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is at least a small amount (*SM*) below the control mean. The null and alternative hypotheses with are

$H_{0i}: \mu_i - \mu_C \geq SM$ vs. $H_{1i}: \mu_i - \mu_C < SM$

$H_{0i}: \mu_i \geq \mu_C + SM$ vs. $H_{1i}: \mu_i < \mu_C + SM$

$H_{0i}: \delta_i \geq SM$ vs. $H_{1i}: \delta_i < SM$

where $SM > 0$.

© NCSS, LLC. All Rights Reserved.

Power Calculations

Denote a continuous 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_{ik} is an item (subject) in cluster *k* of group *i.*

We let σ^2 denote the variance of Y_{ikj}, which is $\sigma^2_{Between} + \sigma^2_{Within}$, where $\sigma^2_{Between}$ is the variation between clusters and σ^2_{Within} is the variation within clusters. Also, let ρ denote the intracluster correlation coefficient (ICC) which is $\sigma^2_{Between}/\left(\sigma^2_{Between} + \sigma^2_{Within}\right)$. This correlation is the simple correlation between any two observations in the same cluster.

For sample size calculation, we assume that the m_{ik} are distributed with a mean cluster size of M_i and a coefficient of variation of cluster sizes of *COV*. The variances of the group means, $\bar{Y}_i$, are approximated by

$$V_i = \frac{\sigma^2 (DE_i)(RE_i)}{K_i M_i}$$

where

$$DE_i = 1 + (M_i - 1)\rho$$

$$RE_i = \frac{1}{1 - (COV)^2 \lambda_i (1 - \lambda_i)}$$

$$\lambda_i = M_i \rho / (M_i \rho + 1 - \rho)$$

DE is called the *Design Effect* and RE is the *Relative Efficiency* of unequal to equal cluster sizes. Both are greater than or equal to one, so both inflate the variance.

Assume that $\delta_i = \mu_i - \mu_c - SM$ is to be tested using a modified two-sample t-test. Assuming that higher values are better, the superiority by a margin test statistic is

$$t = \frac{\bar{Y}_i - \bar{Y}_C - SM}{\sqrt{\hat{V}_i + \hat{V}_C}}$$

has an approximate *t* distribution with degrees of freedom $DF = K_i M_i + K_C M_C - 2$ for a *subject-level* analysis or K_i + K_C – 2 for a *cluster-level* analysis.

Let the noncentrality parameter $\Delta_i = (\delta_i - SM)/\sigma_d$, where $\sigma_d = \sqrt{V_i + V_C}$. We can define the two critical values based on a central t-distribution with DF degrees of freedom as follows.

$$X_1 = t_{\frac{\alpha}{2}, DF}$$

$$X_2 = t_{1-\frac{\alpha}{2}, DF}$$

© NCSS, LLC. All Rights Reserved.

The power can be found from the following to probabilities

$$P_1 = H_{X_1,DF,\Delta_\mathrm{i}}$$

$$P_2 = H_{X_2,DF,\Delta_\mathrm{i}}$$

$$\text{Power } = 1 - (P_2 - P_1)$$

where $H_{X,DF,\Delta}$ is the cumulative probability distribution of the noncentral-t distribution.

The power of a one-sided test can be calculated similarly.

Multiplicity Adjustment

Because *G* t-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.

© NCSS, LLC. All Rights Reserved.

Example 1 – Finding the Sample Size

Suppose that a four-arm, cluster-randomized, superiority by a margin study is to be conducted in which $\mu_1 = \mu_2 = \mu_3 = 4.2$, $\mu_C = 3.2$, SM = 0.032 (10% of the control mean), σ = 3.7, ρ = 0.01, *Mi* = 5, 10, or 15, *COV* = 0.65, *alpha* = 0.025, and number of clusters is to be calculated. Higher means are better. The power is 0.9 calculated for a one-sided, subject-based, superiority test.

The control group multiplier will be set to $\sqrt{G} = \sqrt{3} = 1.732$ since the control group is used for three comparisons in this design.

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 Means Are	**Better (H1: δ > SM)**
Test Statistic	**T-Test Based on Number of Subjects**
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 (Average Cluster Size)	**5 10 15**
COV of Cluster Sizes	**0.65**
SM (Superiority Margin)	**0.32**
Control Mean	**3.2**
Control Items Per Cluster	**M**
Control Cluster Allocation	**1.732**
Set A Number of Groups	**3**
Set A Mean	**4.2**
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**
σ (Standard Deviation)	**3.7**
ρ (Intracluster Correlation)	**0.01**

© NCSS, LLC. All Rights Reserved.

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: T-Test with DF based on number of subjects
Higher Means Are: Better
Hypotheses: H0: δ ≤ SM vs. H1: δ > SM
Number of Groups: 4
Bonferroni Adjustment: Standard Bonferroni (Divisor = 3)

	Power		Number of Clusters		Cluster Size		Sample			Superiority	Standard		Alpha	
Comparison	Target	Actual	Ki	Cluster Allocation	Average Mi	COV	Size Ni	Mean μi	Difference δi	Margin SM	Deviation σ	ICC ρ	Overall	Bonferroni-Adjusted
Control			232	1.732	5	0.65	1160	3.2			3.7	0.01		
vs A1	0.9	0.90028	134	1.000	5	0.65	670	4.2	1	0.32	3.7	0.01	0.025	0.00833
vs A2	0.9	0.90028	134	1.000	5	0.65	670	4.2	1	0.32	3.7	0.01	0.025	0.00833
vs A3	0.9	0.90028	134	1.000	5	0.65	670	4.2	1	0.32	3.7	0.01	0.025	0.00833
Total			634				3170							
Control			125	1.732	10	0.65	1250	3.2			3.7	0.01		
vs A1	0.9	0.90339	72	1.000	10	0.65	720	4.2	1	0.32	3.7	0.01	0.025	0.00833
vs A2	0.9	0.90339	72	1.000	10	0.65	720	4.2	1	0.32	3.7	0.01	0.025	0.00833
vs A3	0.9	0.90339	72	1.000	10	0.65	720	4.2	1	0.32	3.7	0.01	0.025	0.00833
Total			341				3410							
Control			88	1.732	15	0.65	1320	3.2			3.7	0.01		
vs A1	0.9	0.90336	51	1.000	15	0.65	765	4.2	1	0.32	3.7	0.01	0.025	0.00833
vs A2	0.9	0.90336	51	1.000	15	0.65	765	4.2	1	0.32	3.7	0.01	0.025	0.00833
vs A3	0.9	0.90336	51	1.000	15	0.65	765	4.2	1	0.32	3.7	0.01	0.025	0.00833
Total			241				3615							

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.
COV The coefficient of variation of the cluster sizes within the group.
Ni The number of items in the ith group. The total sample size is shown as the last row of the column.
μi The mean of the ith group at which the power is computed. The first row contains μc, the control group mean.
δi The difference between the ith treatment mean and the control mean (μi - μc) at which the power is computed.
SM The margin of superiority in the scale of the mean difference. SM > 0.
σ The standard deviation of the responses within each group.
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.

© NCSS, LLC. All Rights Reserved.

Summary Statements

A parallel, 4-group cluster-randomized design (with one control group and 3 treatment groups) will be used to test whether the mean for each treatment group is superior to the control group mean, with a superiority margin of 0.32 (H0: δ ≤ 0.32 versus H1: δ > 0.32, δ = μi - μc). The superiority-by-a-margin hypotheses will be evaluated using 3 one-sided, two-sample, Bonferroni-adjusted (divisor = 3) t-tests with degrees of freedom based on the number of subjects, with an overall (experiment-wise) Type I error rate (α) of 0.025. The common subject-to-subject standard deviation for all groups is assumed to be 3.7. The coefficient of variation of the cluster size in all clusters is assumed to be 0.65. The control group mean is assumed to be 3.2. 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 5, and the average cluster size for each of the treatment groups is assumed to be 5, 5, and 5. To detect the treatment means 4.2, 4.2, and 4.2 with at least 90% power for each test, the control group cluster count needed will be 232 and the number of needed clusters for the treatment groups will be 134, 134, and 134 (totaling 634 clusters overall).

References

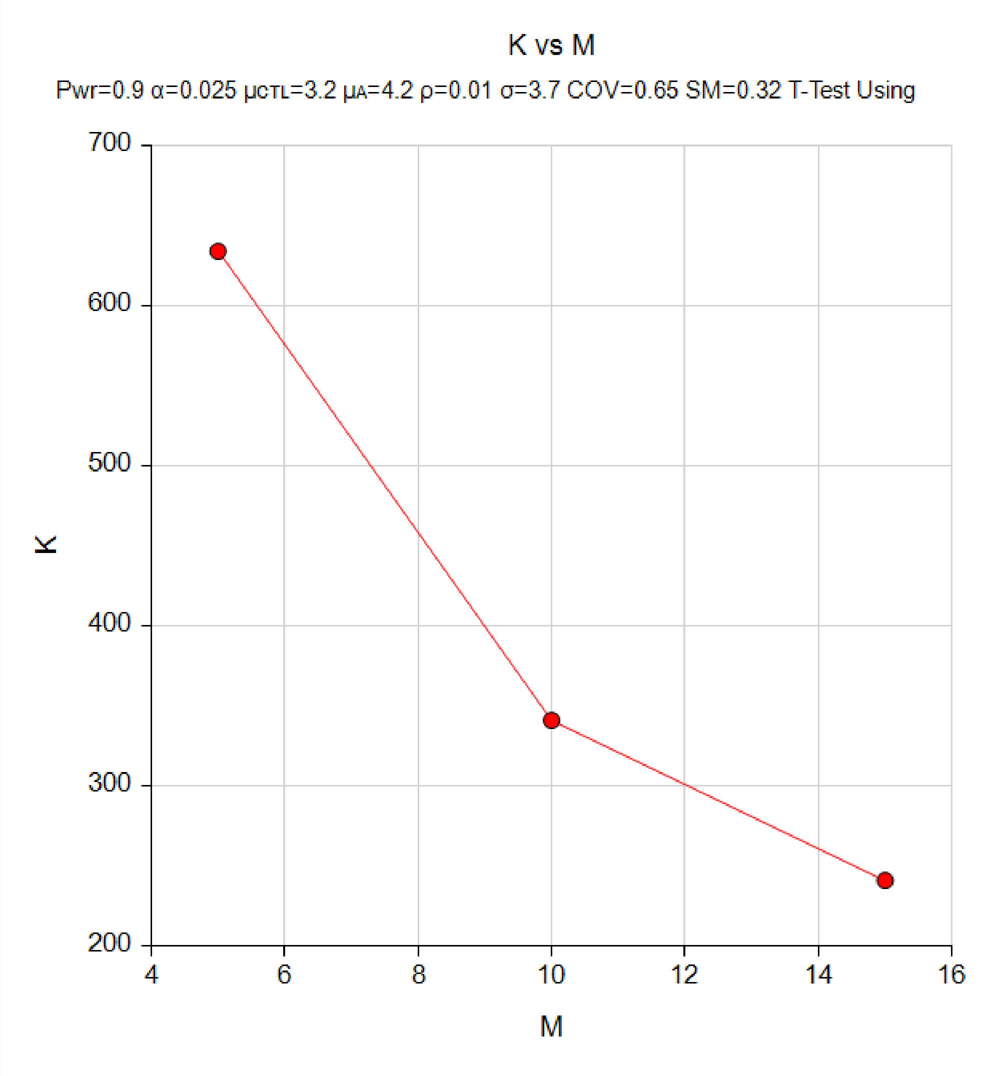
Ahn, C., Heo, M., and Zhang, S. 2015. Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research. CRC Press. New York.
Blackwelder, W.C. 1998. 'Equivalence Trials.' In Encyclopedia of Biostatistics, John Wiley and Sons. New York. Volume 2, 1367-1372.
Campbell, M.J. and Walters, S.J. 2014. How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research. Wiley. New York.
Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. 2018. Sample Size Calculations in Clinical Research, Third Edition. Taylor & Francis/CRC. Boca Raton, Florida.
Donner, A. and Klar, N. 1996. 'Statistical Considerations in the Design and Analysis of Community Intervention Trials'. J. Clin. Epidemiol. Vol 49, No. 4, pages 435-439.
Donner, A. and Klar, N. 2000. Design and Analysis of Cluster Randomization Trials in Health Research. Arnold. London.
Julious, Steven A. 2010. Sample Sizes for Clinical Trials. CRC Press. New York.
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 sample size study. Notice that the results are shown in blocks of five rows at a time. Each block represents an individual treatment.

© NCSS, LLC. All Rights Reserved.

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.

© NCSS, LLC. All Rights Reserved.

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 (**Superiority by a Margin Tests for Two Means in a Cluster-Randomized Design**) to produce the results for the following example.

Suppose that a four-arm, cluster-randomized study is to be conducted in which $\mu_1 = \mu_2 = \mu_3 = 4.2$, $\mu_C = 3.2$, SM = 0.32, σ = 3.7, ρ = 0.01, *Ki* = 91, *Mi* = 10, *COV* = 0.65, and *alpha* = 0.025 / 3 = 0.00833333. The calculated power is 0.90209 for a subject-based test. All groups will have the same number of clusters.

The **Superiority by a Margin Tests for Two Means in a Cluster-Randomized Design** procedure is set up as follows.

Design Tab

Solve For .... **Power**
Higher Means Are .... **Better (H1: δ > SM)**
Test Statistic .... **T-Test Based on Number of Subjects**
Alpha.. **0.008333333**
K1 (Number of Clusters) **91**
M1 (Average Cluster Size)............................. **10**
K2 (Number of Clusters) **K1**
M2 (Average Cluster Size)............................. **M1**
COV of Cluster Sizes.................................... **0.65**
SM (Superiority Margin)................................. **0.32**
δ (Mean Difference = μ1 - μ2)........................ **1**
σ (Standard Deviation).................................. **3.7**
ρ (Intracluster Correlation, ICC)..................... **0.01**

This set of options generates the following report.

Numeric Results for a Test of Mean Difference

Solve For: Power
Groups: 1 = Treatment, 2 = Control
Test Statistic: T-Test with DF based on number of subjects
Higher Means Are: Better
Hypotheses: H0: δ ≤ SM vs. H1: δ > SM

	Number of Clusters			Cluster Size			Sample Size							
Power	K1	K2	K	M1	M2	COV	N1	N2	N	Mean Difference δ	Superiority Margin SM	Standard Deviation σ	ICC ρ	Alpha
0.90209	91	91	182	10	10	0.65	910	910	1820	1	0.32	3.7	0.01	0.00833

The power is computed to be 0.90209.

© NCSS, LLC. All Rights Reserved.

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 Means Are	**Better (H1: δ > SM)**
Test Statistic	**T-Test Based on Number of Subjects**
Overall Alpha	**0.025**
Bonferroni Adjustment	**Standard Bonferroni**
Group Allocation	**Equal (Kc = K1 = K2 = ...)**
Ki (Group Number of Clusters)	**91**
M (Average Cluster Size)	**10**
COV of Cluster Sizes	**0.65**
SM (Superiority Margin)	**0.32**
Control Mean	**3.2**
Control Items Per Cluster	**M**
Set A Number of Groups	**3**
Set A Mean	**4.2**
Set A Items Per Cluster	**M**
Set B Number of Groups	**0**
Set C Number of Groups	**0**
Set D Number of Groups	**0**
More	**Unchecked**
σ (Standard Deviation)	**3.7**
ρ (Intracluster Correlation)	**0.01**

© NCSS, LLC. All Rights Reserved.

Output

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

Numeric Results

Solve For:	Power
Test Type:	T-Test with DF based on number of subjects
Higher Means Are:	Better
Hypotheses:	H0: δ ≤ SM vs. H1: δ > SM
Number of Groups:	4
Bonferroni Adjustment:	Standard Bonferroni (Divisor = 3)

			Cluster Size								Alpha	
Comparison	Power	Number of Clusters Ki	Average Mi	COV	Sample Size Ni	Mean μi	Difference δi	Superiority Margin SM	Standard Deviation σ	ICC ρ	Overall	Bonferroni-Adjusted
Control		91	10	0.65	910	3.2			3.7	0.01		
vs A1	0.90209	91	10	0.65	910	4.2	1	0.32	3.7	0.01	0.025	0.00833
vs A2	0.90209	91	10	0.65	910	4.2	1	0.32	3.7	0.01	0.025	0.00833
vs A3	0.90209	91	10	0.65	910	4.2	1	0.32	3.7	0.01	0.025	0.00833
Total		364			3640							

As you can see, the power is 0.90209 for all treatment groups which matches the power found in the validation run above. The procedure is validated.

© NCSS, LLC. All Rights Reserved.