Chapter 604

Multi-Arm Equivalence Tests for Treatment and Control Means in a Cluster-Randomized Design

Introduction

This module computes power and sample size for multiple equivalence tests of treatment means versus a control mean when the data are obtained from a cluster-randomized design. We could not find any published results about equivalence 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.

Technical Details

Our formulation of 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 $N = N_1 + N_2 + \ldots + N_G + N_C$.

Equivalence Tests

Measurements are made on individuals that have been randomly assigned to the groups. This *parallel-groups* design may be analyzed by a set of TOST equivalence tests to show that the means of the treatment and control groups do not differ by more than a small amount, either positive or negative. To conduct an equivalence test, you must set upper and lower equivalence limits for the difference between a treatment mean and the control mean. These limits, which will be called *EL* and EU, establish an interval of equivalence. When the sample mean difference falls between these limits, the null hypothesis of non-equivalence is rejected and the equivalence of the two group means is concluded.

The statistical hypotheses are written as follows:

$$H_{0i}: \mu_i - \mu_C \le EL \quad \text{or} \quad \mu_i - \mu_C \ge EU \quad \text{vs.} \quad H_{1i}: EL < \mu_i - \mu_C < EU$$

or, if we define $\delta_i = \mu_i - \mu_C$,

$$H_{0i}: \delta_i \le EL \quad \text{or} \quad \delta_i \ge EU \quad \text{vs.} \quad H_{1i}: \delta_i < EU$$

where $EL < 0$ and $EU > 0$. Usually, $EL = -EU$.

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*. 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 $\delta_i = \mu_i - \mu_c - NIM$ is to be tested using two modified two-sample t-tests. The test statistics are

$$t_L = \frac{\bar{Y}_i - \bar{Y}_C - EL}{\sqrt{\hat{V}_i + \hat{V}_C}}$$

and

$$t_U = \frac{\bar{Y}_i - \bar{Y}_C - EU}{\sqrt{\hat{V}_i + \hat{V}_C}}$$

We assume that these statistics have 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.

Define the noncentrality parameters as $\Delta_{Li} = (\delta_i - EL)/\sigma_{di}$ and $\Delta_{Ui} = (\delta_i - EU)/\sigma_{di}$ where $\sigma_{di} = \sqrt{V_i + V_C}$.

The power of this test procedure is given by

$$\text{Power} = \Pr\left(T_L \geq t_{1-\alpha,DF} \text{ and } T_U \leq -t_{1-\alpha,DF}\right)$$

where T_L and T_U are distributed as the bivariate, noncentral *t* distribution with noncentrality parameters Δ_L and Δ_U.

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.

Example 1 – Finding the Sample Size

Suppose that a four-arm, cluster-randomized, equivalence study is to be conducted in which $\mu_1 = \mu_2 = \mu_3 = 5$, $\mu_C = 5$, EL = -1, EU = 1, σ = 3.7, ρ = 0.01, *Mi* = 5, 10, or 15, *COV* = 0.65, *alpha* = 0.05, and the number of clusters is to be calculated. The required power value is 0.9 calculated for a subject-based, equivalence 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**
Test Statistic	**T-Test Based on Number of Subjects**
Power of Each Test	**0.90**
Overall Alpha	**0.05**
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**
EU (Upper Equivalence Limit)	**1**
EL (Lower Equivalence Limit)	**-Upper Limit**
Control Mean	**5**
Control Items Per Cluster	**M**
Control Cluster Allocation	**1.732**
Set A Number of Groups	**3**
Set A Mean	**5**
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**

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
Hypotheses:	H0: δ ≤ EL or δ ≥ EU vs. H1: EL < δ < EU
Number of Groups:	4
Bonferroni Adjustment:	Standard Bonferroni (Divisor = 3)

			Cluster Size					Equivalence Limits				Alpha	
Comparison	Power	Number of Clusters Ki	Cluster Allocation	Average Mi	COV	Sample Size Ni	Mean μi	Difference δi	Lower EL	Upper EU	Standard Deviation σ	ICC ρ	Overall	Bonferroni-Adjusted
Control		114	1.732	5	0.65	570	5				3.7	0.01		
vs A1	0.90401	66	1.000	5	0.65	330	5	0	-1	1	3.7	0.01	0.05	0.01667
vs A2	0.90401	66	1.000	5	0.65	330	5	0	-1	1	3.7	0.01	0.05	0.01667
vs A3	0.90401	66	1.000	5	0.65	330	5	0	-1	1	3.7	0.01	0.05	0.01667
Total		312				1560								
Control		61	1.732	10	0.65	610	5				3.7	0.01		
vs A1	0.90359	35	1.000	10	0.65	350	5	0	-1	1	3.7	0.01	0.05	0.01667
vs A2	0.90359	35	1.000	10	0.65	350	5	0	-1	1	3.7	0.01	0.05	0.01667
vs A3	0.90359	35	1.000	10	0.65	350	5	0	-1	1	3.7	0.01	0.05	0.01667
Total		166				1660								
Control		43	1.732	15	0.65	645	5				3.7	0.01		
vs A1	0.90574	25	1.000	15	0.65	375	5	0	-1	1	3.7	0.01	0.05	0.01667
vs A2	0.90574	25	1.000	15	0.65	375	5	0	-1	1	3.7	0.01	0.05	0.01667
vs A3	0.90574	25	1.000	15	0.65	375	5	0	-1	1	3.7	0.01	0.05	0.01667
Total		118				1770								

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.
EL	The lower equivalence limit for the difference. This is the smallest negative mean difference between each treatment group and the control group that still results in the conclusion that the treatment group is equivalent to the control group.
EU	The upper equivalence limit for the difference. This is the largest positive difference mean difference between each treatment group and the control group that still results in the conclusion that the treatment group is equivalent to the control group.
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.

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 equivalent to the control group mean, with equivalence difference bounds of -1 and 1 (H0: $\delta \leq -1$ or $\delta \geq 1$ versus H1: $-1 < \delta < 1$, $\delta = \mu i - \mu c$). Each of the 3 equivalence comparisons will be made using two 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 0.05. 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 5. 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 5, 5, and 5 with at least 90% power for each test, the control group cluster count needed will be 114 and the number of needed clusters for the treatment groups will be 66, 66, and 66 (totaling 312 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.

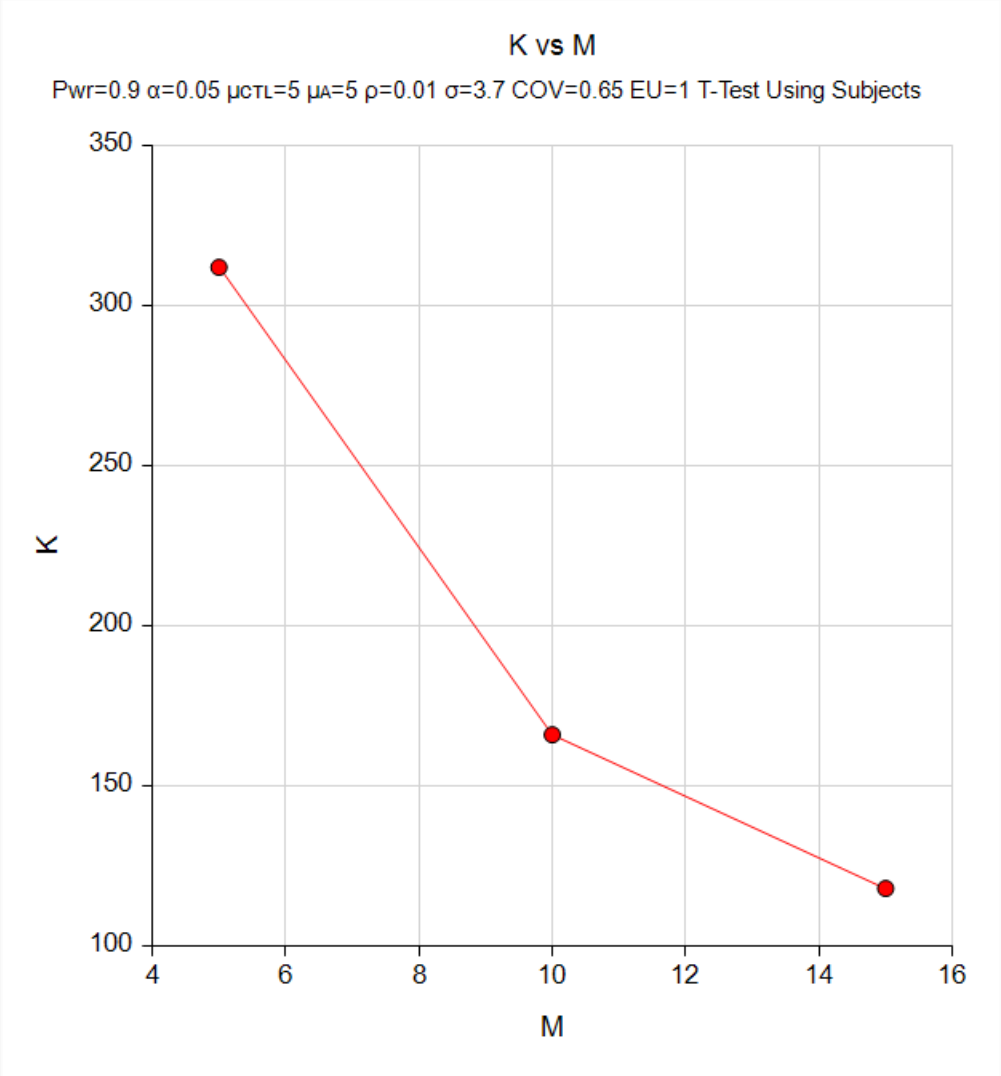
Phillips, Kem F. 1990. 'Power of the Two One-Sided Tests Procedure in Bioequivalence', Journal of Pharmacokinetics and Biopharmaceutics, Volume 18, No. 2, pages 137-144.

Schuirmann, Donald. 1987. 'A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability', Journal of Pharmacokinetics and Biopharmaceutics, Volume 15, Number 6, pages 657-680.

This report shows the numeric results of this sample size study. Notice that the results are shown in blocks of four rows at a time. Each block represents an individual treatment.

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 (**Equivalence 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 $\mu_1 = \mu_2 = \mu_3 = 5$, $\mu_C = 5$, EL = -1, EU = 1, σ = 3.7, ρ = 0.01, *Ki* = 11, *Mi* = 10, *COV* = 0.65, and *alpha* = 0.05 / 3 = 0.016666667. The calculated power is 0.94135 for a subject-based test. All groups will have the same number of clusters.

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

Design Tab

Solve For	**Power**
Test Statistic	**T-Test Based on Number of Subjects**
Alpha	**0.016666667**
K1 (Number of Clusters)	**50**
M1 (Average Cluster Size)	**10**
K2 (Number of Clusters)	**K1**
M2 (Average Cluster Size)	**M1**
COV of Cluster Sizes	**0.65**
EU (Upper Equivalence Limit)	**1**
EL (Lower Equivalence Limit)	**-Upper Limit**
δ (Mean Difference = μ1 - μ2)	**0**
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
Hypotheses: H0: δ ≤ EL or δ ≥ EU vs. H1: EL < δ < EU

	Number of Clusters			Cluster Size			Sample Size		Mean Difference	Equivalence Limits		Standard Deviation	ICC	
Power	K1	K2	K	M1	M2	COV	N1	N2	δ	Lower EL	Upper EU	σ	ρ	Alpha
0.94135	50	50	100	10	10	0.65	500	500	0	-1	1	3.7	0.01	0.01667

The power is computed to be 0.94135.

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**
Test Statistic	**T-Test Based on Number of Subjects**
Overall Alpha	**0.05**
Bonferroni Adjustment	**Standard Bonferroni**
Group Allocation	**Equal (Kc = K1 = K2 = ...)**
Ki (Group Number of Clusters)	**50**
M (Average Cluster Size)	**10**
COV of Cluster Sizes	**0.65**
EU (Upper Equivalence Limit)	**1**
EL (Lower Equivalence Limit)	**-Upper Limit**
Control Mean	**5**
Control Items Per Cluster	**M**
Set A Number of Groups	**3**
Set A Mean	**5**
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**

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
Hypotheses:	H0: δ ≤ EL or δ ≥ EU vs. H1: EL < δ < EU
Number of Groups:	4
Bonferroni Adjustment:	Standard Bonferroni (Divisor = 3)

			Cluster Size					Equivalence Limits				Alpha	
Comparison	Power	Number of Clusters Ki	Average Mi	COV	Sample Size Ni	Mean μi	Difference δi	Lower EL	Upper EU	Standard Deviation σ	ICC ρ	Overall	Bonferroni-Adjusted
Control		50	10	0.65	500	5				3.7	0.01		
vs A1	0.94135	50	10	0.65	500	5	0	-1	1	3.7	0.01	0.05	0.01667
vs A2	0.94135	50	10	0.65	500	5	0	-1	1	3.7	0.01	0.05	0.01667
vs A3	0.94135	50	10	0.65	500	5	0	-1	1	3.7	0.01	0.05	0.01667
Total		200			2000								

As you can see, the power is 0.94135 for all treatment groups which matches the power found in the validation run above. The procedure is validated.