Chapter 483

Non-Inferiority Tests for Two Means in a Cluster-Randomized Design

Introduction

This procedure computes power and sample size for a *non-inferiority* test in cluster-randomized designs in which the outcome is a continuous normal random variable.

Cluster-randomized designs are those in which whole clusters of subjects (classes, hospitals, communities, etc.) are put into the treatment group or the control group. In this case, the means of two groups, made up of K_i clusters of M_{ij} individuals each, are to be tested using a modified z test, or t-test, in which the clusters are treated as subjects. Generally speaking, the larger the cluster sizes and the higher the correlation among subjects within the same cluster, the larger will be the overall sample size necessary to detect an effect with the same power.

It should be noted that we could not find any published results about non-inferiority 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. We look forward to results that substantiate our approach.

The Statistical Hypotheses

Non-inferiority tests are examples of directional (one-sided) tests. This program module provides the input and output in formats that are convenient for these types of tests. This section will review the specifics of non-inferiority testing.

Remember that in the usual t-test setting, the null (H_0) and alternative (H_a) hypotheses for one-sided tests are defined as follows, assuming that $\delta = \mu_1 - \mu_2$ is to be tested.

$$H_0: \delta \leq 0 \text{ versus. } H_a: \delta > 0$$

Rejecting this test implies that the mean difference is larger than the value δ. This test is called an *upper-tailed test* because it is rejected in samples in which the difference between the sample means is larger than *D*.

Following is an example of a *lower-tailed test*.

$$H_0: \delta \geq 0 \text{ versus. } H_a: \delta < 0$$

Non-inferiority tests are special cases of the above directional tests. It will be convenient to adopt the following specialized notation for the discussion of these tests.

Parameter	PASS Input/Output	Interpretation
μ_1	Not used	*Mean* of population 1. Population 1 is assumed to consist of those who have received the new treatment.
μ_2	Not used	*Mean* of population 2. Population 2 is assumed to consist of those who have received the reference treatment.
ε	NIM	*Margin of non-inferiority*. This is a tolerance value that defines the magnitude of the amount that is not of practical importance. This may be thought of as the largest change from the baseline that is considered to be trivial. The absolute value is shown to emphasize that this is a magnitude. The sign of the value will be determined by the specific design that is being used.
δ	δ	*True difference*. This is the value of $\mu_1 - \mu_2$, the difference between the means.

Note that the actual values of μ_1 and μ_2 are not needed. Only their difference is needed for power and sample size calculations.

Non-Inferiority Tests

A *non-inferiority test* tests that the treatment mean is not worse than the reference mean by more than the non-inferiority margin. The actual direction of the hypothesis depends on the response variable being studied.

Case 1: High Values Good, Non-Inferiority Test

In this case, higher values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is no less than a small amount below the reference mean. The value of δ is often set to zero. The following are equivalent sets of hypotheses.

H_0: $\delta \le -\varepsilon$ versus H_a: $\delta > -\varepsilon$, $\varepsilon > 0$

H_0: $\mu_1 - \mu_2 \le -\varepsilon$ versus H_a: $\mu_1 - \mu_2 > -\varepsilon$, $\varepsilon > 0$

H_0: $\mu_1 \le \mu_2 - \varepsilon$ versus H_a: $\mu_1 > \mu_2 - \varepsilon$, $\varepsilon > 0$

Case 2: High Values Bad, Non-Inferiority Test

In this case, lower values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is no more than a small amount above the reference mean. The value of δ is often set to zero. The following are equivalent sets of hypotheses.

H_0: $\delta \ge \varepsilon$ versus H_a: $\delta < \varepsilon$, $\varepsilon > 0$

H_0: $\mu_1 - \mu_2 \ge \varepsilon$ versus H_a: $\mu_1 - \mu_2 < \varepsilon$, $\varepsilon > 0$

H_0: $\mu_1 \ge \mu_2 + \varepsilon$ versus H_a: $\mu_1 < \mu_2 + \varepsilon$, $\varepsilon > 0$

Technical Details

Our formulation is a combination of non-inferiority formulas given by Chow, Shao, Wang, and Lokhnygina (2018) pages 50-51 and the cluster-randomized design formulas given in Campbell and Walters (2014) and Ahn, Heo, and Zhang (2015). Denote an observation by Y_{ijk} where i = 1, 2 gives the group, j = 1,2,…,K_i gives the cluster within group i, and k = 1, 2, …, m_{ij} denotes an individual in cluster j of group i.

We let σ^2 denote the variance of Y_{ijk}, 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 simply correlation between any two observations in the same cluster.

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

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

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

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

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

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 $\delta = \mu_1 - \mu_2$ is to be tested using a modified two-sample t-test. Assuming that higher values are better, the non-inferiority test statistic is

$$t = \frac{\bar{Y}_1 - \bar{Y}_2 - \varepsilon}{\sqrt{\hat{V}_1 + \hat{V}_2}}$$

We assume this statistic has an approximate t distribution with degrees of freedom $DF = K_1M_1 + K_2M_2 - 2$ for a *subject-level* analysis or $K_1 + K_2 - 2$ for a *cluster-level* analysis.

Define the noncentrality parameter as $\Delta = (\delta - \varepsilon)/\sigma_d$, where $\sigma_d = \sqrt{V_1 + V_2}$. We can define the critical value based on a central t-distribution with DF degrees of freedom as follows.

$$X = t_{\alpha,DF}$$

The power can be found from the following to probabilities

$$\text{Power} = 1 - H_{X,DF,\Delta}$$

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

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Design tab. For more information about the options of other tabs, go to the Procedure Window chapter.

Design Tab

The Design tab contains most of the parameters and options that you will be concerned with.

Solve For

Solve For

This option specifies the parameter to be solved for from the other parameters. The parameters that may be selected are *δ*, *Power*, *K1*, and *M1*.

Under most situations, you will select either *Power* to calculate power or *K1* to calculate the number of clusters. Occasionally, you may want to fix the number of clusters and find the necessary cluster size.

Note that the value selected here always appears as the vertical axis on the charts.

The program is set up to calculate power directly. To find appropriate values of the other parameters, a binary search is made using an iterative procedure until an appropriate value is found.

Test

Higher Means Are

This option defines whether higher values of the response variable are to be considered better or worse. The choice here determines the direction of the non-inferiority test.

If Higher Means Are *Better* the null hypothesis is $\delta \leq$ -NIM and the alternative hypothesis is $\delta >$ -NIM. If Higher Means Are *Worse* the null hypothesis is $\delta \geq$ NIM and the alternative hypothesis is $\delta <$ NIM.

Test Statistic

Specify which t-test statistic you are going to use: a t-test based on the number of subjects or a t-test in which the cluster means are treated as subjects.

- **T-Test Based on Number of Subjects**

  This uses the methodology shown in the recent books by Campbell and Walters (2014) and Ahn, Heo, and Zhang (2015). In this case, power is based on a t-test in which the variance is inflated to adjust for the clustering and the degrees of freedom is based on to the number of subjects.

- **T-Test Based on Number of Clusters**

  This uses the original methodology of Donner and Klar (1996). In this case, power is based on a t-test in which the variance is also inflated to adjust for the clustering, but the degrees of freedom are based on the number of clusters. Donner and Klar ignored the impact of COV, so, if you want to match their results, you should set the COV to zero.

Power and Alpha

Power

This option specifies one or more values for power. Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

If your only interest is in determining the appropriate sample size for a confidence interval, set power to 0.5.

Alpha

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected.

Values must be between zero and one. Usually, the value of 0.025 is used for alpha of one-sided test and this has become a standard.

You may enter a range of values such as *0.025 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size – Number of Clusters & Cluster Size

Group 1 (Treatment)

K1 (Number of Clusters)

Enter a value (or range of values) for the number of clusters in the treatment group. You may enter a range of values such as *10 to 20 by 2*. The sample size for this group is equal to the number of clusters times the average cluster size.

M1 (Average Cluster Size)

This is the average number of subjects per cluster in group one. This value must be a positive number that is at least one. You can use a list of values such as *100 150 200*.

Group 2 (Control)

K2 (Number of Clusters)

This is the number of clusters in group two. This value must be a positive number. The sample size for this group is equal to the number of clusters times the number of subjects per cluster.

If you simply want a multiple of the value for group one, you would enter the multiple followed by *K1*, with no blanks. If you want to use *K1* directly, you do not have to premultiply by *1*. For example, all of the following are valid entries:*10 K1 2K1 0.5K1 K1*.

You can use a list of values such as *10 20 30* or *K1 2K1 3K1*.

M2 (Average Cluster Size)

This is the average number of subjects per cluster in group two. This value must be at least one.

If you simply want a multiple of the value for group one, you would enter the multiple followed by *M1*, with no blanks. If you want to use *M1* directly, you do not have to premultiply by *1*. For example, all of the following are valid entries: *10M1 2M1 0.5M1 M1*.

You can use a list of values such as *10 20 30* or *M1 2M1 3M1*.

Coefficient of Variation of Cluster Sizes

COV of Cluster Sizes

Enter the *coefficient of variation* of the cluster sizes (number of subjects). This value must be zero or a positive number. The COV of X is defined as the standard deviation of X divided by the mean of X.

Campbell and Walters (2014) page 71 give guidance on the possible values of COV. They indicate that as the average cluster size increases, COV tends toward 0.65. They say that typical values of COV range from 0.4 to 0.9.

You can use a list of values such as *0.4 0.6 0.8*.

Standard Deviation

The standard deviation, calculated by the sample formula (divide by n-1), is a measure of the variability. When no other information is available, Campbell and Walters (2014) page 71 suggest using (Maximum Cluster Size - Minimum Cluster Size) / 4.

All Cluster Sizes Equal

When all cluster sizes are equal, the coefficient of variation is zero.

Effect Size

NIM (Non-Inferiority Margin)

This is the magnitude of the margin of non-inferiority. It must be a positive number.

When higher means are better, this value is the distance below the reference mean that is still considered non-inferior.

When higher means are worse, this value is the distance above the reference mean that is still considered non-inferior.

When a series of values is entered, PASS will generate a separate calculation result for each value of the series.

δ (Mean Difference = μ1 – μ2)

This is the actual difference between the treatment mean and the reference mean at which power is calculated. For non-inferiority tests, this value is often set to zero.

When higher means are better, $\delta > -\text{NIM}$. When higher means are worse, $\delta < \text{NIM}$.

σ (Standard Deviation)

Enter the subject-to-subject standard deviation. This standard deviation applies for both groups.

Note that σ must be a positive number. You can enter a single value such as *5* or a series of values such as *1 3 5 7 9* or *1 to 9 by 2*.

Press the small 'σ' button to the right to obtain calculation options for estimating the standard deviation.

ρ (Intracluster Correlation, ICC)

This is the value of the intracluster correlation coefficient. It may be interpreted as the correlation between any two observations in the same cluster. It may also be thought of as the proportion of the variation in response that can be accounted for by the between-cluster variation.

Possible values are from 0 to just below 1. Typical values are between 0.0001 and 0.05.

You may enter a single value or a list of values.

Example 1 – Calculating Power

Suppose that a non-inferiority test is to be conducted on data obtained from a cluster-randomized design in which *NIM*= 1; δ = 0; σ = 4; ρ = 0.0, 0.01, and 0.10; *M1* and *M2* = 10; *COV* = 0.65; *alpha* = 0.025; and *K1* and *K2* = 10, 20, or 40. Power is to be calculated assuming higher means are better.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority Tests for Two Means in a Cluster-Randomized Design** procedure window by expanding **Means**, then **Two Independent Means**, then clicking on **Cluster-Randomized**, and then clicking on **Non-Inferiority Tests for Two Means in a Cluster-Randomized Design**. You may then make the appropriate entries as listed below, or open **Example 1** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Higher Means Are	**Better (Ha: δ > -NIM)**
Test Statistic	**T-Test Based on Number of Subjects**
Alpha	**0.025**
K1 (Number of Clusters)	**10 20 40**
M1 (Average Cluster Size)	**10**
K2 (Number of Clusters)	**K1**
M2 (Average Cluster Size)	**M1**
COV of Cluster Sizes	**0.65**
NIM (Non-Inferiority Margin)	**1**
δ (Mean Difference = μ1 - μ2)	**0**
σ (Standard Deviation)	**4**
ρ (Intracluster Correlation, ICC)	**0 0.01 0.1**

Annotated Output

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

Numeric Results

Numeric Results for a Test of Mean Difference
Test Statistic: T-Test with DF based on number of subjects
Higher Means are Better
Hypotheses: H0: δ ≤ -NIM vs. H1: δ > -NIM

Power	Subj Cnt Gr 1 N1	Subj Cnt Gr 2 N2	Clus Cnt Gr 1 K1	Clus Cnt Gr 2 K2	Clus Size Gr 1 M1	Clus Size Gr 2 M2	COV Clus Sizes COV	Diff μ1-μ2 δ	N.I. Margin -NIM	Std Dev σ	ICC ρ	Alpha
0.4204	100	100	10	10	10	10	0.650	0.00	-1.00	4.00	0.000	0.025
0.7033	200	200	20	20	10	10	0.650	0.00	-1.00	4.00	0.000	0.025
0.9423	400	400	40	40	10	10	0.650	0.00	-1.00	4.00	0.000	0.025
0.3802	100	100	10	10	10	10	0.650	0.00	-1.00	4.00	0.010	0.025
0.6504	200	200	20	20	10	10	0.650	0.00	-1.00	4.00	0.010	0.025
0.9139	400	400	40	40	10	10	0.650	0.00	-1.00	4.00	0.010	0.025
0.2258	100	100	10	10	10	10	0.650	0.00	-1.00	4.00	0.100	0.025
0.4018	200	200	20	20	10	10	0.650	0.00	-1.00	4.00	0.100	0.025
0.6795	400	400	40	40	10	10	0.650	0.00	-1.00	4.00	0.100	0.025

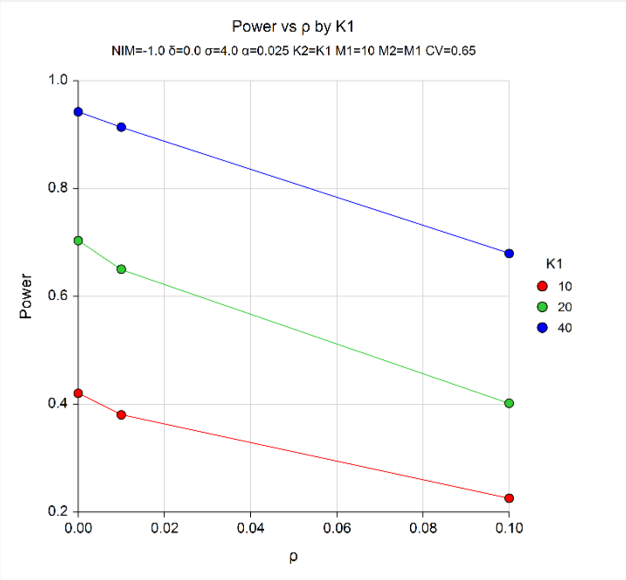
References
Campbell, M.J. and Walters, S.J. 2014. How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research. Wiley. New York.
Julious, Steven A. 2010. Sample Sizes for Clinical Trials. CRC Press. New York.
Ahn, C., Heo, M., and Zhang, S. 2015. Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research. CRC Press. New York.
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.
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.

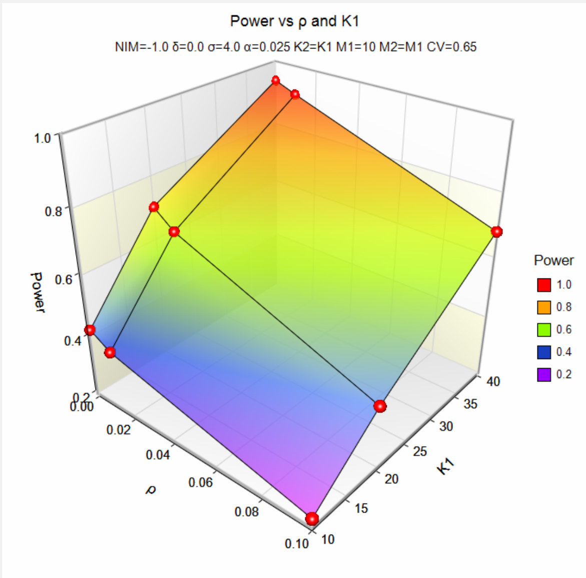
Report Definitions
Power is the probability of rejecting a false null hypothesis. It should be close to one.
N1 and N2 are the number of subjects in groups 1 and 2, respectively.
K1 and K2 are the number of clusters in groups 1 and 2, respectively.
M1 and M2 are the average number of items (subjects) per cluster in groups 1 and 2, respectively.
NIM is the margin of non-inferiority. Since higher means are better, this value is negative and is the distance below the group two (reference) mean that is still considered non-inferior.
COV is the coefficient of variation of the cluster sizes.
δ is the mean difference (μ1 - μ2) in the response at which the power is calculated.
σ is the standard deviation of the subject responses.
ρ (ICC) is the intracluster correlation.
Alpha is the probability of rejecting a true null hypothesis, that is, rejecting when the means are actually equal.

Summary Statements
Sample sizes of 100 in group one and 100 in group two, which were obtained by sampling 10 clusters with an average of 10 subjects each in group one and 10 clusters with an average of 10 subjects each in group two, achieve 42% power to detect non-inferiority. The margin of non-inferiority is -1.00. The true difference between the means is assumed to be 0.00. The standard deviation of subjects is 4.00. The intracluster correlation coefficient is 0.000. The coefficient of variation of cluster sizes is 0.650. A one-sided, two-sample t-test was used with a significance level of 0.025. This test used degrees of freedom based on the number of subjects.

This report shows the power for each of the scenarios.

Plots Section





These plots show the results of the various scenarios specified.

Example 2 – Validation using Chow, Shao, Wang, and Lokhnygina (2018)

We could not find a validation example for this test, so we will use a validation in which M1 = M2 = 1. Chow, Shao, Wang, and Lokhnygina (2018) page 53 has an example of a sample size calculation for a non-inferiority trial. Their example obtains a sample size of 51 in each group when $\delta = 0$, NIM = 0.05, $\sigma = 0.1$, alpha = 0.05, and power = 0.80. Because M1 = 1, the values of ρ and COV are ignored. They obtain a value of 51 for K1 and K2.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority Tests for Two Means in a Cluster-Randomized Design** procedure window by expanding **Means**, then **Two Independent Means**, then clicking on **Cluster-Randomized**, and then clicking on **Non-Inferiority Tests for Two Means in a Cluster-Randomized Design**. You may then make the appropriate entries as listed below, or open **Example 2** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**K1 (Number of Clusters)**
Higher Means Are	**Better (Ha: δ > -NIM)**
Test Statistic	**T-Test Based on Number of Subjects**
Power	**0.8**
Alpha	**0.05**
M1 (Average Cluster Size)	**1**
K2 (Number of Clusters)	**K1**
M2 (Average Cluster Size)	**M1**
COV of Cluster Sizes	**0**
NIM (Non-Inferiority Margin)	**0.05**
δ (Mean Difference = μ1 - μ2)	**0**
σ (Standard Deviation)	**0.1**
ρ (Intracluster Correlation, ICC)	**0**

Output

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

Numeric Results

Numeric Results for a Test of Mean Difference
Test Statistic: T-Test with DF based on number of subjects
Higher Means are Better
Hypotheses: H0: δ ≤ -NIM vs. H1: δ > -NIM

Power	Subj Cnt Gr 1 N1	Subj Cnt Gr 2 N2	Clus Cnt Gr 1 K1	Clus Cnt Gr 2 K2	Clus Size Gr 1 M1	Clus Size Gr 2 M2	COV Clus Sizes COV	Diff μ1-μ2 δ	N.I. Margin -NIM	Std Dev σ	ICC ρ	Alpha
0.8059	51	51	51	51	1	1	0.000	0.00	-0.05	0.10	0.000	0.050

PASS calculates the same sample sizes as Chow, Shao, Wang, and Lokhnygina (2018).