

## 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.

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

Rejecting this test implies that the mean difference is larger than the value  $\delta$ . 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*.

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

*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.

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

<b>Parameter</b>	<b>PASS Input/Output</b>	<b>Interpretation</b>
$\mu_1$	Not used	<i>Mean</i> of population 1. Population 1 is assumed to consist of those who have received the new treatment.
$\mu_2$	Not used	<i>Mean</i> of population 2. Population 2 is assumed to consist of those who have received the reference treatment.
$\varepsilon$	NIM	<i>Margin of non-inferiority</i> . 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.
$\delta$	$\delta$	<i>True difference</i> . This is the value of $\mu_1 - \mu_2$ , the difference between the means.

Note that the actual values of  $\mu_1$  and  $\mu_2$  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  $\delta$  is often set to zero. The following are equivalent sets of hypotheses.

$$\begin{aligned}
 H_0: \delta \leq -\varepsilon & \quad \text{versus} \quad H_a: \delta > -\varepsilon, \quad \varepsilon > 0 \\
 H_0: \mu_1 - \mu_2 \leq -\varepsilon & \quad \text{versus} \quad H_a: \mu_1 - \mu_2 > -\varepsilon, \quad \varepsilon > 0 \\
 H_0: \mu_1 \leq \mu_2 - \varepsilon & \quad \text{versus} \quad H_a: \mu_1 > \mu_2 - \varepsilon, \quad \varepsilon > 0
 \end{aligned}$$

### 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  $\delta$  is often set to zero. The following are equivalent sets of hypotheses.

$$\begin{aligned}
 H_0: \delta \geq \varepsilon & \quad \text{versus} \quad H_a: \delta < \varepsilon, \quad \varepsilon > 0 \\
 H_0: \mu_1 - \mu_2 \geq \varepsilon & \quad \text{versus} \quad H_a: \mu_1 - \mu_2 < \varepsilon, \quad \varepsilon > 0 \\
 H_0: \mu_1 \geq \mu_2 + \varepsilon & \quad \text{versus} \quad H_a: \mu_1 < \mu_2 + \varepsilon, \quad \varepsilon > 0
 \end{aligned}$$

## Technical Details

Our formulation is a combination of non-inferiority formulas given by Chow, Shao, Wang, and Likhnygina (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,  $j = 1, 2, \dots, K_i$  gives the cluster within group  $i$ , and  $k = 1, 2, \dots, m_{ij}$  denotes an individual in cluster  $j$  of group  $i$ .

We let  $\sigma^2$  denote the variance of  $Y_{ijk}$ , which is  $\sigma_{Between}^2 + \sigma_{Within}^2$ , where  $\sigma_{Between}^2$  is the variation between clusters and  $\sigma_{Within}^2$  is the variation within clusters. Also, let  $\rho$  denote the intracluster correlation coefficient (ICC) which is  $\sigma_{Between}^2 / (\sigma_{Between}^2 + \sigma_{Within}^2)$ . This correlation is simply the 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  $DF = K_1 M_1 + K_2 M_2 - 2$  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.

## Example 1 – Calculating Power

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

### 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 .....	<b>Power</b>
Higher Means Are .....	<b>Better (Ha: <math>\delta &gt; -NIM</math>)</b>
Test Statistic .....	<b>T-Test Based on Number of Subjects</b>
Alpha.....	<b>0.025</b>
K1 (Number of Clusters) .....	<b>10 20 40</b>
M1 (Average Cluster Size).....	<b>10</b>
K2 (Number of Clusters) .....	<b>K1</b>
M2 (Average Cluster Size).....	<b>M1</b>
COV of Cluster Sizes .....	<b>0.65</b>
NIM (Non-Inferiority Margin) .....	<b>1</b>
$\delta$ (Mean Difference = $\mu_1 - \mu_2$ ).....	<b>0</b>
$\sigma$ (Standard Deviation).....	<b>4</b>
$\rho$ (Intracluster Correlation, ICC).....	<b>0 0.01 0.1</b>

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

## Output

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

## Numeric Reports

## 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:  $H_0: \delta \leq -NIM$  vs.  $H_1: \delta > -NIM$

Power	Number of Clusters			Cluster Size			Sample Size			Mean Difference $\delta$	Non-Inferiority Margin -NIM	Standard Deviation $\sigma$	ICC $\rho$	Alpha
	K1	K2	K	M1	M2	COV	N1	N2	N					
0.4204	10	10	20	10	10	0.65	100	100	200	0	-1	4	0.00	0.025
0.7033	20	20	40	10	10	0.65	200	200	400	0	-1	4	0.00	0.025
0.9423	40	40	80	10	10	0.65	400	400	800	0	-1	4	0.00	0.025
0.3802	10	10	20	10	10	0.65	100	100	200	0	-1	4	0.01	0.025
0.6504	20	20	40	10	10	0.65	200	200	400	0	-1	4	0.01	0.025
0.9139	40	40	80	10	10	0.65	400	400	800	0	-1	4	0.01	0.025
0.2258	10	10	20	10	10	0.65	100	100	200	0	-1	4	0.10	0.025
0.4018	20	20	40	10	10	0.65	200	200	400	0	-1	4	0.10	0.025
0.6795	40	40	80	10	10	0.65	400	400	800	0	-1	4	0.10	0.025

Power The probability of rejecting a false null hypothesis when the alternative hypothesis is true.  
 K1, K2, and K The number of clusters in groups 1 and 2, and their total.  
 M1 and M2 The average number of items (subjects) per cluster in groups 1 and 2, respectively.  
 COV The coefficient of variation of the cluster sizes.  
 N1, N2, and N The number of subjects in groups 1 and 2, and their total.  
 -NIM The margin of non-inferiority. Since higher means are better, this value is negative and is the distance below the group 2 (control) mean that is still considered non-inferior.  
 $\delta$  The mean difference in the response at which the power is calculated.  $\delta = \mu_1 - \mu_2$ .  
 $\sigma$  The standard deviation of the subject responses.  
 $\rho$  The intraclass correlation (ICC). The correlation between a pair of subjects within a cluster.  
 Alpha The probability of rejecting a true null hypothesis.

## Summary Statements

A parallel, two-group cluster-randomized design (where higher means are considered to be better) will be used to test whether the Group 1 (treatment) mean ( $\mu_1$ ) is non-inferior to the Group 2 (control) mean ( $\mu_2$ ), with a non-inferiority margin of -1 ( $H_0: \delta \leq -1$  versus  $H_1: \delta > -1$ ,  $\delta = \mu_1 - \mu_2$ ). The comparison will be made using a one-sided t-test with the degrees of freedom based on the total number of subjects (see Campbell and Walters, 2014, and Ahn, Heo, and Zhang, 2015), with a Type I error rate ( $\alpha$ ) of 0.025. The common subject-to-subject standard deviation for both groups is assumed to be 4, the intraclass correlation coefficient is assumed to be 0, and the coefficient of variation of cluster sizes is assumed to be 0.65. To detect a mean difference ( $\mu_1 - \mu_2$ ) of 0, with 10 clusters of 10 subjects per cluster in Group 1 (totaling 100 subjects) and 10 clusters of 10 subjects per cluster in Group 2 (totaling 100 subjects), the power is 0.4204.

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

**References**

Ahn, C., Heo, M., and Zhang, S. 2015. Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research. CRC Press. New York.

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, 3rd Edition. Chapman & Hall/CRC. Boca Raton, FL. Pages 86-88.

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.

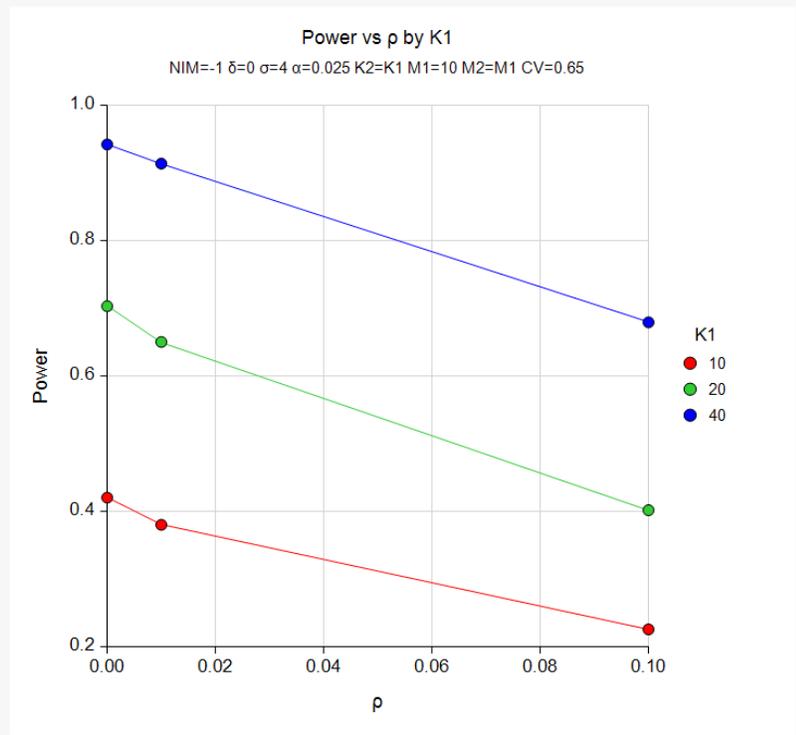
Donner, A. and Klar, N. 2000. Design and Analysis of Cluster Randomization Trials in Health Research. Arnold. London.

Julious, Steven A. 2010. Sample Sizes for Clinical Trials. CRC Press. New York.

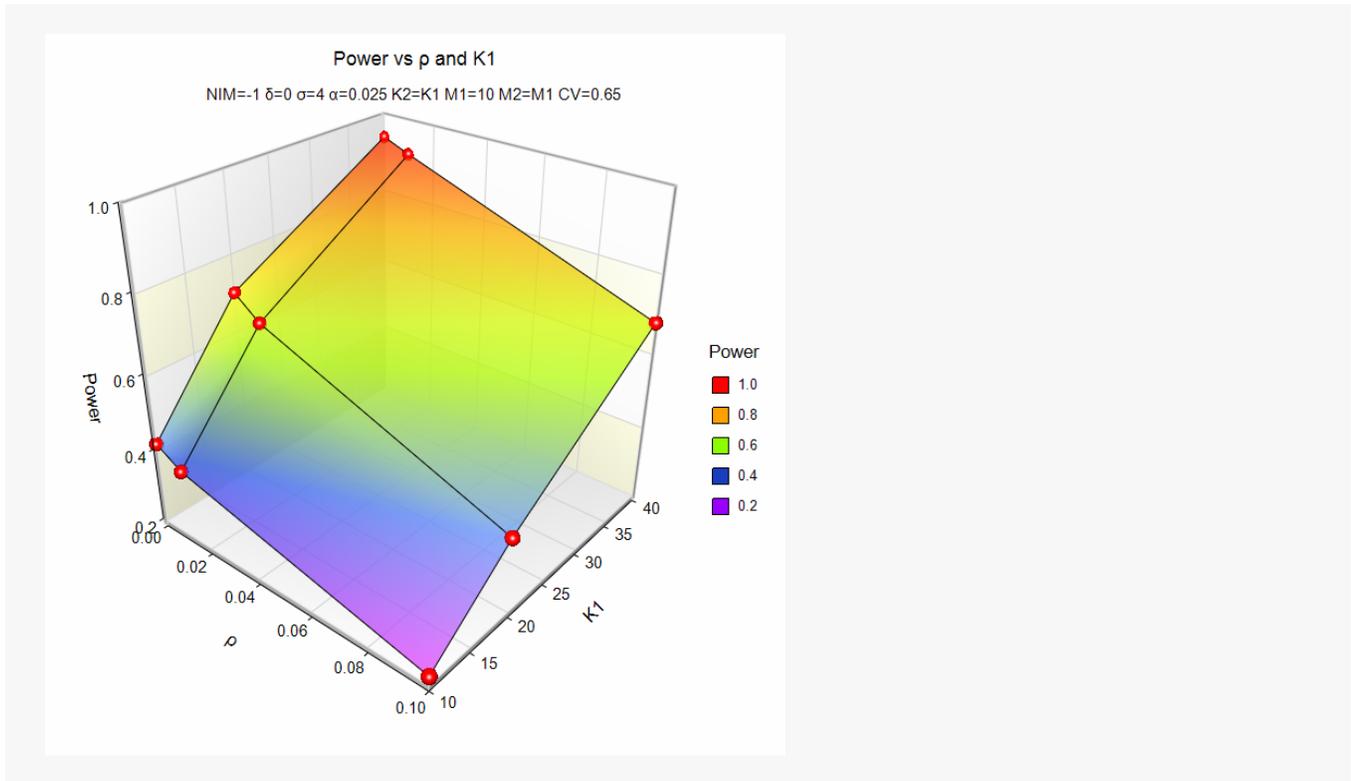
This report shows the power for each of the scenarios.

**Plots Section**

**Plots**



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



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  $\rho$  and COV are ignored. They obtain a value of 51 for  $K1$  and  $K2$ .

### 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 .....	<b>K1 (Number of Clusters)</b>
Higher Means Are .....	<b>Better (Ha: <math>\delta &gt; -NIM</math>)</b>
Test Statistic .....	<b>T-Test Based on Number of Subjects</b>
Power.....	<b>0.8</b>
Alpha.....	<b>0.05</b>
M1 (Average Cluster Size).....	<b>1</b>
K2 (Number of Clusters) .....	<b>K1</b>
M2 (Average Cluster Size).....	<b>M1</b>
COV of Cluster Sizes .....	<b>0</b>
NIM (Non-Inferiority Margin) .....	<b>0.05</b>
$\delta$ (Mean Difference = $\mu_1 - \mu_2$ ).....	<b>0</b>
$\sigma$ (Standard Deviation).....	<b>0.1</b>
$\rho$ (Intraclass Correlation, ICC).....	<b>0</b>

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

## Output

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

### Numeric Results for a Test of Mean Difference

Solve For: [K1 \(Number of Clusters\)](#)  
 Groups: 1 = Treatment, 2 = Control  
 Test Statistic: T-Test with DF Based on Number of Subjects  
 Higher Means Are: Better  
 Hypotheses:  $H_0: \delta \leq -NIM$  vs.  $H_1: \delta > -NIM$

Power	Number of Clusters			Cluster Size			Sample Size			Mean Difference $\delta$	Non-Inferiority Margin -NIM	Standard Deviation $\sigma$	ICC $\rho$	Alpha
	K1	K2	K	M1	M2	COV	N1	N2	N					
0.8059	51	51	102	1	1	0	51	51	102	0	-0.05	0.1	0	0.05

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