Chapter 510

# Non-Inferiority Tests for the Difference Between Two Means in a 2x2 Cross-Over Design

## Introduction

This procedure computes power and sample size for non-inferiority tests in 2x2 cross-over designs in which the outcome is a continuous normal random variable. The details of sample size calculation for the 2x2 cross-over design are presented in the 2x2 Cross-Over Designs chapter and they will not be duplicated here. This chapter only discusses those changes necessary for non-inferiority tests. Sample size formulas for non-inferiority tests of cross-over designs are presented in Chow et al. (2003) pages 63-68.

## Cross-Over Designs

Senn (2002) defines a *cross-over* design as one in which each subject receives all treatments, and the objective is to study differences among the treatments. The name *cross-over* comes from the most common case in which there are only two treatments. In this case, each subject *crosses over* from one treatment to the other. It is assumed that there is a *washout* period between treatments during which the response returns back to its baseline value. If this does not occur, there is said to be a *carry-over* effect.

A 2x2 cross-over design refers to two treatments (periods) and two *sequences* (treatment orderings). One sequence receives treatment A followed by treatment B. The other sequence receives B and then A. The design includes a washout period between responses to make certain that the effects of the first drug do not carry-over to the second. Thus, the groups in this design are defined by the sequence in which the two drugs are administered, not by the treatments they receive.

Cross-over designs are employed because, if the no-carryover assumption is met, treatment differences are measured within a subject rather than between subjects—making a more precise measurement. Examples of the situations that might use a cross-over design are the comparison of anti-inflammatory drugs in arthritis and the comparison of hypotensive agents in essential hypertension. In both of these cases, symptoms are expected to return to their usual baseline level shortly after the treatment is stopped.

## Cross-Over Analysis

The following discussion summarizes the presentation of Chow and Liu (1999). The general linear model for the standard 2x2 cross-over design is

$$Y_{ijk} = \mu + S_{ik} + P_j + \mu_{(j,k)} + C_{(j-1,k)} + e_{ijk}$$

where *i* represents a subject (1 to $N_k$), *j* represents the period (1 or 2), and *k* represents the sequence (1 or 2). The $S_{ik}$ represent the random effects of the subjects. The $P_j$ represent the effects of the two periods. The $\mu_{(j,k)}$ represent the means of the two treatments. In the case of the 2x2 cross-over design

$$\mu_{(j,k)} = \begin{cases} \mu_1 & \text{if } k = j \\ \mu_2 & \text{if } k \neq j \end{cases}$$

where the subscripts 1 and 2 represent treatments A and B, respectively.

The $C_{(j-1,k)}$ represent the carry-over effects. In the case of the 2x2 cross-over design

$$C_{(j-1,k)} = \begin{cases} C_1 & \text{if } j = 2, k = 1 \\ C_2 & \text{if } j = 2, k = 2 \\ 0 & \text{otherwise} \end{cases}$$

where the subscripts 1 and 2 represent treatments A and B, respectively.

Assuming that the average effect of the subjects is zero, the four means from the 2x2 cross-over design can be summarized using the following table.

| **Sequence** | **Period 1** | **Period 2** |
|---|---|---|
| 1 (AB) | $\mu_{11} = \mu + P_1 + \mu_1$ | $\mu_{21} = \mu + P_2 + \mu_2 + C_1$ |
| 2 (BA) | $\mu_{12} = \mu + P_1 + \mu_2$ | $\mu_{22} = \mu + P_2 + \mu_1 + C_2$ |

where $P_1 + P_2 = 0$ and $C_1 + C_2 = 0$.

# The Statistical Hypotheses

Both non-inferiority and superiority tests are examples of directional (one-sided) tests and their power and sample size can be calculated using the 2x2 Cross-Over Design procedure. However, at the urging of our users, we have developed this module which provides the input and output in formats that are convenient for these types of tests. This section reviews the specifics of non-inferiority and superiority testing.

Remember that in the usual t-test setting, the null ($H_0$) and alternative ($H_1$) hypotheses for one-sided tests are defined as

$$H_0: \mu_X \leq A \quad \text{versus} \quad H_1: \mu_X > A$$

Rejecting $H_0$ implies that the mean is larger than the value *A*. This test is called an *upper-tailed test* because it is rejected in samples in which the difference in sample means is larger than *A.*

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

$$H_0: \mu_X \geq A \quad \text{versus} \quad H_1: \mu_X < A$$

*Non-inferiority* and *superiority* 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** |
|---|---|---|
| $\mu_T$ | Not used | *Treatment mean*. This is the treatment mean. |
| $\mu_R$ | Not used | *Reference mean*. This is the mean of a reference population. |
| $M_{NI}$ | 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. |
| $\delta$ | δ1 | *True difference*. This is the value of $\mu_T - \mu_R$, the difference between the treatment and reference means. This is the value at which the power is calculated. |

Note that the actual values of $\mu_T$ and $\mu_R$ 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 equivalence 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.

$$H_0: \mu_1 \leq \mu_2 - |M_{NI}| \quad \text{versus} \quad H_1: \mu_1 > \mu_2 - |M_{NI}|$$

$$H_0: \mu_1 - \mu_2 \leq -|M_{NI}| \quad \text{versus} \quad H_1: \mu_1 - \mu_2 > -|M_{NI}|$$

$$H_0: \delta \leq -|M_{NI}| \quad \text{versus} \quad H_1: \delta > -|M_{NI}|$$

### 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.

$$H_0: \mu_1 \geq \mu_2 + |M_{NI}| \quad \text{versus} \quad H_1: \mu_1 < \mu_2 + |M_{NI}|$$

$$H_0: \mu_1 - \mu_2 \geq |M_{NI}| \quad \text{versus} \quad H_1: \mu_1 - \mu_2 < |M_{NI}|$$

$$H_0: \delta \geq |M_{NI}| \quad \text{versus} \quad H_1: \delta < |M_{NI}|$$

## Test Statistic

This section describes the test statistic that is used to perform the hypothesis test.

### T-Test

A *t*-test is used to analyze the data. When the data are balanced between sequences, the two-sided *t*-test is equivalent to an analysis of variance F-test. The test assumes that the data are a simple random sample from a population of normally distributed values that have the same variance. This assumption implies that the differences are continuous and normal. The calculation of the t-statistic proceeds as follow

$$t_d = \frac{(\bar{x}_T - \bar{x}_R) - \varepsilon}{\hat{\sigma}_w \sqrt{\frac{2}{N}}}$$

where $\hat{\sigma}_w^2$ is the within mean square error from the appropriate ANOVA table.

The significance of the test statistic is determined by computing the p-value. If this p-value is less than a specified level (usually 0.05), the hypothesis is rejected. That is, the one-sided null hypothesis is rejected at the $\alpha$ significance level if $t_d > t_{\alpha,N-2}$. Otherwise, no conclusion can be reached.

## Computing the Within-Subject Variance ($\sigma_w^2$)

The ANOVA F-test is calculated using a standard repeated-measures analysis of variance table in which the between factor is the sequence and the within factor is the treatment. The within mean square error provides an estimate of the within-subject variance, $\sigma_w^2$, where

$$\sigma_w^2 = \text{Variance}\left(e_{ijk}\right)$$

If prior studies used a *t*-test rather than an ANOVA to analyze the data, you may not have a direct estimate of $\sigma_w^2$. Instead, you may have an estimate of the variance of the period differences from the *t*-test ($\hat{\sigma}_P^2$), an estimate of the variance of the paired differences ($\hat{\sigma}_D^2$), or an estimate of the variances of the paired variables ($\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$) and the correlation between the paired variables ($\hat{\rho}$). The within-subject variance, $\sigma_w^2$, is functionally related to these other variances as described below. Any of these different variances may be entered directly into this procedure.

### Using the Variance of the Period Differences ($\sigma_P^2$)

The variance of the period differences for each subject within each sequence ($\sigma_P^2$) is defined as

$$\sigma_P^2 = \text{Variance}\left(\frac{Y_{i2k} - Y_{i1k}}{2}\right).$$

$\sigma_P^2$ has a functional relationship with the within-subject population variance ($\sigma_w^2$), namely

$$\sigma_P^2 = \frac{\sigma_w^2}{2},$$

such that

$$\sigma_w^2 = 2\sigma_P^2 \,.$$

The within-subject standard deviation ($\sigma_w$) is then

$$\sigma_w = \sqrt{2\sigma_P^2} \,.$$

## Using the Variance of the Paired Differences ($\sigma_D^2$)

The variance of the paired differences ($\sigma_D^2$) is defined as

$$\sigma_D^2 = \text{Variance}(Y_{i2k} - Y_{i1k}) \,.$$

$\sigma_D^2$ has a functional relationship with the within-subject population variance ($\sigma_w^2$), namely

$$\sigma_D^2 = 2\sigma_w^2 \,,$$

such that

$$\sigma_w^2 = \frac{\sigma_D^2}{2} \,.$$

The within-subject standard deviation ($\sigma_w$) is then

$$\sigma_w = \sqrt{\frac{\sigma_D^2}{2}} \,.$$

## Using the Variances of the Paired Variables ($\sigma_1^2$ and $\sigma_2^2$) and the Correlation Between the Paired Variables (ρ)

The variances of the paired variables ($\sigma_1^2$ and $\sigma_2^2$) and the correlation between the paired variables ($\rho$) are defined as

$$\sigma_1^2 = \text{Variance}(Y_{i1k})$$

$$\sigma_2^2 = \text{Variance}(Y_{i2k})$$

$$\rho = \text{Correlation}(Y_{i1k}, Y_{i2k})$$

The variance of paired differences ($\sigma_D^2$) can be computed from $\sigma_1^2$, $\sigma_2^2$ and $\rho$ as

$$\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 \,,$$

such that the within-subject population variance ($\sigma_w^2$) can be computed as

$$\sigma_w^2 = \frac{\sigma_D^2}{2} = \frac{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}{2} .$$

The within-subject standard deviation ($\sigma_w$) is then

$$\sigma_w = \sqrt{\frac{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}{2}} .$$

If $\sigma_1^2 = \sigma_2^2 = \sigma_x^2$, then with

$$\sigma_x^2 = \text{Variance}\left(Y_{ijk}\right) ,$$

the formula for $\sigma_w^2$ reduces to

$$\sigma_w^2 = \sigma_x^2(1 - \rho) .$$

The within-subject standard deviation ($\sigma_w$) is then

$$\sigma_w = \sqrt{\sigma_x^2(1 - \rho)} .$$

## Computing the Power

The power is calculated as follows.

1. Find $t_\alpha$ such that $1 - T_{df}(t_\alpha) = \alpha$, where $T_{df}(x)$ is the area under a central-*t* curve to the left of *x* and *df* = *N* - 2.
2. Calculate the noncentrality parameter: $\lambda = \frac{(\delta - \varepsilon)\sqrt{N}}{\sigma_w\sqrt{2}}$.
3. Calculate: Power = $1 - T'_{df,\lambda}(t_\alpha)$, where $T'_{df,\lambda}(x)$ is the area under a noncentral-*t* curve with degrees of freedom *df* and noncentrality parameter $\lambda$ to the left of *x*.

# Example 1 – Power Analysis

Suppose you want to consider the power of a balanced, cross-over design that will be analyzed using the t-test approach. You want to compute the power when the margin of equivalence is either 5 or 10 at several sample sizes between 5 and 50. The true difference between the means under H0 is assumed to be 0. Similar experiments have had an *Sw* of 10. The significance level is 0.025.

## 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 | **Power** |
| Higher Means Are | **Better (H1: δ > -NIM)** |
| Alpha | **0.025** |
| N (Total Sample Size) | **5 10 15 20 30 40 50** |
| NIM (Non-Inferiority Margin) | **5 10** |
| δ1 (Actual Difference to Detect) | **0** |
| Standard Deviation Input Type | **Enter the Within-Subject Population SD** |
| σw (Within-Subject Population SD) | **10** |

## Output

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

### Numeric Results

Solve For: Power
Higher Means Are: Better
Hypotheses: H0: δ ≤ -NIM vs. H1: δ > -NIM

| | Total | | | | | |
|---|---|---|---|---|---|---|
| | Sample | Non-Inferiority | Actual | Standard | | |
| | Size | Margin | Difference | Deviation | | |
| Power | N | -NIM | δ1 | σw | Alpha | Beta |
| 0.08310 | 5 | -5 | 0 | 10 | 0.025 | 0.91690 |
| 0.16563 | 10 | -5 | 0 | 10 | 0.025 | 0.83437 |
| 0.24493 | 15 | -5 | 0 | 10 | 0.025 | 0.75507 |
| 0.32175 | 20 | -5 | 0 | 10 | 0.025 | 0.67825 |
| 0.46414 | 30 | -5 | 0 | 10 | 0.025 | 0.53586 |
| 0.58682 | 40 | -5 | 0 | 10 | 0.025 | 0.41318 |
| 0.68785 | 50 | -5 | 0 | 10 | 0.025 | 0.31215 |
| 0.20131 | 5 | -10 | 0 | 10 | 0.025 | 0.79869 |
| 0.50245 | 10 | -10 | 0 | 10 | 0.025 | 0.49755 |
| 0.71650 | 15 | -10 | 0 | 10 | 0.025 | 0.28350 |
| 0.84845 | 20 | -10 | 0 | 10 | 0.025 | 0.15155 |
| 0.96222 | 30 | -10 | 0 | 10 | 0.025 | 0.03778 |
| 0.99173 | 40 | -10 | 0 | 10 | 0.025 | 0.00827 |
| 0.99835 | 50 | -10 | 0 | 10 | 0.025 | 0.00165 |

Power The probability of rejecting a false null hypothesis when the alternative hypothesis is true.
N The total sample size drawn from all sequences. The sample is divided equally among sequences.
-NIM The magnitude and direction of the margin of non-inferiority. Since higher means are better, this value is negative and is the distance below the reference mean that is still considered non-inferior.
δ The difference in means. Difference (δ) = Treatment Mean (μT) - Reference Mean (μR).
δ1 The actual mean difference under the alternative hypothesis at which the power is computed.
σw The within-subject population standard deviation. σw = √[var(e_ijk)]. σw is estimated as the square root of the within mean square error (WMSE) (i.e., σw = √[WMSE]) from a repeated measures ANOVA analysis of a prior cross-over design.
Alpha The probability of rejecting a true null hypothesis.
Beta The probability of failing to reject the null hypothesis when the alternative hypothesis is true.

### Summary Statements

A 2×2 cross-over design (where higher means are considered to be better) will be used to test whether the treatment mean (μT) is non-inferior to the reference mean (μR), with a non-inferiority margin of -5 (H0: μT - μR ≤ -5 versus H1: μT - μR > -5). The comparison will be made using a one-sided t-test, with a Type I error rate (α) of 0.025. The within-subject population standard deviation is assumed to be 10. To detect a difference in means (μT - μR) of 0, with a total sample size of 5 (allocated equally to the two sequences), the power is 0.0831.

### Dropout-Inflated Sample Size

| Dropout Rate | Sample Size<br>N | Dropout-<br>Inflated<br>Enrollment<br>Sample Size<br>N' | Expected<br>Number of<br>Dropouts<br>D |
|---|---|---|---|
| 20% | 5 | 7 | 2 |
| 20% | 10 | 13 | 3 |
| 20% | 15 | 19 | 4 |
| 20% | 20 | 25 | 5 |
| 20% | 30 | 38 | 8 |
| 20% | 40 | 50 | 10 |
| 20% | 50 | 63 | 13 |

Dropout Rate The percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e., will be treated as "missing"). Abbreviated as DR.

N The evaluable sample size at which power is computed (as entered by the user). If N subjects are evaluated out of the N' subjects that are enrolled in the study, the design will achieve the stated power.

N' The total number of subjects that should be enrolled in the study in order to obtain N evaluable subjects, based on the assumed dropout rate. N' is calculated by inflating N using the formula N' = N / (1 - DR), with N' always rounded up. (See Julious, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. (2018) pages 32-33.)

D The expected number of dropouts. D = N' - N.

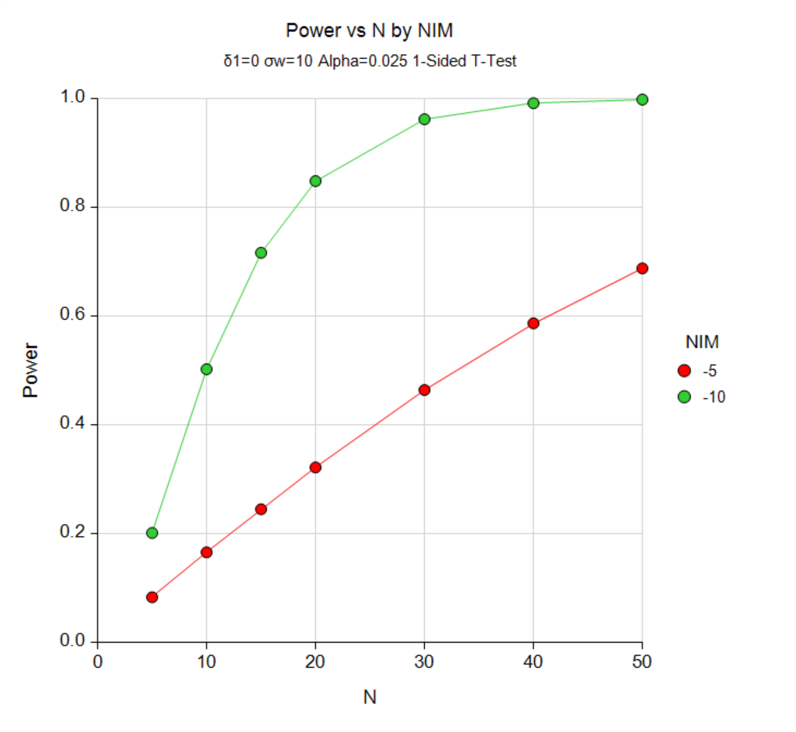
### Dropout Summary Statements

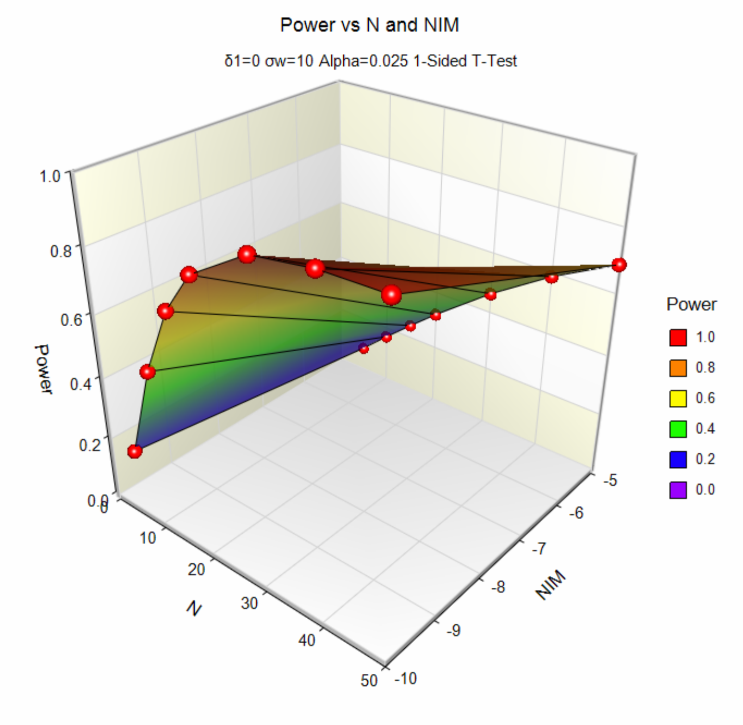
Anticipating a 20% dropout rate, 7 subjects should be enrolled to obtain a final sample size of 5 subjects.

### References

Chow, S.C. and Liu, J.P. 1999. Design and Analysis of Bioavailability and Bioequivalence Studies. Marcel Dekker. New York
Chow, S.C., Shao, J., and Wang, H. 2003. Sample Size Calculations in Clinical Research. Marcel Dekker. New York.
Julious, Steven A. 2004. 'Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data.' Statistics in Medicine, 23:1921-1986.
Senn, Stephen. 2002. Cross-over Trials in Clinical Research. Second Edition. John Wiley & Sons. New York.

## Plots





This report shows the values of each of the parameters, one scenario per row. The plots show the relationship between sample size and power. We see that a sample size of about 20 is needed to achieve 80% power when NIM = -10.

# Example 2 – Finding the Sample Size

Continuing with Example 1, suppose the researchers want to find the exact sample size necessary to achieve 90% power for both values of δ1.

## 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 ......................................................Sample Size
Higher Means Are.........................................Better (H1: δ > -NIM)
Power............................................................0.90
Alpha............................................................0.025
NIM (Non-Inferiority Margin) ..........................5 10
δ1 (Actual Difference to Detect).....................0
Standard Deviation Input Type ......................Enter the Within-Subject Population SD
σw (Within-Subject Population SD)................10
```

## Output

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

```
Numeric Results

Solve For:            Sample Size
Higher Means Are:     Better
Hypotheses:           H0: δ ≤ -NIM   vs.   H1: δ > -NIM

             Total
            Sample   Non-Inferiority      Actual    Standard
              Size            Margin  Difference   Deviation
Power            N              -NIM          δ1          σw     Alpha      Beta
0.90648         88                -5           0          10     0.025   0.09352
0.91139         24               -10           0          10     0.025   0.08861
```

This report shows the exact sample size necessary for each scenario.

Note that the search for N is conducted across only even values of N since the design is assumed to be balanced.

# Example 3 – Validation using Julious (2004)

Julious (2004) page 1953 presents an example in which δ1 = 0.0, NIM = 10, $\sigma_w$ = 20.00, alpha = 0.025, and beta = 0.10. Julious obtains a sample size of 86.

## 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 3** 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 > -NIM)
Power............................................................0.90
Alpha............................................................0.025
NIM (Non-Inferiority Margin) ..........................10
δ1 (Actual Difference to Detect).....................0
Standard Deviation Input Type ......................Enter the Within-Subject Population SD
σw (Within-Subject Population SD)................20
```

## Output

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

```
Numeric Results

Solve For:          Sample Size
Higher Means Are:   Better
Hypotheses:         H0: δ ≤ -NIM   vs.   H1: δ > -NIM

                Total
               Sample   Non-Inferiority      Actual    Standard
                 Size            Margin  Difference   Deviation
Power               N              -NIM          δ1          σw     Alpha      Beta
0.90648            88               -10           0          20     0.025   0.09352
```

**PASS** obtained a sample size of 88, two higher than that obtained by Julious (2004). However, if you look at the power achieved by an N of 86, you will find that it is 0.899997—slightly less than the goal of 0.90.