Chapter 507

Mann-Whitney U or Wilcoxon Rank-Sum Tests for Superiority by a Margin

Introduction

This procedure provides sample size and power calculations for one-sided two-sample Mann-Whitney U or Wilcoxon Rank-Sum Tests for superiority by a margin. This test is the nonparametric alternative to the traditional equal-variance two-sample *t*-test. Other names for this test are the Mann-Whitney-Wilcoxon test or the Wilcoxon-Mann-Whitney test.

Measurements are made on individuals that have been randomly assigned to one of two groups. This is sometimes referred to as a *parallel-groups* design. This design is used in situations such as the comparison of the income level of two regions, the nitrogen content of two lakes, or the effectiveness of two drugs.

The details of sample size calculation for the two-sample design are presented in the Mann-Whitney U or Wilcoxon Rank-Sum Tests chapter and they will not be duplicated here. This chapter only discusses those changes necessary for superiority tests.

The Statistical Hypotheses

Remember that in the usual *t*-test setting, the null (H0) and alternative (H1) hypotheses for one-sided tests are defined as

$$H_0: \mu_1 - \mu_2 \leq \delta_0 \quad \text{versus} \quad H_1: \mu_1 - \mu_2 > \delta_0$$

or equivalently

$$H_0: \delta \leq \delta_0 \quad \text{versus} \quad H_1: \delta > \delta_0.$$

Rejecting this test implies that the mean difference is larger than the value δ_0. 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 δ_0.

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

$$H_0: \mu_1 - \mu_2 \geq \delta_0 \quad \text{versus} \quad H_1: \mu_1 - \mu_2 < \delta_0$$

or equivalently

$$H_0: \delta \geq \delta_0 \quad \text{versus} \quad H_1: \delta < \delta_0.$$

Superiority by a Margin 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.
M_S	SM	*Margin of superiority*. This is a tolerance value that defines the magnitude of difference that is not of practical importance. This may be thought of as the smallest difference from the reference value that is considered to be of practical significance. This value is assumed to be a positive number.
δ	δ	*Actual difference*. This is the value of $\mu_1 - \mu_2$, the difference between the means. This is the value at which the power is calculated.

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.

Superiority by a Margin Tests

A *superiority by a margin test* tests that the treatment mean is better than the reference mean by more than the superiority margin. The actual direction of the hypothesis depends on the response variable being studied.

Case 1: High Values Good

In this case, higher values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is greater than the reference mean by at least the margin of superiority. The value of δ at which power is calculated must be greater than $|M_S|$. The null and alternative hypotheses with $\delta_0 = |M_S|$ are

$H_0: \mu_1 \leq \mu_2 + |M_S|$ versus $H_1: \mu_1 > \mu_2 + |M_S|$

$H_0: \mu_1 - \mu_2 \leq |M_S|$ versus $H_1: \mu_1 - \mu_2 > |M_S|$

$H_0: \delta \leq |M_S|$ versus $H_1: \delta > |M_S|$

Case 2: High Values Bad

In this case, higher values are worse. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is less than the reference mean by at least the margin of superiority. The value of δ at which power is calculated must be less than $-|M_S|$. The null and alternative hypotheses with $\delta_0 = -|M_S|$ are

$H_0: \mu_1 \geq \mu_2 - |M_S|$ versus $H_1: \mu_1 < \mu_2 - |M_S|$

$H_0: \mu_1 - \mu_2 \geq -|M_S|$ versus $H_1: \mu_1 - \mu_2 < -|M_S|$

$H_0: \delta \geq -|M_S|$ versus $H_1: \delta < -|M_S|$

Example

A superiority test example will set the stage for the discussion of the terminology that follows. Suppose that a test is to be conducted to determine if a new cancer treatment substantially improves mean bone density. The adjusted mean bone density (AMBD) in the population of interest is 0.002300 gm/cm with a standard deviation of 0.000300 gm/cm. Clinicians decide that if the treatment increases AMBD by more than 5% (0.000115 gm/cm), it provides a significant health benefit.

The hypothesis of interest is whether the mean AMBD in the treated group is more than 0.000115 above that of the reference group. The statistical test will be set up so that if the null hypothesis is rejected, the conclusion will be that the new treatment is superior. The value 0.000115 gm/cm is called the *margin of superiority*.

Mann-Whitney U or Wilcoxon Rank-Sum Test Statistic

This test is the nonparametric substitute for the equal-variance t-test. Two key assumptions are that the distributions are at least ordinal and that they are identical under H0. This means that ties (repeated values) are not acceptable. When ties are present, you can use approximations, but the theoretic results no longer hold.

The Mann-Whitney test statistic is defined as follows in Gibbons (1985).

$$z = \frac{W_1 - \frac{N_1(N_1 + N_2 + 1)}{2} + C}{s_W}$$

where

$$W_1 = \sum_{k=1}^{N_1} Rank\left(X_{1k}\right)$$

The ranks are determined after combining the two samples. The standard deviation is calculated as

$$s_W = \sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12} - \frac{N_1 N_2 \sum_{i=1} (t_i^3 - t_i)}{12(N_1 + N_2)(N_1 + N_2 - 1)}}$$

where t_i is the number of observations tied at value one, t_2 is the number of observations tied at some value two, and so forth.

The correction factor, *C*, is 0.5 if the rest of the numerator is negative or -0.5 otherwise. The value of *z* is then compared to the normal distribution.

Computing the Power

The power calculation for the Mann-Whitney U or Wilcoxon Rank-Sum Test is the same as that for the two-sample equal-variance *t*-test except that an adjustment is made to the sample size based on an assumed data distribution as described in Al-Sunduqchi and Guenther (1990). The sample size n_i' used in power calculations is equal to

$$n_i' = n_i / W,$$

where W is the Wilcoxon adjustment factor based on the assumed data distribution.

The adjustments are as follows:

Distribution	W
Double Exponential	$2/3$
Logistic	$9/\pi^2$
Normal	$\pi/3$

When $\sigma_1 = \sigma_2 = \sigma$, the power of the equal-variance *t*-test is calculated as follows.

1. Find t_α such that $1 - T_{df}(t_\alpha) = \alpha$, where $T_{df}(x)$ is the area to the left of *x* under a central-*t* distribution with degrees of freedom, $df = n_1' + n_2' - 2$.
2. Calculate: $\sigma_{\bar{X}} = \sigma\sqrt{\frac{1}{n_1'} + \frac{1}{n_2'}}$.
3. Calculate the noncentrality parameter: $\lambda = \frac{\delta - \delta_0}{\sigma_{\bar{X}}}$.
4. Calculate: $Power = 1 - T'_{df,\lambda}(t_\alpha)$, where $T'_{df,\lambda}(x)$ is the area to the left of *x* under a noncentral-*t* distribution with degrees of freedom, $df = n_1' + n_2' - 2$, and noncentrality parameter, λ.

When solving for something other than power, PASS uses this same power calculation formulation, but performs a search to determine that parameter.

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 calculated from the values of the other parameters.

Select *Sample Size* when you want to determine the sample size needed to achieve a given power and alpha.

Select *Power* when you want to calculate the power of an experiment that has already been run.

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 superiority test.

- **Better (H1: δ > SM)**
  If higher means are Better, the null hypothesis is H0: $\delta \leq SM$, and the alternative hypothesis is H1: $\delta > SM$.

- **Worse (H1: δ < -SM)**
  If higher means are Worse, the null hypothesis is H0: $\delta \geq -SM$, and the alternative hypothesis is H1: $\delta < -SM$.

Data Distribution

This option makes appropriate sample size adjustments for the Mann-Whitney U or Wilcoxon Rank-Sum test. Results by Al-Sunduqchi and Guenther (1990) indicate that power calculations for the Mann-Whitney U or Wilcoxon Rank-Sum test may be made using the standard t-test formulations with a simple adjustment to the sample size. The size of the adjustment depends upon the actual distribution of the data. They give sample size adjustment factors for three distributions. Select a distribution similar in shape to your data.

If Ni is the group sample size and W is the adjustment factor, then the distribution-adjusted group sample size is

$$Ni' = Ni/W$$

The options are:

- **Uniform**
  The sample size adjustment factor, W, is equal to “1”. This selection gives the same result as the one-sample t-test.

- **Double Exponential**
  The sample size adjustment factor, W, is equal to “2/3”.

- **Logistic**
  The sample size adjustment factor, W, is equal to “$9/\pi^2$”.

- **Normal**
  The sample size adjustment factor, W, is equal to “$\pi/3$”.

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. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of inferiority when the null hypothesis should be 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 here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

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. In this procedure, a type-I error occurs when you reject the null hypothesis of inferiority when in fact the mean is not non-inferior.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

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

Sample Size (When Solving for Sample Size)

Group Allocation

Select the option that describes the constraints on *N1* or *N2* or both.

The options are

- **Equal (N1 = N2)**
  This selection is used when you wish to have equal sample sizes in each group. Since you are solving for both sample sizes at once, no additional sample size parameters need to be entered.

- **Enter N2, solve for N1**
  Select this option when you wish to fix *N2* at some value (or values), and then solve only for *N1*. Please note that for some values of *N2*, there may not be a value of *N1* that is large enough to obtain the desired power.

- **Enter R = N2/N1, solve for N1 and N2**
  For this choice, you set a value for the ratio of *N2* to *N1*, and then PASS determines the needed *N1* and *N2*, with this ratio, to obtain the desired power. An equivalent representation of the ratio, *R*, is
  $$N2 = R * N1.$$

- **Enter percentage in Group 1, solve for N1 and N2**
  For this choice, you set a value for the percentage of the total sample size that is in Group 1, and then PASS determines the needed *N1* and *N2* with this percentage to obtain the desired power.

N2 (Sample Size, Group 2)

This option is displayed if Group Allocation = "Enter N2, solve for N1"

N2 is the number of items or individuals sampled from the Group 2 population.

N2 must be ≥ 2. You can enter a single value or a series of values.

R (Group Sample Size Ratio)

This option is displayed only if Group Allocation = "Enter R = N2/N1, solve for N1 and N2."

R is the ratio of *N2* to *N1*. That is,

$$R = N2 / N1.$$

Use this value to fix the ratio of *N2* to *N1* while solving for *N1* and *N2*. Only sample size combinations with this ratio are considered.

N2 is related to *N1* by the formula:

$$N2 = [R \times N1],$$

where the value *[Y]* is the next integer ≥ *Y*.

For example, setting *R* = 2.0 results in a Group 2 sample size that is double the sample size in Group 1 (e.g., *N1* = 10 and *N2* = 20, or *N1* = 50 and *N2* = 100).

R must be greater than 0. If *R* < 1, then *N2* will be less than *N1*; if *R* > 1, then *N2* will be greater than *N1*. You can enter a single or a series of values.

Percent in Group 1

This option is displayed only if Group Allocation = "Enter percentage in Group 1, solve for N1 and N2."

Use this value to fix the percentage of the total sample size allocated to Group 1 while solving for *N1* and *N2*. Only sample size combinations with this Group 1 percentage are considered. Small variations from the specified percentage may occur due to the discrete nature of sample sizes.

The Percent in Group 1 must be greater than 0 and less than 100. You can enter a single or a series of values.

Sample Size (When Not Solving for Sample Size)

Group Allocation

Select the option that describes how individuals in the study will be allocated to Group 1 and to Group 2.

The options are

- **Equal (N1 = N2)**
  This selection is used when you wish to have equal sample sizes in each group. A single per group sample size will be entered.

- **Enter N1 and N2 individually**
  This choice permits you to enter different values for *N1* and *N2*.

- **Enter N1 and R, where N2 = R * N1**
  Choose this option to specify a value (or values) for *N1*, and obtain *N2* as a ratio (multiple) of *N1*.

- **Enter total sample size and percentage in Group 1**
  Choose this option to specify a value (or values) for the total sample size (*N*), obtain *N1* as a percentage of *N*, and then *N2* as *N* - *N1*.

Sample Size Per Group

This option is displayed only if Group Allocation = "Equal (N1 = N2)."

The Sample Size Per Group is the number of items or individuals sampled from each of the Group 1 and Group 2 populations. Since the sample sizes are the same in each group, this value is the value for *N1*, and also the value for *N2*.

The Sample Size Per Group must be ≥ 2. You can enter a single value or a series of values.

N1 (Sample Size, Group 1)

*This option is displayed if Group Allocation = "Enter N1 and N2 individually" or "Enter N1 and R, where N2 = R * N1."*

N1 is the number of items or individuals sampled from the Group 1 population.

N1 must be ≥ 2. You can enter a single value or a series of values.

N2 (Sample Size, Group 2)

This option is displayed only if Group Allocation = "Enter N1 and N2 individually."

N2 is the number of items or individuals sampled from the Group 2 population.

N2 must be ≥ 2. You can enter a single value or a series of values.

R (Group Sample Size Ratio)

*This option is displayed only if Group Allocation = "Enter N1 and R, where N2 = R * N1."*

R is the ratio of *N2* to *N1*. That is,

$$R = N2/N1$$

Use this value to obtain *N2* as a multiple (or proportion) of *N1*.

N2 is calculated from *N1* using the formula:

$$N2=[R \times N1],$$

where the value *[Y]* is the next integer ≥ *Y*.

For example, setting *R* = 2.0 results in a Group 2 sample size that is double the sample size in Group 1.

R must be greater than 0. If *R* < 1, then *N2* will be less than N1; if *R* > 1, then *N2* will be greater than *N1*. You can enter a single value or a series of values.

Total Sample Size (N)

This option is displayed only if Group Allocation = "Enter total sample size and percentage in Group 1."

This is the total sample size, or the sum of the two group sample sizes. This value, along with the percentage of the total sample size in Group 1, implicitly defines *N1* and *N2*.

The total sample size must be greater than one, but practically, must be greater than 3, since each group sample size needs to be at least 2.

You can enter a single value or a series of values.

Percent in Group 1

This option is displayed only if Group Allocation = "Enter total sample size and percentage in Group 1."

This value fixes the percentage of the total sample size allocated to Group 1. Small variations from the specified percentage may occur due to the discrete nature of sample sizes.

The Percent in Group 1 must be greater than 0 and less than 100. You can enter a single value or a series of values.

Effect Size – Mean Difference

SM (Superiority Margin)

This is the magnitude of the margin of superiority. It must be entered as a positive number. If a negative value is entered, the absolute value is used.

When higher means are better, this value is the distance above the reference mean that is required to be considered superior. When higher means are worse, this value is the distance below the reference mean that is required to be considered superior.

δ (Actual Difference to Detect)

This is the actual difference between the treatment mean and the reference mean at which the power is calculated. When higher means are better, δ > SM. When higher means are worse, δ < -SM.

Effect Size – Standard Deviation

σ (Standard Deviation)

The standard deviation entered here is the assumed standard deviation for both the Group 1 population and the Group 2 population. σ must be a positive number.

When σ is not known, you must supply an estimate. Press the small 'σ' button to the right to obtain calculation options for estimating the standard deviation.

Example 1 – Power Analysis

Suppose that a test is to be conducted to determine if a new cancer treatment improves bone density. The adjusted mean bone density (AMBD) in the population of interest is 0.002300 gm/cm with a standard deviation of 0.000300 gm/cm. Clinicians decide that if the treatment increases AMBD by more than 5% (0.000115 gm/cm), it generates a significant health benefit. They also want to consider what would happen if the margin of superiority is set to 2.5% (0.0000575 gm/cm).

The researchers will be performing a Mann-Whitney-Wilcoxon test instead of the *t*-test because it is anticipated that the distribution of the two populations is not Normal. The researchers assume that the Logistic distribution shape most closely resembles what they expect to observe from the data.

The analysis will be a superiority test at the 0.025 significance level. Power is to be calculated assuming that the new treatment has 7.5% improvement on AMBD. Several sample sizes between 10 and 800 will be analyzed. The researchers want to achieve a power of at least 90%. All numbers have been multiplied by 10000 to make the reports and plots easier to read.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Mann-Whitney U or Wilcoxon Rank-Sum Tests for Superiority by a Margin** procedure window by expanding **Means**, then **Two Independent Means**, then clicking on **Superiority by a Margin**, and then clicking on **Mann-Whitney U or Wilcoxon Rank-Sum Tests for Superiority by a Margin**. 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 (H1: δ > SM)**
Data Distribution	**Logistic**
Alpha	**0.025**
Group Allocation	**Equal (N1 = N2)**
Sample Size Per Group	**10 50 100 200 300 500 600 800**
SM (Superiority Margin)	**0.575 1.15**
δ (Actual Difference to Detect)	**1.725**
σ (Standard Deviation)	**3**

Annotated Output

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

Numeric Results and Plots

Numeric Results

δ = μ1 - μ2 = μT – μR
Higher Means are Better
Hypotheses: H0: δ ≤ SM vs. H1: δ > SM
Data Distribution: Logistic

Power	N1	N2	N	SM	δ	σ	Alpha
0.12553	10	10	20	0.575	1.725	3.0	0.025
0.50552	50	50	100	0.575	1.725	3.0	0.025
0.80438	100	100	200	0.575	1.725	3.0	0.025

(report continues)

Mann-Whitney U or Wilcoxon Rank-Sum Tests for Superiority by a Margin

References

Al-Sunduqchi, Mahdi S. 1990. Determining the Appropriate Sample Size for Inferences Based on the Wilcoxon Statistics. Ph.D. dissertation under the direction of William C. Guenther, Dept. of Statistics, University of Wyoming, Laramie, Wyoming.

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.

Julious, Steven A. 2004. 'Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data.' Statistics in Medicine, 23:1921-1986.

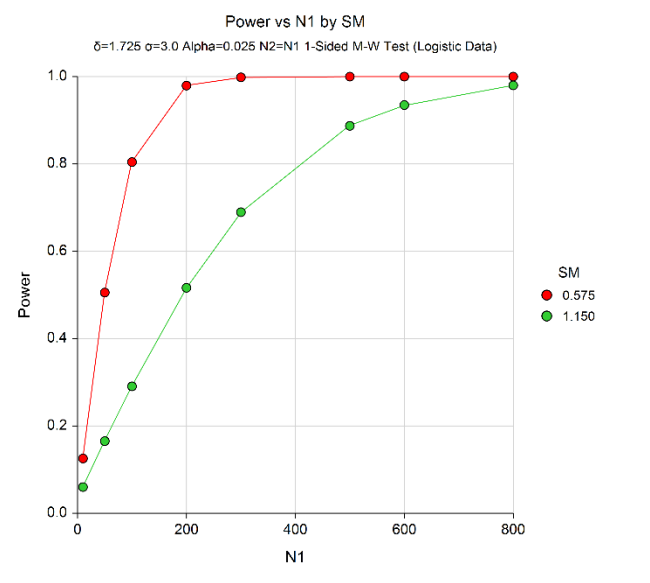
Report Definitions

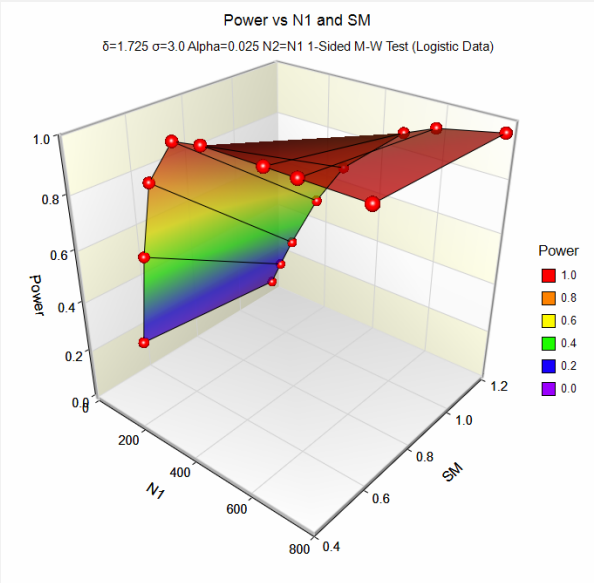
Power is the probability of rejecting a false null hypothesis.
N1 and N2 are the number of items sampled from each population.
N = N1 + N2 is the total sample size.
SM is the magnitude of the margin of superiority. Since higher means are better, this value is positive and is the distance above the reference mean that is required to be considered superior.
δ = μ1 - μ2 = μT - μR is the difference between the treatment and reference means at which power and sample size calculations are made.
σ is the assumed population standard deviation for each of the two groups.
Alpha is the probability of rejecting a true null hypothesis.

Summary Statements

Group sample sizes of 10 and 10 achieve 13% power to detect superiority using a one-sided, Mann-Whitney U or Wilcoxon Rank-Sum test assuming that the actual data distribution is logistic. The margin of superiority is 0.575. The actual difference between the means is assumed to be 1.725. The significance level (alpha) of the test is 0.025. The data are drawn from populations with a standard deviation of 3.0 in both groups.

Chart Section





The above report shows that for SM = 1.15, the sample size necessary to obtain 90% power is about 600 per group. However, if SM = 0.575, the required sample size is only about 150 per group.

Example 2 – Finding the Sample Size

Continuing with Example 1, the researchers want to know the exact sample size for each value of SM to achieve 90% power.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Mann-Whitney U or Wilcoxon Rank-Sum Tests for Superiority by a Margin** procedure window by expanding **Means**, then **Two Independent Means**, then clicking on **Superiority by a Margin**, and then clicking on **Mann-Whitney U or Wilcoxon Rank-Sum Tests for Superiority by a Margin**. 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	**Sample Size**
Higher Means Are	**Better (H1: δ > SM)**
Data Distribution	**Logistic**
Power	**0.90**
Alpha	**0.025**
Group Allocation	**Equal (N1 = N2)**
SM (Superiority Margin)	**0.575 1.15**
δ (Actual Difference to Detect)	**1.725**
σ (Standard Deviation)	**3**

Output

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

Numeric Results

Numeric Results

δ = μ1 - μ2 = μT – μR
Higher Means are Better
Hypotheses: H0: δ ≤ SM vs. H1: δ > SM
Data Distribution: Logistic

Target Power	Actual Power	N1	N2	N	SM	δ	σ	Alpha
0.90	0.90004	132	132	264	0.575	1.725	3.0	0.025
0.90	0.90036	523	523	1046	1.150	1.725	3.0	0.025

This report shows the exact sample size requirement for each value of SM.

Example 3 – Validation

This procedure uses the same mechanics as the Mann-Whitney U or Wilcoxon Rank-Sum Tests for Non-Inferiority procedure. We refer the user to Example 3 of Chapter 504 for the validation.