Chapter 504

Mann-Whitney U or Wilcoxon Rank-Sum Tests for Non-Inferiority

Introduction

This procedure provides sample size and power calculations for one-sided two-sample Mann-Whitney U or Wilcoxon Rank-Sum Tests for non-inferiority. 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 non-inferiority 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

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

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.
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 sign of the value will be determined by the specific design that is being used.
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.

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

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 at which power is calculated is often set to zero. The null and alternative hypotheses with $\delta_0 = -|M_{NI}|$ are

$H_0{:}\,\mu_1 \le \mu_2 - |M_{NI}|$ versus $H_1{:}\,\mu_1 > \mu_2 - |M_{NI}|$

$H_0{:}\,\mu_1 - \mu_2 \le -|M_{NI}|$ versus $H_1{:}\,\mu_1 - \mu_2 > -|M_{NI}|$

$H_0{:}\,\delta \le -|M_{NI}|$ versus $H_1{:}\,\delta > -|M_{NI}|$

Case 2: High Values Bad

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 at which power is calculated is often set to zero. The null and alternative hypotheses with $\delta_0 = |M_{NI}|$ are

$H_0{:}\,\mu_1 \ge \mu_2 + |M_{NI}|$ versus $H_1{:}\,\mu_1 < \mu_2 + |M_{NI}|$

$H_0{:}\,\mu_1 - \mu_2 \ge |M_{NI}|$ versus $H_1{:}\,\mu_1 - \mu_2 < |M_{NI}|$

$H_0{:}\,\delta \ge |M_{NI}|$ versus $H_1{:}\,\delta < |M_{NI}|$

Example

A non-inferiority 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 adversely affects 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 reduces AMBD by more than 5% (0.000115 gm/cm), it poses a significant health threat.

The hypothesis of interest is whether the mean AMBD in the treated group is more than 0.000115 below 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 non-inferior. The value 0.000115 gm/cm is called the *margin of non-inferiority.*

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).

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

where

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

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

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

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). For a Mann-Whitney U or Wilcoxon Rank-Sum Test group sample size of n_i, the adjusted 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.

Example 1 – Power Analysis

Suppose that a test is to be conducted to determine if a new cancer treatment adversely affects 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 reduces AMBD by more than 5% (0.000115 gm/cm), it poses a significant health threat. They also want to consider what would happen if the margin of equivalence 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.

Following accepted procedure, the analysis will be a non-inferiority test at the 0.025 significance level. Power is to be calculated assuming that the new treatment has no effect 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

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)**
Data Distribution	**Logistic**
Alpha	**0.025**
Group Allocation	**Equal (N1 = N2)**
Sample Size Per Group	**10 50 100 200 300 500 600 800**
NIM (Non-Inferiority Margin)	**0.575 1.15**
δ (Actual Difference to Detect)	**0**
σ (Standard Deviation)	**3**

Output

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

Numeric Results

Solve For:	Power
Test Type:	Two-Sample Mann-Whitney U or Wilcoxon Rank-Sum Test
Difference:	δ = μ1 - μ2 = μT - μR
Higher Means Are:	Better
Hypotheses:	H0: δ ≤ -NIM vs. H1: δ > -NIM
Data Distribution:	Logistic

	Sample Size			Non-Inferiority Margin	Mean Difference	Standard Deviation	
Power	N1	N2	N	-NIM	δ	σ	Alpha
0.06013	10	10	20	-0.575	0	3	0.025
0.16527	50	50	100	-0.575	0	3	0.025
0.29072	100	100	200	-0.575	0	3	0.025
0.51646	200	200	400	-0.575	0	3	0.025
0.68956	300	300	600	-0.575	0	3	0.025
0.88726	500	500	1000	-0.575	0	3	0.025
0.93488	600	600	1200	-0.575	0	3	0.025
0.97995	800	800	1600	-0.575	0	3	0.025
0.12553	10	10	20	-1.150	0	3	0.025
0.50552	50	50	100	-1.150	0	3	0.025
0.80438	100	100	200	-1.150	0	3	0.025
0.97945	200	200	400	-1.150	0	3	0.025
0.99839	300	300	600	-1.150	0	3	0.025
0.99999	500	500	1000	-1.150	0	3	0.025
1.00000	600	600	1200	-1.150	0	3	0.025
1.00000	800	800	1600	-1.150	0	3	0.025

Power The probability of rejecting a false null hypothesis when the alternative hypothesis is true.
N1 The sample size from group 1.
N2 The sample size from group 2.
N The total sample size from both groups. N = N1 + N2.
-NIM The magnitude and direction of the margin of non-inferiority. Since higher means are better, this value is negative and is the maximum distance below μ2 that μ1 can be and still conclude that group 1 is non-inferior to group 2.
δ The difference between the group means at which power and sample size calculations are made. δ = μ1 - μ2.
σ The assumed standard deviation for each of the two groups.
Alpha The probability of rejecting a true null hypothesis.

Summary Statements

A parallel two-group design will be used to test whether the Group 1 (treatment) location (distribution center) is non-inferior to the Group 2 (reference) location, with a non-inferiority margin of -0.575 (H0: δ ≤ -0.575 versus H1: δ > -0.575, δ = μ1 - μ2). The comparison will be made using a one-sided, two-sample Mann-Whitney U or Wilcoxon Rank-Sum test, with a Type I error rate (α) of 0.025. The common standard deviation for both groups is assumed to be 3, and the underlying data distribution is assumed to be Logistic. To detect a difference in means of 0, with sample sizes of 10 in Group 1 and 10 in Group 2, the power is 0.06013.

Dropout-Inflated Sample Size

	Sample Size			Dropout-Inflated Enrollment Sample Size			Expected Number of Dropouts		
Dropout Rate	**N1**	**N2**	**N**	**N1'**	**N2'**	**N'**	**D1**	**D2**	**D**
20%	10	10	20	13	13	26	3	3	6
20%	50	50	100	63	63	126	13	13	26
20%	100	100	200	125	125	250	25	25	50
20%	200	200	400	250	250	500	50	50	100
20%	300	300	600	375	375	750	75	75	150
20%	500	500	1000	625	625	1250	125	125	250
20%	600	600	1200	750	750	1500	150	150	300
20%	800	800	1600	1000	1000	2000	200	200	400

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.

N1, N2, and N The evaluable sample sizes at which power is computed (as entered by the user). If N1 and N2 subjects are evaluated out of the N1' and N2' subjects that are enrolled in the study, the design will achieve the stated power.

N1', N2', and N' The number of subjects that should be enrolled in the study in order to obtain N1, N2, and N evaluable subjects, based on the assumed dropout rate. N1' and N2' are calculated by inflating N1 and N2 using the formulas N1' = N1 / (1 - DR) and N2' = N2 / (1 - DR), with N1' and N2' 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.)

D1, D2, and D The expected number of dropouts. D1 = N1' - N1, D2 = N2' - N2, and D = D1 + D2.

Dropout Summary Statements

Anticipating a 20% dropout rate, 13 subjects should be enrolled in Group 1, and 13 in Group 2, to obtain final group sample sizes of 10 and 10, respectively.

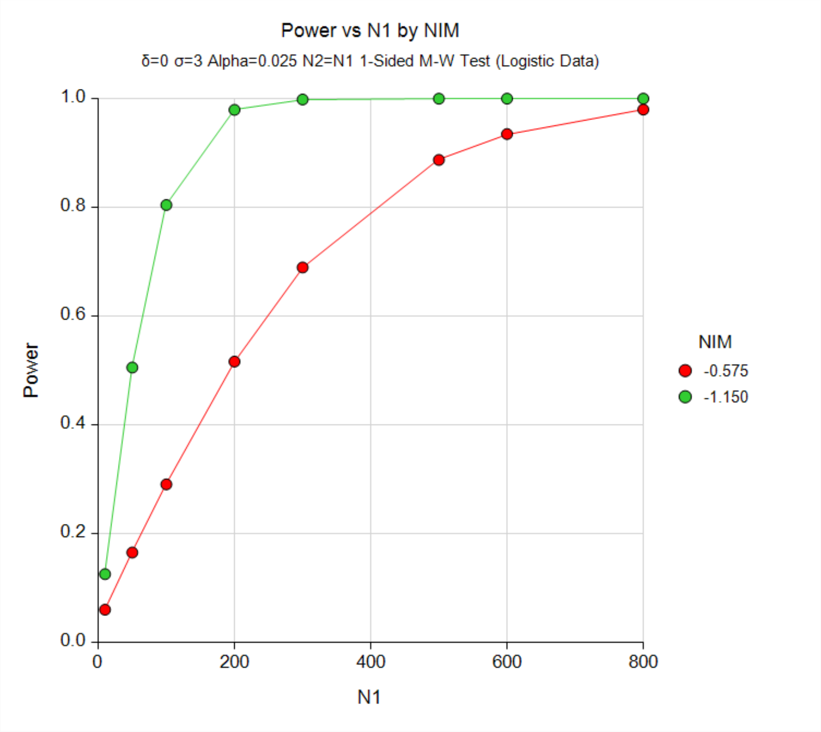
References

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.

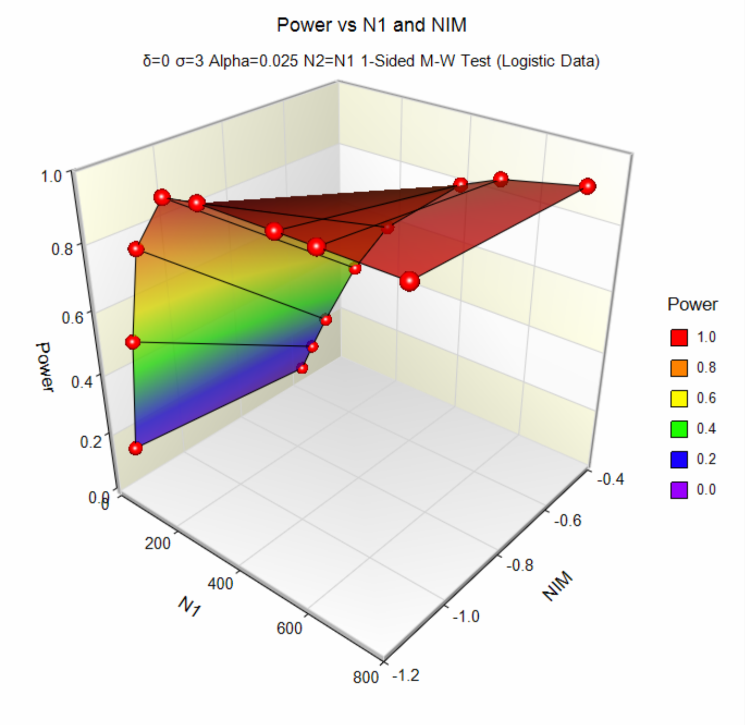
Julious, Steven A. 2004. 'Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data.' Statistics in Medicine, 23:1921-1986.

Plots



Power vs N1 by NIM

δ=0 σ=3 Alpha=0.025 N2=N1 1-Sided M-W Test (Logistic Data)



Power vs N1 and NIM

δ=0 σ=3 Alpha=0.025 N2=N1 1-Sided M-W Test (Logistic Data)

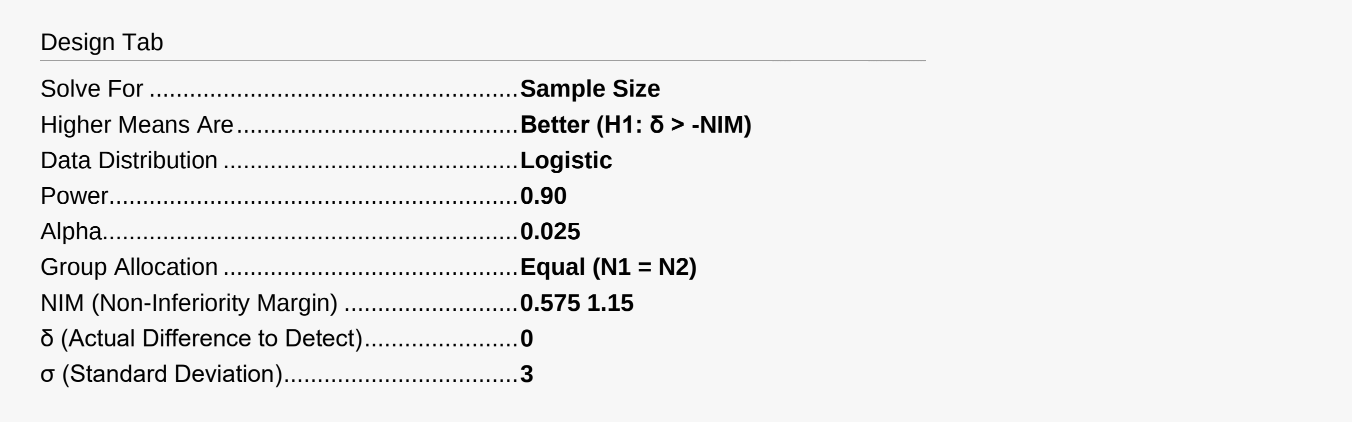
The above report shows that for NIM = 1.15, the sample size necessary to obtain 90% power is about 125 per group. However, if NIM = 0.575, the required sample size is about 525 per group.

Example 2 – Finding the Sample Size

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

Setup

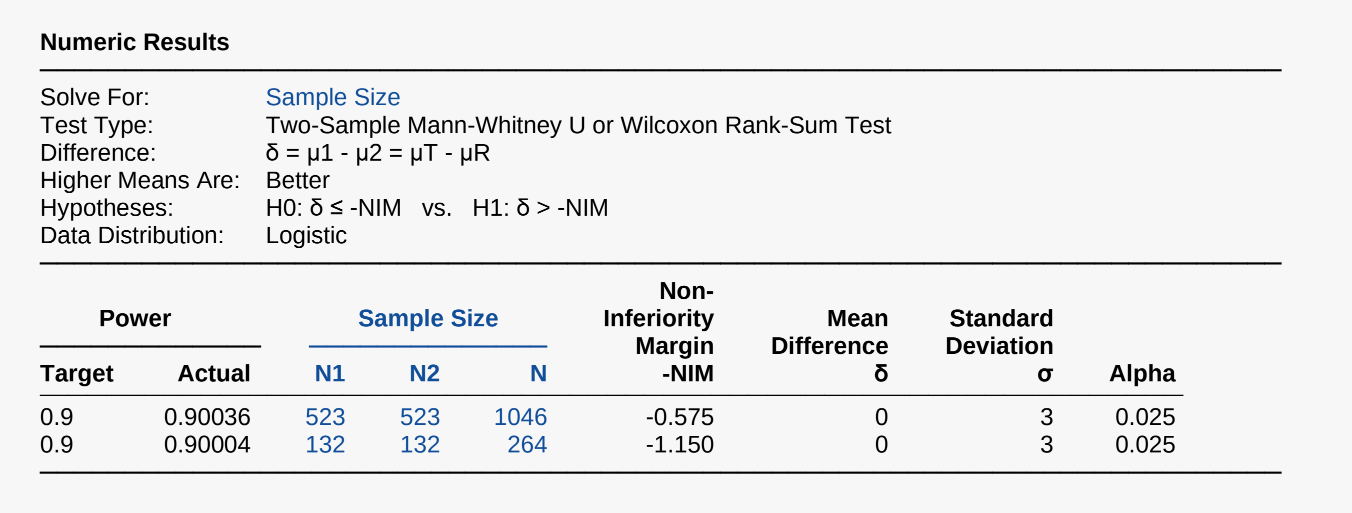
If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.



```
Design Tab
Solve For ......................................................Sample Size
Higher Means Are.........................................Better (H1: δ > -NIM)
Data Distribution ...........................................Logistic
Power............................................................0.90
Alpha............................................................0.025
Group Allocation ...........................................Equal (N1 = N2)
NIM (Non-Inferiority Margin) ..........................0.575 1.15
δ (Actual Difference to Detect).......................0
σ (Standard Deviation)..................................3
```

Output

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



```
Numeric Results

Solve For:           Sample Size
Test Type:           Two-Sample Mann-Whitney U or Wilcoxon Rank-Sum Test
Difference:          δ = μ1 - μ2 = μT - μR
Higher Means Are:    Better
Hypotheses:          H0: δ ≤ -NIM   vs.   H1: δ > -NIM
Data Distribution:   Logistic

                                                Non-
      Power              Sample Size        Inferiority       Mean     Standard
 ----------------    ------------------       Margin     Difference   Deviation
 Target    Actual     N1     N2       N          -NIM            δ           σ     Alpha

 0.9      0.90036    523    523    1046        -0.575            0           3     0.025
 0.9      0.90004    132    132     264        -1.150            0           3     0.025
```

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

Example 3 – Validation using Chow, Shao, Wang, and Lokhnygina (2018)

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 δ = 0, NIM = 0.05, σ = 0.1, Alpha = 0.05, and Power = 0.80.

The Mann-Whitney U or Wilcoxon Rank-Sum test power calculations are the same as the *t*-test except for non-inferiority except for an adjustment factor for the assumed data distribution. If we set the data distribution to Normal, we should get a result of N1 = N2 = 51 × π/3 = 53.407, which rounds up to 54.

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)**
Data Distribution .... **Normal**
Power.. **0.80**
Alpha.. **0.05**
Group Allocation .... **Equal (N1 = N2)**
NIM (Non-Inferiority Margin) **0.05**
δ (Actual Difference to Detect)....................... **0**
σ (Standard Deviation).................................. **0.1**

Output

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

Numeric Results

Solve For: Sample Size
Test Type: Two-Sample Mann-Whitney U or Wilcoxon Rank-Sum Test
Difference: δ = μ1 - μ2 = μT - μR
Higher Means Are: Better
Hypotheses: H0: δ ≤ -NIM vs. H1: δ > -NIM
Data Distribution: Normal

Power		Sample Size			Non-Inferiority Margin	Mean Difference	Standard Deviation	
Target	**Actual**	**N1**	**N2**	**N**	**-NIM**	**δ**	**σ**	**Alpha**
0.8	0.8059	54	54	108	-0.1	0	0.1	0.05

The sample size of 54 in each group matches the expected result.