

Chapter 499

Paired Wilcoxon Signed-Rank Tests for Non-Inferiority

Introduction

This procedure computes power and sample size for non-inferiority tests in paired-data designs in which the paired t -test assumptions are violated and is the *Paired Wilcoxon Signed-Rank Test* is used.

Paired Designs

Paired data may occur because two measurements are made on the same subject or because measurements are made on two subjects that have been matched according to other variables. Hypothesis tests on paired data can be analyzed by considering the difference between the paired items as the response. The distribution of differences is usually symmetric. In fact, the distribution must be symmetric if the individual distributions of the two items are identical. Hence, the paired t -test is appropriate for paired data even when the distributions of the individual items are not normal.

The Statistical Hypotheses

Both non-inferiority and superiority tests are examples of directional (one-sided) tests and their power and sample size could be calculated using the One-Sample T-Tests procedure. However, at the urging of our users, we have developed this procedure which provides the input and output options that are convenient for non-inferiority tests for paired data. This section will review the specifics of non-inferiority testing.

Remember that in the usual t -test setting with δ defined as the mean paired difference, the null (H_0) and alternative (H_1) hypotheses for one-sided upper-tail tests are defined as

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

Rejecting H_0 implies that the mean is larger than the value δ_0 . This test is called an *upper-tail test* because H_0 is rejected in samples in which the mean of paired differences is larger than δ_0 .

The *lower-tail test* is

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

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

Paired Wilcoxon Signed-Rank Tests for Non-Inferiority

<u>Parameter</u>	<u>PASS Input/Output</u>	<u>Interpretation</u>
δ	δ	<i>Population mean paired difference.</i> This is the mean of paired differences in the population. This parameter will be estimated by the study.
δ_1	δ_1	<i>Actual paired difference at which power is calculated.</i> This is the value of the mean paired difference at which power is calculated.
M_{NI}	NIM	<i>Margin of non-inferiority.</i> This is a tolerance value that defines the magnitude of difference that is not of practical importance. This may be thought of as the largest difference that is considered to be trivial. The sign is determined by the specific design. This value is used to compute the bound, δ_0 .

Non-Inferiority Tests

A *non-inferiority test* tests that the mean difference is not less than (or greater than) zero by more than a small non-inferiority margin. The actual direction of the hypothesis depends on the whether higher values of the response are good or bad.

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 mean difference no less than a small amount below zero. 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

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

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 mean of the treatment group is no more than a small amount above the reference value. 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

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

Wilcoxon Signed-Rank Test Statistic

The Wilcoxon signed-rank test is a popular, nonparametric substitute for the t -test. It assumes that the data follow a symmetric distribution. The test is computed using the following steps.

1. Subtract the hypothesized mean, δ_0 , from each data value. Rank the values according to their absolute values.
2. Compute the sum of the positive ranks S_p and the sum of the negative ranks S_n . The test statistic, W_R , is the minimum of S_p and S_n .
3. Compute the mean and standard deviation of W_R using the formulas

$$\mu_{W_R} = \frac{n(n+1)}{4}$$

$$\sigma_{W_R} = \sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}$$

where t represents the number of times the i^{th} value occurs.

Paired Wilcoxon Signed-Rank Tests for Non-Inferiority

4. Compute the z -value using

$$z_W = \frac{W_R - \mu_{W_R}}{\sigma_{W_R}}$$

The significance of the test statistic is determined by computing the p-value using the standard normal distribution. If this p-value is less than a specified level (usually 0.05), the null hypothesis is rejected in favor of the alternative hypothesis. Otherwise, no conclusion can be reached.

Power Calculation for the Wilcoxon Signed-Rank Test

The power calculation for the Wilcoxon signed-rank test is the same as that for the one-sample 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' used in power calculations is equal to

$$n' = n/W,$$

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

The adjustments are as follows:

Distribution	W
Uniform	1
Double Exponential	2/3
Logistic	$9/\pi^2$
Normal	$\pi/3$

The power is calculated as follows for a directional alternative (one-tailed test) in which $\delta_1 > \delta_0$.

1. Find t_α such that $1 - T_{df}(t_\alpha) = \alpha$, where $T_{df}(t_\alpha)$ is the area under a central- t curve to the left of x and $df = n' - 1$.
2. Calculate: $X_1 = \delta_0 + t_\alpha \frac{\sigma}{\sqrt{n'}}$.
3. Calculate the noncentrality parameter: $\lambda = \frac{\delta_1 - \delta_0}{\frac{\sigma}{\sqrt{n'}}}$.
4. Calculate: $t_1 = \frac{X_1 - \delta_1}{\frac{\sigma}{\sqrt{n'}}} + \lambda$.
5. Power = $1 - T'_{df,\lambda}(t_1)$, where $T'_{df,\lambda}(x)$ is the area to the left of x under a noncentral- t curve with degrees of freedom df and noncentrality parameter λ .

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 will be of interest.

Solve For

Solve For

This option specifies the parameter to be calculated from the values of the other parameters. Under most conditions, you would select either *Power* or *Sample Size*.

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

Select *Power* when you want to calculate the power.

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. Note that δ is the population mean of paired differences.

- **Better (H1: $\delta > -NIM$)**

If higher means are Better the null hypothesis is $H_0: \delta \leq -NIM$, and the alternative hypothesis is $H_1: \delta > -NIM$.

- **Worse (H1: $\delta < NIM$)**

If higher means are Worse the null hypothesis is $H_0: \delta \geq NIM$, and the alternative hypothesis is $H_1: \delta < NIM$.

Data Distribution

This option makes appropriate sample size adjustments for the Wilcoxon Signed-Rank test. Results by Al-Sundugchi and Guenther (1990) indicate that power calculations for the Wilcoxon Signed-Rank 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 four distributions.

The options are as follows:

- **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/π²”.

Paired Wilcoxon Signed-Rank Tests for Non-Inferiority

- **Normal**

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

Population Size

This is the number of subjects in the population. Usually, you assume that samples are drawn from a very large (infinite) population. Occasionally, however, situations arise in which the population of interest is of limited size. In these cases, appropriate adjustments must be made.

When a finite population size is specified, the standard deviation is reduced according to the formula:

$$\sigma_1^2 = \left(1 - \frac{n}{N}\right) \sigma^2$$

where n is the sample size, N is the population size, σ is the original standard deviation, and σ_1 is the new standard deviation.

The quantity n/N is often called the sampling fraction. The quantity $\left(1 - \frac{n}{N}\right)$ is called the *finite population correction factor*.

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 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.

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

N (Sample Size)

This option specifies one or more values of the sample size, the number of individuals in the study. This value must be an integer greater than one. You may enter a list of values using the syntax *50 100 150 200 250* or *50 to 250 by 50*.

Paired Wilcoxon Signed-Rank Tests for Non-Inferiority

Effect Size – Means

NIM (Non-Inferiority Margin)

This is the magnitude of the margin of non-inferiority for the mean of paired differences. It should 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 below zero for which the new treatment will still be considered non-inferior to the standard treatment.

When higher means are worse, this value is the distance above zero for which the new treatment will still be considered non-inferior to the standard treatment.

δ_1 (Mean of Paired Differences)

Enter a value (or range of values) for the mean of paired differences at which power and sample size are calculated. This value indicates the minimum detectible paired difference.

For non-inferiority tests, this value is often set to 0. When higher means are better, this value should be greater than -NIM. When higher means are worse, this value should be less than NIM.

Effect Size – Standard Deviation

σ (Std Dev of Paired Differences)

Enter an estimate of the standard deviation of the paired differences (must be positive).

When this value is not known, you must supply an estimate of it. **PASS** includes a special tool for estimating the standard deviation. This tool may be loaded by pressing the *SD* button. Refer to the Standard Deviation Estimator chapter for further details.

Paired Wilcoxon Signed-Rank Tests for Non-Inferiority

Example 1 – Power Analysis

Suppose that a test is to be conducted to determine if a new cancer treatment adversely affects the mean bone density. The adjusted mean bone density (AMBD) in the population of interest is 0.002300 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 non-inferiority is set to 2.5% (0.0000575 gm/cm). The standard deviation of paired differences is 0.000300 gm/cm.

Following accepted procedure, the analysis will be a non-inferiority test using the Wilcoxon signed-rank test assuming a Normal data distribution 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 20 and 300 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 **Paired Wilcoxon Signed-Rank Tests for Non-Inferiority** procedure window by expanding **Means**, then **Paired Means**, then clicking on **Non-Inferiority**, and then clicking on **Paired Wilcoxon Signed-Rank Tests for Non-Inferiority**. 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**.

<u>Option</u>	<u>Value</u>
Design Tab	
Solve For	Power
Higher Means Are	Better (H1: $\delta > -NIM$)
Data Distribution	Normal
Population Size	Infinite
Alpha	0.025
N (Sample Size)	20 40 60 80 100 150 200 300
NIM (Non-Inferiority Margin)	0.575 1.15
δ_1 (Mean of Paired Differences)	0
σ (Std Dev of Paired Differences)	3

Annotated Output

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

Numeric Results

Higher Means are Better

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

Data Distribution: Normal

	Non-Inferiority Margin	Mean of Paired Differences	Standard Deviation		
Power	N	-NIM	δ_1	σ	Beta
0.12134	20	-0.575	0.000	3.0	0.025 0.87866
0.20927	40	-0.575	0.000	3.0	0.025 0.79073
0.29540	60	-0.575	0.000	3.0	0.025 0.70460
0.37811	80	-0.575	0.000	3.0	0.025 0.62189
0.45584	100	-0.575	0.000	3.0	0.025 0.54416
0.62419	150	-0.575	0.000	3.0	0.025 0.37581
0.74810	200	-0.575	0.000	3.0	0.025 0.25190
0.89804	300	-0.575	0.000	3.0	0.025 0.10196
0.35274	20	-1.150	0.000	3.0	0.025 0.64726

(report continues)

Paired Wilcoxon Signed-Rank Tests for Non-Inferiority

Report Definitions

Power is the probability of rejecting a false null hypothesis. It should be close to one.

N is the sample size, the number of subjects (or pairs) in the study.

-NIM is 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 value that is still considered non-inferior.

δ is the mean of paired differences.

δ_1 is the value of the mean of paired differences at which power and sample size are calculated.

σ is the standard deviation of paired differences for the population.

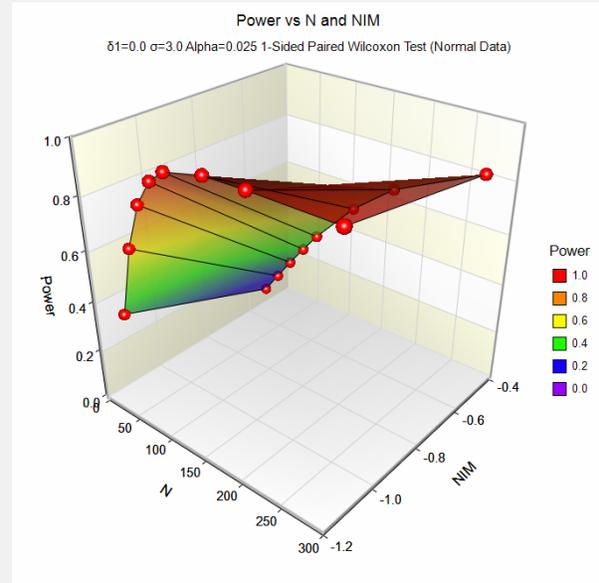
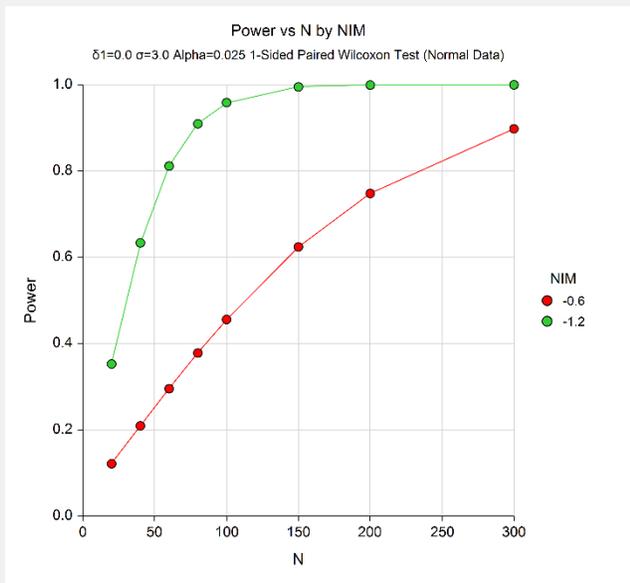
Alpha is the probability of rejecting a true null hypothesis. It should be small.

Beta is the probability of accepting a false null hypothesis. It should be small.

Summary Statements

A sample size of 20 achieves 12% power to detect non-inferiority using a one-sided paired Wilcoxon Signed-Rank test assuming that the actual distribution of paired differences is normal when the non-inferiority margin is -0.575 and the actual mean of paired differences is 0.000.

The an estimated standard deviation of paired differences is 3.0. The significance level (alpha) of the test is 0.025.



The above report shows that for NIM = 1.15, the sample size necessary to obtain 90% power is just under 80. However, if NIM = 0.575, the required sample size is about 300.

Example 2 – Finding the Sample Size

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

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Paired Wilcoxon Signed-Rank Tests for Non-Inferiority** procedure window by expanding **Means**, then **Paired Means**, then clicking on **Non-Inferiority**, and then clicking on **Paired Wilcoxon Signed-Rank Tests for Non-Inferiority**. 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**.

<u>Option</u>	<u>Value</u>
Design Tab	
Solve For	Sample Size
Higher Means Are	Better (H1: $\delta > -NIM$)
Data Distribution	Normal
Population Size	Infinite
Power	0.90
Alpha	0.025
NIM (Non-Inferiority Margin)	0.575 1.15
δ_1 (Mean of Paired Differences)	0
σ (Std Dev of Paired Differences)	3

Output

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

Numeric Results

Numeric Results							
Higher Means are Better							
Hypotheses: $H_0: \delta \leq -NIM$ vs. $H_1: \delta > -NIM$							
Data Distribution: Normal							
	Non-Inferiority Margin		Mean of Paired Differences	Standard Deviation			
Power	N	-NIM	δ_1	σ	Alpha	Beta	
0.90005	302	-0.575	0.000	3.0	0.025	0.09995	
0.90215	78	-1.150	0.000	3.0	0.025	0.09785	

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 46 has an example of a sample size calculation for a non-inferiority trial using a paired t -test. Their example obtains a sample size of 8 when $\delta_1 = 0.5$, $NIM = 0.5$, $\sigma = 1$, $\text{Alpha} = 0.05$, and $\text{Power} = 0.80$.

The paired Wilcoxon Signed-Rank test power calculations are the same as the paired t -test except for an adjustment factor for the assumed data distribution. If we assume a uniform data distribution, we should get the same value of $N = 8$. If we assume a Normal data distribution, then the expected sample size would be $N = 8 \times \pi/3 = 9$ after rounding up.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Paired Wilcoxon Signed-Rank Tests for Non-Inferiority** procedure window by expanding **Means**, then **Paired Means**, then clicking on **Non-Inferiority**, and then clicking on **Paired Wilcoxon Signed-Rank Tests for Non-Inferiority**. You may then make the appropriate entries as listed below, or open **Example 3 (a or b)** by going to the **File** menu and choosing **Open Example Template**.

<u>Option</u>	<u>Value</u>
Design Tab	
Solve For	Sample Size
Higher Means Are.....	Better (H1: $\delta > -NIM$)
Data Distribution	Uniform
Population Size.....	Infinite
Power.....	0.80
Alpha.....	0.05
NIM (Non-Inferiority Margin)	0.5
δ_1 (Mean of Paired Differences).....	0.5
σ (Std Dev of Paired Differences).....	1

Output

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

Numeric Results							
Higher Means are Better							
Hypotheses: $H_0: \delta \leq -NIM$ vs. $H_1: \delta > -NIM$							
Data Distribution: Uniform							
		Non-Inferiority Margin -NIM	Mean of Paired Differences δ_1	Standard Deviation σ		Alpha	Beta
Power	N						
0.81502	8	-0.5	0.5	1.0		0.050	0.18498

PASS also obtains a sample size of 8 with the uniform distribution.

Paired Wilcoxon Signed-Rank Tests for Non-Inferiority

If we now assume a Normal data distribution and solve for sample size, the results match our expected outcome.

Numeric Results

Higher Means are Better

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

Data Distribution: Normal

	Non-Inferiority Margin -NIM	Mean of Paired Differences δ_1	Standard Deviation σ	Alpha	Beta	
Power	N					
0.81502	9	-0.5	0.5	1.0	0.050	0.18498

The sample size of 9 matches the expected result.