Chapter 416

Wilcoxon Signed-Rank Tests for Non-Inferiority

Introduction

This procedure computes power and sample size for non-inferiority tests in one-sample designs in which the one-sample *t*-test assumptions are violated and is the *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.

In paired designs, the variable of interest is the difference between two individual measurements. Although the non-inferiority hypothesis refers to the difference between two individual means, the actual values of those means are not needed. All that is needed is their difference.

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 Wilcoxon Signed-Rank 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. 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_1) hypotheses for one-sided upper-tail tests are defined as

$$H_0: \mu \leq \mu_0 \quad \text{versus} \quad H_1: \mu > \mu_0$$

Rejecting H_0 implies that the mean is larger than the value μ_0. This test is called an *upper-tail test* because H_0 is rejected in samples in which the sample mean is larger than μ_0.

The *lower-tail test* is

$$H_0: \mu \geq \mu_0 \quad \text{versus} \quad H_1: \mu < \mu_0$$

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.

Parameter	PASS Input/Output	Interpretation
μ	μ	*Population mean*. If the data are paired differences, this is the mean of those differences. This parameter will be estimated by the study.
μ_1	$\mu 1$	*Actual population mean at which power is calculated*. This is the assumed population mean used in all calculations.
μ_0	$\mu 0$	*Non-Inferiority Mean*. This is the smallest (or largest) value of the mean for which the new treatment will still be considered non-inferior to the reference.
μ_R	μR	*Reference value*. Usually, this is the mean of a reference population. If the data are paired differences, this is the hypothesized value of the mean difference.
M_{NI}	NIM	*Margin of non-inferiority*. 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 from the reference value that is considered to be trivial. This value is assumed to be a positive number.

Non-Inferiority Tests

A *non-inferiority test* tests that the mean is not worse than that of the baseline (reference) population 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 of the treatment group is no less than a small amount below the reference value. The value of μ_1 at which power is calculated is often set to zero for paired tests.

Equivalent sets of the null and alternative hypotheses are

$H_0: \mu \leq \mu_0$ versus $H_1: \mu > \mu_0$

$H_0: \mu \leq \mu_R - M_{NI}$ versus $H_1: \mu > \mu_R - M_{NI}$

$H_0: \mu - \mu_R \leq -M_{NI}$ versus $H_1: \mu - \mu_R > -M_{NI}$

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 μ_1 at which power is calculated is often set to zero for paired tests.

Equivalent sets of the null and alternative hypotheses are

$H_0: \mu \geq \mu_0$ versus $H_1: \mu < \mu_0$

$H_0: \mu \geq \mu_R + M_{NI}$ versus $H_1: \mu < \mu_R + M_{NI}$

$H_0: \mu - \mu_R \geq M_{NI}$ versus $H_1: \mu - \mu_R < M_{NI}$

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 the 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 AMBD in the treated group is greater than 0.002300 – 0.000115 = 0.002185. The statistical test will be set up so that if the null hypothesis that the AMBD is less than or equal to 0.002185 is rejected, the conclusion will be that the new treatment is non-inferior, at least in terms of AMBD. The value 0.000115 gm/cm is called the *margin of non-inferiority*.

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, μ_0, from each data value. Rank the values according to their absolute values.
2. Compute the sum of the positive ranks *Sp* and the sum of the negative ranks *Sn*. The test statistic, W_R, is the minimum of *Sp* and *Sn*.
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.
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 $\mu_1 > \mu_0$.

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: $X_1 = \mu_0 + t_\alpha \frac{\sigma}{\sqrt{n'}}$.
3. Calculate the noncentrality parameter: $\lambda = \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n'}}}$.
4. Calculate: $t_1 = \frac{X_1 - \mu_1}{\frac{\sigma}{\sqrt{n'}}} + \lambda$.
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.

- **Better (H1: μ > μ0)**
  If higher means are Better the null hypothesis is H0: μ ≤ μ0, and the alternative hypothesis is H1: μ > μ0.
  μ0 = μR - NIM, where μR is the baseline, standard, or reference mean and NIM is the margin of non-inferiority.

- **Worse (H1: μ < μ0)**
  If higher means are Worse the null hypothesis is H0: μ ≥ μ0, and the alternative hypothesis is H1: μ < μ0.
  μ0 = μR + NIM, where μR is the baseline, standard, or reference mean and NIM is the margin of non-inferiority.

Data Distribution

This option makes appropriate sample size adjustments for the Wilcoxon Signed-Rank test. Results by Al-Sunduqchi 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 "$9/\pi^2$".

- **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*.

Effect Size – Means

μ0 (Non-Inferiority Mean)

Enter a value (or range of values) for the non-inferiority mean. Since higher means are worse, this is the largest value of the mean for which the new treatment will still be considered non-inferior to the reference.

Define μR as the baseline, standard, or reference mean and NIM as the margin of non-inferiority. When higher means are better, the non-inferiority mean is calculated as $\mu0 = \mu R - NIM$ and μ0 can be any number that satisfies $\mu0 < \mu1$. When higher means are worse, the non-inferiority mean is calculated as $\mu0 = \mu R + NIM$ and μ0 can be any number that satisfies $\mu0 > \mu1$.

μ1 (Actual Mean)

Enter a value (or range of values) for the actual mean at which power and sample size are calculated. Care should be taken that this value is consistent with whether higher means are better or worse. $\mu1 > \mu0$ when higher means are better. $\mu1 < \mu0$ when higher means are worse.

Effect Size – Standard Deviation

σ (Standard Deviation)

This option specifies one or more values of the standard deviation. This must be a positive value. Be sure to use the standard deviation of *X* and not the standard deviation of the mean (the standard error). If you are doing a paired test, this is the standard deviation of the differences.

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.

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 (μ_R) 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% (M_{NI} = 0.000115 gm/cm such that $\mu_0 = \mu_R - M_{NI}$ = 0.002185), 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 such that $\mu_0 = \mu_R - M_{NI}$ = 0.0022425).

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 **Wilcoxon Signed-Rank Tests for Non-Inferiority** procedure window by expanding **Means**, then **One Mean**, then clicking on **Non-Inferiority**, and then clicking on **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**.

Option	Value
Design Tab	
Solve For	**Power**
Higher Means Are	**Better (H1: μ > μ0)**
Data Distribution	**Normal**
Population Size	**Infinite**
Alpha	**0.025**
N (Sample Size)	**20 40 60 80 100 150 200 300**
μ0 (Non-Inferiority Mean)	**21.85 22.425**
μ1 (Actual Mean)	**23**
σ (Standard Deviation)	**3**

Annotated Output

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

Numeric Results

Higher Means are Better
Hypotheses: H0: μ ≤ μ0 vs. H1: μ > μ0
Data Distribution: Normal

Power	N	Non-Inferiority Mean μ0	Actual Mean μ1	Standard Deviation σ	Alpha	Beta
0.35274	20	21.9	23.0	3.0	0.025	0.64726
0.63360	40	21.9	23.0	3.0	0.025	0.36640
0.81170	60	21.9	23.0	3.0	0.025	0.18830
0.90968	80	21.9	23.0	3.0	0.025	0.09032
0.95888	100	21.9	23.0	3.0	0.025	0.04112
0.99524	150	21.9	23.0	3.0	0.025	0.00476
0.99951	200	21.9	23.0	3.0	0.025	0.00049
1.00000	300	21.9	23.0	3.0	0.025	0.00000

(report continues)

Wilcoxon Signed-Rank Tests for Non-Inferiority

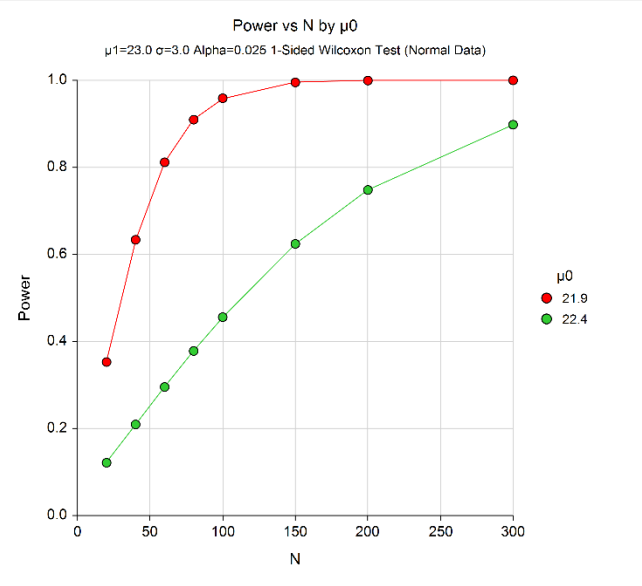
Report Definitions

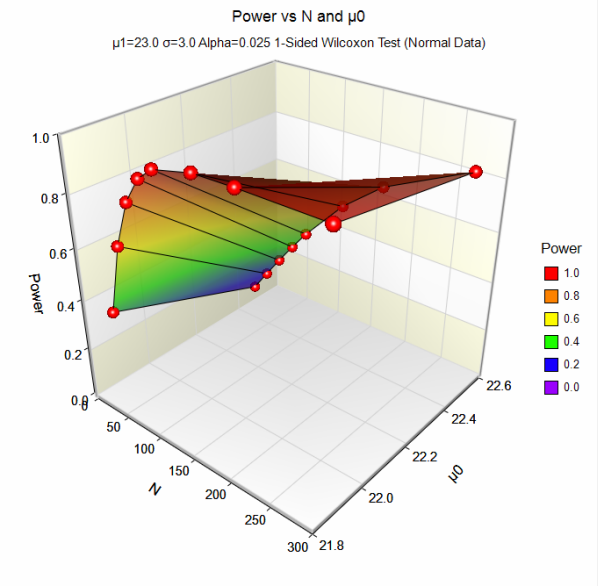
Power is the probability of rejecting the null hypothesis when it is false. It should be close to one.
N is the sample size, the number of subjects (or pairs) in the study.
μ0 = μR - NIM is the non-inferiority mean since higher means are better, where μR is the baseline, standard, or reference mean and NIM is the margin of non-inferiority.
μ1 is the actual value of the population mean at which power and sample size are calculated.
σ is the standard deviation of the response (or standard deviation of differences for paired data). It measures the variability in the population.
Alpha is the probability of rejecting the null hypothesis when it is true, which is the probability of a false positive.
Beta is the probability of accepting the null hypothesis when it is false, which is the probability of a false negative.

Summary Statements

A sample size of 20 achieves 35% power to detect non-inferiority using a one-sided Wilcoxon Signed-Rank test assuming that the actual data distribution is normal when the non-inferiority mean is 21.9 and the actual mean is 23.0. The data are drawn from a single population with an estimated standard deviation of 3.0. The significance level (alpha) of the test is 0.025.

Chart Section





The above report shows that for μ_0 = 21.85 (NIM = 1.15), the sample size necessary to obtain 90% power is just under 80. However, if μ_0 = 22.425 (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 **Wilcoxon Signed-Rank Tests for Non-Inferiority** procedure window by expanding **Means**, then **One Mean**, then clicking on **Non-Inferiority**, and then clicking on **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**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Higher Means Are	**Better (H1: μ > μ0)**
Data Distribution	**Normal**
Population Size	**Infinite**
Power	**0.90**
Alpha	**0.025**
μ0 (Non-Inferiority Mean)	**21.85 22.425**
μ1 (Actual Mean)	**23**
σ (Standard Deviation)	**3**

Output

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

Numeric Results

Numeric Results

Higher Means are Better
Hypotheses: H0: μ ≤ μ0 vs. H1: μ > μ0
Data Distribution: Normal

Power	N	Non-Inferiority Mean μ0	Actual Mean μ1	Standard Deviation σ	Alpha	Beta
0.90215	78	21.9	23.0	3.0	0.025	0.09785
0.90005	302	22.4	23.0	3.0	0.025	0.09995

This report shows the exact sample size requirement for each value of μ_0.

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 *t*-test where μ_R = 1.5 and M_{NI} = 0.5. Their example obtains a sample size of 8 when μ_1 = 2, μ_0 = 1, σ = 1, Alpha = 0.05, and Power = 0.80.

The Wilcoxon Signed-Rank test power calculations are the same as the one-sample *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 × π/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 **Wilcoxon Signed-Rank Tests for Non-Inferiority** procedure window by expanding **Means**, then **One Mean**, then clicking on **Non-Inferiority**, and then clicking on **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**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Higher Means Are	**Better (H1: μ > μ0)**
Data Distribution	**Uniform**
Population Size	**Infinite**
Power	**0.80**
Alpha	**0.05**
μ0 (Non-Inferiority Mean)	**1**
μ1 (Actual Mean)	**2**
σ (Standard Deviation)	**1**

Output

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

Numeric Results

Higher Means are Better
Hypotheses: H0: μ ≤ μ0 vs. H1: μ > μ0
Data Distribution: Uniform

Power	N	Non-Inferiority Mean μ0	Actual Mean μ1	Standard Deviation σ	Alpha	Beta
0.81502	8	1.0	2.0	1.0	0.050	0.18498

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

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: H0: μ ≤ μ0 vs. H1: μ > μ0
Data Distribution: Normal

Power	N	Non-Inferiority Mean μ0	Actual Mean μ1	Standard Deviation σ	Alpha	Beta
0.81502	9	1.0	2.0	1.0	0.050	0.18498

The sample size of 9 matches the expected result.