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.

The details of sample size calculation for the paired design are presented in the Paired T-Tests 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 a single mean are presented in Chow et al. (2018) page 43.

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.

Parameter	PASS Input/Output	Interpretation
δ	δ	*Population mean paired difference*. This is the mean of paired differences in the population. This parameter will be estimated by the study.
δ_1	$\delta 1$	*Actual paired difference at which power is calculated*. This is the value of the mean paired difference at which power is calculated.
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 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 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

$H_0: \delta \leq -|M_{NI}|$ 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 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

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

Paired 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 *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.

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:

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

where *n* is the sample size, *N* is the population size, σ is the original standard deviation, and σ' 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*.

The Standard Deviation of Paired Differences (σ)

If you have results from a previous (or pilot) study, use the estimate of the standard deviation of paired differences, σ, from the study. Another reasonable (but somewhat rough) estimate of σ may be obtained using the range of paired differences as

$$\sigma = \frac{Range}{4}$$

If you have estimates of the expected standard deviations of the paired variables (σ_1 and σ_2) and the Pearson correlation between the paired variables (ρ), the standard deviation of paired differences (σ) may be calculated using the equation

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$$

such that

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}\,.$$

If $\sigma_1 = \sigma_2 = \sigma_x$, then this formula reduces to

$$\sigma^2 = 2\sigma_x^2(1-\rho)$$

such that

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

If you have an estimate of the within-subject population standard deviation (σ_w), then σ may be calculated using the equation

$$\sigma^2 = 2\sigma_w^2$$

such that

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

σ_w is often estimated by the square root of the within mean square error (WMSE) from a repeated measures ANOVA.

Power Calculation for the Paired 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.

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

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	**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**
δ1 (Mean of Paired Differences)	**0**
Standard Deviation Input Type	**Enter the SD of Paired Differences**
σ (SD of Paired Differences)	**3**

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
Data Distribution: Normal

		Paired Differences				
Power	Sample Size N	Non-Inferiority Margin -NIM	Mean δ1	Standard Deviation σ	Alpha	Beta
0.12134	20	-0.575	0	3	0.025	0.87866
0.20927	40	-0.575	0	3	0.025	0.79073
0.29540	60	-0.575	0	3	0.025	0.70460
0.37811	80	-0.575	0	3	0.025	0.62189
0.45584	100	-0.575	0	3	0.025	0.54416
0.62419	150	-0.575	0	3	0.025	0.37581
0.74810	200	-0.575	0	3	0.025	0.25190
0.89804	300	-0.575	0	3	0.025	0.10196
0.35274	20	-1.150	0	3	0.025	0.64726
0.63360	40	-1.150	0	3	0.025	0.36640
0.81170	60	-1.150	0	3	0.025	0.18830
0.90968	80	-1.150	0	3	0.025	0.09032
0.95888	100	-1.150	0	3	0.025	0.04112
0.99524	150	-1.150	0	3	0.025	0.00476
0.99951	200	-1.150	0	3	0.025	0.00049
1.00000	300	-1.150	0	3	0.025	0.00000

Power The probability of rejecting a false null hypothesis when the alternative hypothesis is true.
N The sample size, the number of subjects (or pairs) in the study.
-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 value that is still considered non-inferior.
δ The population mean of paired differences.
δ1 The value of the mean of paired differences at which power and sample size are calculated.
σ The standard deviation of paired differences for the population.
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 paired design (where higher means are considered to be better) will be used to test whether treatment 1 is non-inferior to treatment 2 by testing whether the paired difference in distributions (δ) is greater than -0.575 (H0: δ ≤ -0.575 versus H1: δ > -0.575). The comparison will be made using a one-sided, paired-difference Wilcoxon Signed-Rank test, with a Type I error rate (α) of 0.025. The underlying data distribution of paired differences is assumed to be Normal. The underlying standard deviation of the paired difference distribution is assumed to be 3. To detect a paired mean difference of 0 with a sample size of 20 pairs, the power is 0.12134.

Dropout-Inflated Sample Size

Dropout Rate	Sample Size N	Dropout-Inflated Enrollment Sample Size N'	Expected Number of Dropouts D
20%	20	25	5
20%	40	50	10
20%	60	75	15
20%	80	100	20
20%	100	125	25
20%	150	188	38
20%	200	250	50
20%	300	375	75

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, 25 subjects should be enrolled to obtain a final sample size of 20 subjects.

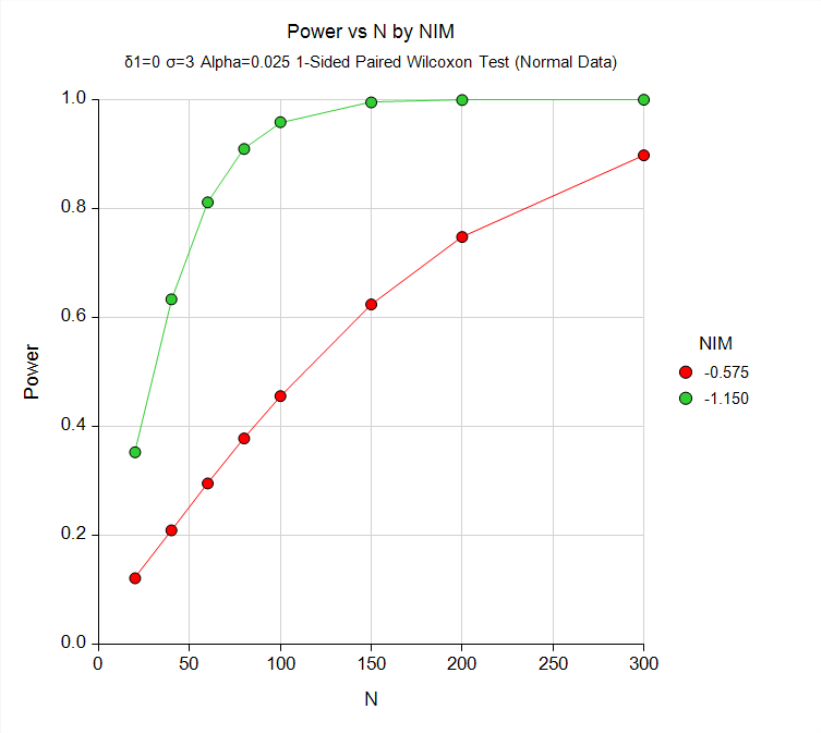
References

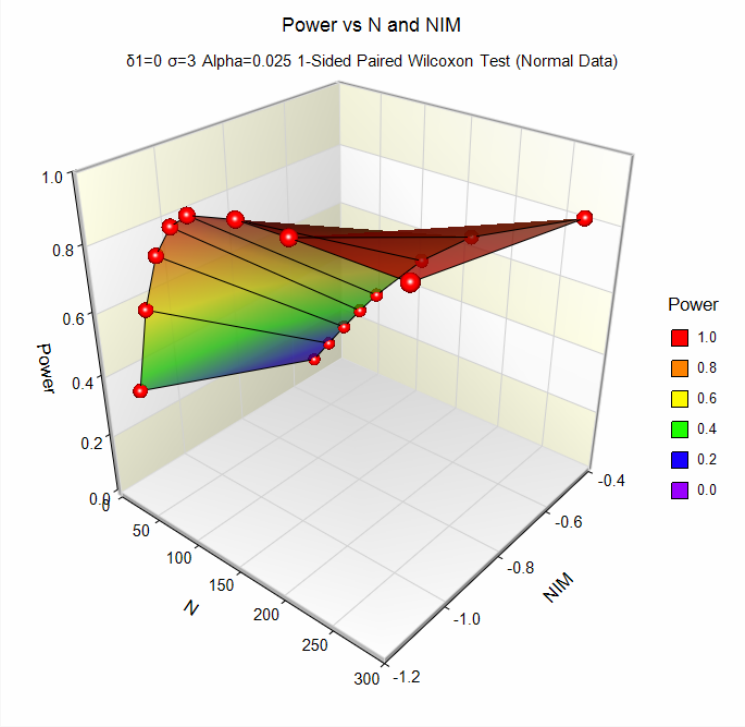
Al-Sunduqchi, Mahdi S. 1990. Determining the Appropriate Sample Size for Inferences Based on the Wilcoxon Statistics. Ph.D. dissertation under the direction of William C. Guenther, Dept. of Statistics, University of Wyoming, Laramie, Wyoming.

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





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

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	**Normal**
Population Size	**Infinite**
Power	**0.90**
Alpha	**0.025**
NIM (Non-Inferiority Margin)	**0.575 1.15**
δ1 (Mean of Paired Differences)	**0**
Standard Deviation Input Type	**Enter the SD of Paired Differences**
σ (SD of Paired Differences)	**3**

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
Data Distribution:	Normal

		Paired Differences				
Power	Sample Size N	Non-Inferiority Margin -NIM	Mean δ1	Standard Deviation σ	Alpha	Beta
0.90005	302	-0.575	0	3	0.025	0.09995
0.90215	78	-1.150	0	3	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 δ1 = 0.5, NIM = 0.5, σ = 1, Alpha = 0.05, and 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 × π/3 = 9 after rounding up.

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 (a and b)** settings files. 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	**Uniform**
Population Size	**Infinite**
Power	**0.80**
Alpha	**0.05**
NIM (Non-Inferiority Margin)	**0.5**
δ1 (Mean of Paired Differences)	**0.5**
Standard Deviation Input Type	**Enter the SD of Paired Differences**
σ (SD of Paired Differences)	**1**

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
Data Distribution:	Uniform

		Paired Differences				
Power	Sample Size N	Non-Inferiority Margin -NIM	Mean δ1	Standard Deviation σ	Alpha	Beta
0.81502	8	-0.5	0.5	1	0.05	0.18498

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

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

Numeric Results

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

		Paired Differences				
Power	Sample Size N	Non-Inferiority Margin -NIM	Mean δ1	Standard Deviation σ	Alpha	Beta
0.81502	9	-0.5	0.5	1	0.05	0.18498

The sample size of 9 matches the expected result.