Chapter 160

Non-Inferiority Tests for the Difference Between Two Correlated Proportions

Introduction

This module computes power and sample size for non-inferiority tests of the difference in which two dichotomous responses are measured on each subject. When one is interested in showing that the true proportions are different, the data are often analyzed with McNemar's test. However, we are interested in showing non-inferiority rather than difference. For example, suppose a diagnostic procedure is accurate, but is expensive to apply or has serious side effects. A replacement procedure is sought which is no less accurate but is less expensive or has fewer side effects. In this case, we are not interested in showing that the two diagnostic procedures are different, but rather that the second is no worse than the first. *Non-inferiority tests* were designed for this situation.

These tests are often divided into two categories: *equivalence* (two-sided) tests and *non-inferiority* (one-sided) tests. Here, the term *equivalence tests* means that we want to show that two diagnostic procedures are equivalent—that is, their accuracy is about the same. This requires a two-sided hypothesis test. On the other hand, *non-inferiority tests* are used when we want to show that a new (experimental) procedure is no worse than the existing (reference or gold-standard) one. This requires a one-sided hypothesis test. The procedures discussed in this chapter deal with the non-inferiority (one-sided) case.

Technical Details

The results of a study in which two dichotomous responses are measured on each subject can be displayed in a 2-by-2 table in which one response is shown across the columns and the other is shown down the rows. In the discussion to follow, the columns of the table represent the standard (reference or control) response and the rows represent the treatment (experimental) response. The outcome probabilities can be classified into the following table.

Experimental Diagnosis	**Standard Diagnosis**		
	Yes	No	**Total**
Yes	p_{11}	p_{10}	P_T
No	p_{01}	p_{00}	$1 - P_T$
Total	P_S	$1 - P_S$	1

In this table, $p_{ij} = p_{Treatment,Standard}$. That is, the first subscript represents the response of the new, experimental procedure while the second subscript represents the response of the standard procedure. Thus, p_{01} represents the proportion having a negative treatment response and a positive standard response.

Sensitivity, Specificity, and Prevalence

To aid in interpretation, analysts have developed a few proportions that summarize the table. Three of the most popular ratios are *sensitivity*, *specificity,* and *prevalence*.

Sensitivity

Sensitivity is the proportion of subjects with a positive standard response who also have a positive experimental response. In terms of proportions from a 2-by-2 table,

$$\text{Sensitivity} = \frac{p_{11}}{(p_{01} + p_{11})} = \frac{p_{11}}{P_S}$$

Specificity

Specificity is the proportion of subjects with a negative standard response who also have a negative experimental response. In terms of proportions from a 2-by-2 table,

$$\text{Specificity} = \frac{p_{00}}{(p_{10} + p_{00})}$$

Prevalence

Prevalence is the overall proportion of individuals with the disease (or feature of interest). In terms of proportions from a 2-by-2 table,

$$\text{Prevalence} = P_S$$

Table Probabilities

The outcome counts from a sample of *n* subjects can be classified into the following table.

Experimental Diagnosis	**Standard Diagnosis**		**Total**
	Yes	No	
Yes	n_{11}	n_{10}	n_T
No	n_{01}	n_{00}	$n - n_T$
Total	n_S	$n - n_S$	n

Note that $n_{11} + n_{00}$ is the number of matches (*concordant pairs*) and $n_{01} + n_{10}$ is the number of *discordant pairs*.

The hypothesis of interest concerns the two marginal probabilities P_T and P_S. P_S represents the accuracy or success of the standard test and P_T represents the accuracy or success of the new, experimental test. Non-inferiority is defined in terms of either the difference of these two proportions, $D = P_T - P_S$, or the relative risk ratio, $R = P_T/P_S$. The choice between D and R will usually lead to different sample sizes to achieve the same power.

Non-Inferiority Hypotheses using Differences

This section is based on Liu, Hsueh, Hsieh and Chen (2002). Refer to that paper for complete details.

If we define M_{NI} as the positive non-inferiority margin, then the null and alternative hypotheses of non-inferiority in terms of the difference are

$$H_0: P_T - P_S \leq -M_{NI} \quad \text{versus} \quad H_1: P_T - P_S > -M_{NI},$$

or equivalently, with $D_0 = -M_{NI}$,

$$H_0: P_T - P_S \leq D_0 \quad \text{versus} \quad H_1: P_T - P_S > D_0.$$

To demonstrate non-inferiority, one desires to reject the null hypothesis and thus conclude that the experimental treatment is not worse than the standard by as much or more than M_{NI}. In the context of the preceding statement and as stated earlier, M_{NI} is defined to be positive. The choice of an appropriate M_{NI} may be difficult. It should be clinically meaningful so that clinicians are willing to conclude that the experimental treatment is acceptable if the difference is no less than M_{NI}. From a statistical perspective, M_{NI} should be less than the effect, if known, of the standard treatment compared to placebo.

Liu et al. (2002) discuss the RMLE-based (score) method for constructing these confidence intervals. This method is based on (developed by, described by) Nam (1997).

Asymptotic Tests

An asymptotic test is given by

$$Z_{NI} = \frac{\widehat{D} + M_{NI}}{\hat{\sigma}} = \frac{c + nM_{NI}}{\sqrt{d - n\widehat{D}^2}} \geq z_{1-\alpha}$$

where

$$\widehat{D} = \frac{n_T}{n} - \frac{n_S}{n} = \frac{n_{10}}{n} - \frac{n_{01}}{n}$$

$$d = n_{10} + n_{01}$$

$$c = n_{10} - n_{01}$$

and $z_{1-\alpha}$ is the standard normal deviate having α in the right tail.

An estimate of $\hat{\sigma}$ based on the RMLE-based (score) procedure of Nam (1997) uses the estimates

$$\tilde{\sigma}_L = \sqrt{\frac{\tilde{p}_{L,10} + \tilde{p}_{L,01} - M_{NI}^2}{n}}$$

where

$$\tilde{p}_{L,10} = \frac{-\tilde{a}_L + \sqrt{\tilde{a}_L^2 - 8\tilde{b}_L}}{4}$$

$$\tilde{p}_{L,01} = \tilde{p}_{L,10} - M_{NI}$$

$$\tilde{a}_{L,01} = -\hat{D}(1 - M_{NI}) - 2(\hat{p}_{01} + M_{NI})$$

$$\tilde{b}_{L,01} = M_{NI}(1 + M_{NI})\hat{p}_{01}$$

Power Formula

The power when the actual difference is D_1 can be evaluated exactly using the multinomial distribution. However, when the sample size is above a user-set level, we use a normal approximation to this distribution which leads to

$$\text{Power} = \begin{cases} 1 - \Phi(c_{NI}) & \text{if } D_1 > -M_{NI} \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} 1 - \Phi(c_{NI}) & \text{if } D_1 > D_0 \\ 0 & \text{otherwise} \end{cases}$$

where

$$c_{NI} = \frac{z_\alpha}{w_L} - \frac{D_1 + M_{NI}}{\sigma} = \frac{z_\alpha}{w_L} - \frac{D_1 - D_0}{\sigma}$$

$$\sigma = \sqrt{\frac{p_{01} + p_{10} - D_1^2}{n}}$$

$$w_L = \sqrt{\frac{2p_{01} + D_1 - D_1^2}{2\bar{p}_{L,01} - M_{NI} - M_{NI}^2}} = \sqrt{\frac{2p_{01} + D_1 - D_1^2}{2\bar{p}_{L,01} + D_0 - D_0^2}}$$

$$\bar{p}_{L,01} = \frac{-a_L + \sqrt{a_L^2 - 8b_L}}{4}$$

$$a_L = -D_1(1 - M_{NI}) - 2(p_{01} + M_{NI}) = -D_1(1 + D_0) - 2(p_{01} - D_0)$$

$$b_L = M_{NI}(1 + M_{NI})p_{01} = -D_0(1 - D_0)p_{01}$$

Nuisance Parameter

The 2-by-2 table includes four parameters, p_{11}, p_{10}, p_{01}, and p_{00}, but the power calculations only require two parameters: P_S and D_1. A third parameter is defined implicitly since the sum of the four parameters is one. Thus, one parameter (known as a nuisance parameter) remains unaccounted for. This parameter must be addressed to fully specify the problem. This fourth parameter can be specified using any one of the following: p_{11}, p_{10}, p_{01}, p_{00}, $p_{10} + p_{01}$, $p_{11} + p_{00}$, the sensitivity of the experimental response, p_{11}/P_S, or the within-subject correlation, ρ.

It may be difficult to specify a reasonable value for the nuisance parameter since its value may not be even approximately known until after the study is conducted. Because of this, we suggest that you calculate power or sample size for a range of values of the nuisance parameter. This will allow you to determine how sensitive the results are to its value.

Estimating P11, P01, and P10 using Pt, Ps, and ρ

Sometimes, obtaining estimates of *P*11, *P*01, and/or *P*10 is problematic. This problem is solved by using the marginal probabilities and the within-subject correlation coefficient, which may be easier to estimate. As outlined in Zhang, Cao, and Ahn (2017), the relationship between *P*11, *Pt*, *Ps* and the correlation is

$$\rho = \frac{P11 - P_s P_t}{\sqrt{P_s P_t (1 - P_s)(1 - P_t)}}$$

Using this relationship, values of ρ can be entered and transformed to the corresponding value of *P*11 using the equation

$$P11 = \rho\sqrt{P_s P_t (1 - P_s)(1 - P_t)} + P_s P_t$$

The only concern is that values of ρ be used that limit *P*11, *P*01, *P*10, and *P*00 to be between 0 and 1. The lower and upper limits of the correlation are

$$\rho_L = \max\left\{-\sqrt{\frac{P_s P_t}{(1 - P_s)(1 - P_t)}}, -\sqrt{\frac{(1 - P_s)(1 - P_t)}{P_s P_t}}\right\}$$

$$\rho_U = \min\left\{\sqrt{\frac{P_s(1 - P_t)}{P_t(1 - P_s)}}, \sqrt{\frac{P_t(1 - P_s)}{P_s(1 - P_t)}}\right\}$$

*P*11, along with *Pt* and *Ps*, can then be used to calculate *P*01 and *P*10.

Example 1 – Finding Power

A clinical trial will be conducted to show that a non-invasive MRI test is not inferior to the invasive CTAP reference test. Historical data suggest that the CTAP test is 80% accurate. After careful discussion, the researchers decide that if the MRI test is 75% accurate or better, it will be considered non-inferior. They decide to use a difference test statistic. Thus, the non-inferiority difference is 0.05. They want to study the power for various sample sizes between 20 and 1000 at the 5% significance level.

They use P01 as the nuisance parameter and look at two values: 0.05 and 0.10.

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**
Power Calculation Method	**Normal Approximation**
Alpha	**0.05**
N (Sample Size)	**20 100 200 300 450 600 800 1000**
D0 (Non-Inferiority Difference)	**-0.05**
D1 (Actual Difference)	**0.0**
Ps (Standard Proportion)	**0.80**
Nuisance Parameter Type	**P01 (% -Trt +Std)**
Nuisance Parameter Value	**0.05 0.10**

Output

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

Numeric Reports

Numeric Results

Solve For: Power
Hypotheses: H0: Pt - Ps ≤ D0 vs. H1: Pt - Ps > D0

		Differences		Proportions			
Power*	Sample Size N	Non-Inferiority D0	Actual D1	Treatment Pt	Standard Ps	Nuisance Parameter P01	Alpha
0.14284	20	-0.05	0	0.8	0.8	0.05	0.05
0.12028	20	-0.05	0	0.8	0.8	0.10	0.05
0.42323	100	-0.05	0	0.8	0.8	0.05	0.05
0.28926	100	-0.05	0	0.8	0.8	0.10	0.05
0.67771	200	-0.05	0	0.8	0.8	0.05	0.05
0.46318	200	-0.05	0	0.8	0.8	0.10	0.05
0.83244	300	-0.05	0	0.8	0.8	0.05	0.05
0.60369	300	-0.05	0	0.8	0.8	0.10	0.05
0.94287	450	-0.05	0	0.8	0.8	0.05	0.05
0.75745	450	-0.05	0	0.8	0.8	0.10	0.05
0.98206	600	-0.05	0	0.8	0.8	0.05	0.05
0.85657	600	-0.05	0	0.8	0.8	0.10	0.05
0.99651	800	-0.05	0	0.8	0.8	0.05	0.05
0.93172	800	-0.05	0	0.8	0.8	0.10	0.05
0.99937	1000	-0.05	0	0.8	0.8	0.05	0.05
0.96870	1000	-0.05	0	0.8	0.8	0.10	0.05

* Power was computed using the normal approximation method.

Power	The probability of rejecting a false null hypothesis when the alternative hypothesis is true.
N	The number of subjects, the sample size.
D0	The difference at which the treatment will be considered non-inferior to the standard.
D1	The actual difference at which the power is calculated. D1 = Pt - Ps.
Pt	The response proportion in the treatment (experimental) group.
Ps	The response proportion in the standard (baseline, reference, or control) group.
Nuisance Parameter	A value that is needed but is not a direct part of the hypotheses.
Alpha	The probability of rejecting a true null hypothesis.

Summary Statements

A paired design will be used to test whether the treatment proportion (Pt) is non-inferior to the standard proportion (Ps), with a non-inferiority difference bound of -0.05 (H0: Pt - Ps ≤ -0.05 versus H1: Pt - Ps > -0.05). The comparison will be made using an RMLE-based score test, with a Type I error rate (α) of 0.05. The nuisance parameter (P01) is assumed to be 0.05. To detect a difference between the proportions of 0 (Pt = 0.8, Ps = 0.8) with a sample size of 20 pairs, the power is 0.14284.

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%	100	125	25
20%	200	250	50
20%	300	375	75
20%	450	563	113
20%	600	750	150
20%	800	1000	200
20%	1000	1250	250

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

Lewis, J.A. 1999. 'Statistical principles for clinical trials (ICH E9) an introductory note on an international guideline.' Statistics in Medicine, 18, pages 1903-1942.

Liu, J., Hsueh, H., Hsieh, E., and Chen, J.J. 2002. 'Tests for equivalence or non-inferiority for paired binary data', Statistics in Medicine, Volume 21, pages 231-245.

Nam, Jun-mo. 1997. 'Establishing equivalence of two treatments and sample size requirements in matched-pairs design', Biometrics, Volume 53, pages 1422-1430.

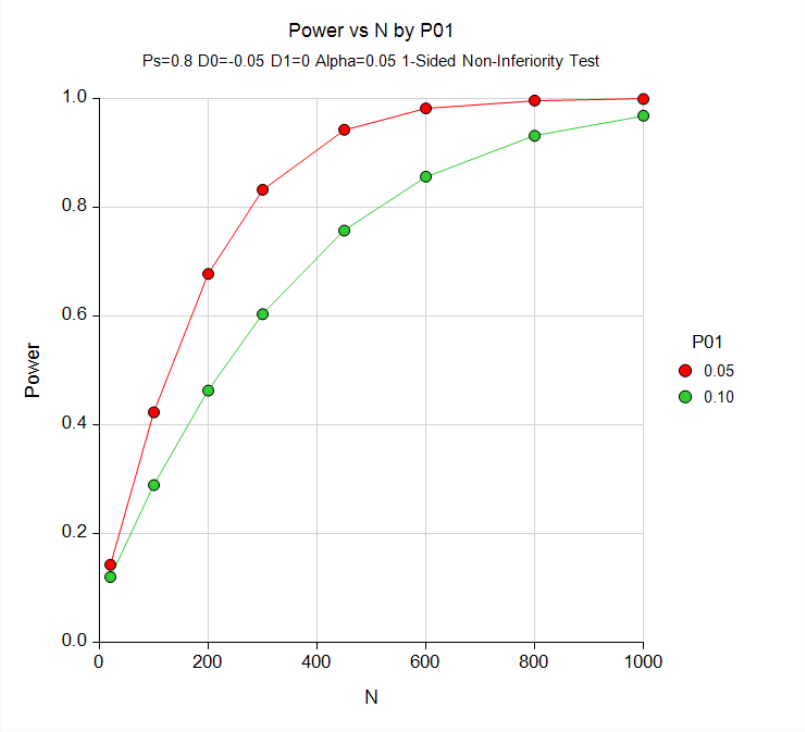
Nam, Jun-mo and Blackwelder, W.C. 2002. 'Analysis of the ratio of marginal probabilities in a matched-pair setting', Statistics in Medicine, Volume 21, pages 689-699.

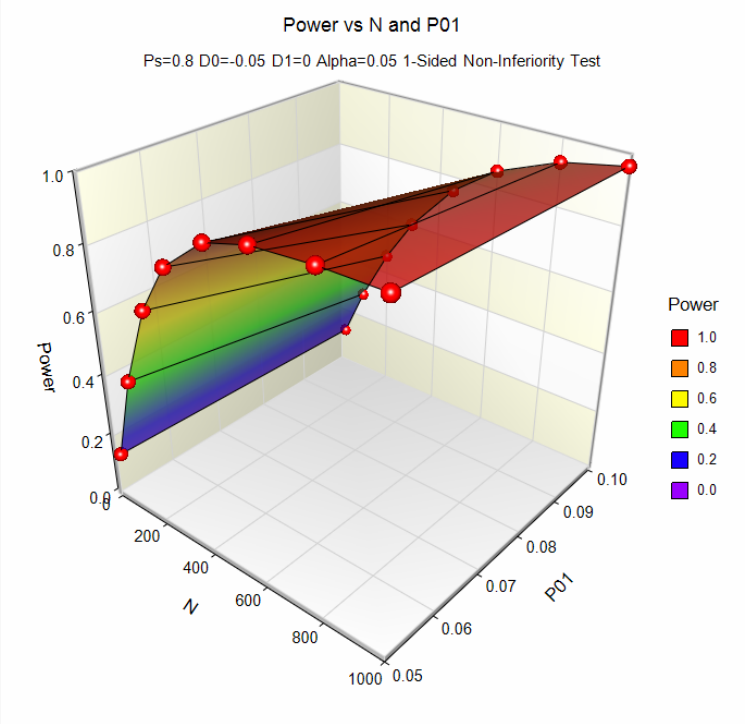
Zhang, S., Cao, J., Ahn, C. 2017. 'Inference and sample size calculation for clinical trials with incomplete observations of paired binary outcomes'. Statistics in Medicine. Volume 36. Pages 581-591.

This report shows the power for the indicated scenarios.

Plots Section

Plots





These plots show the power versus the sample size for the two values of P01. In this example, we see that the value of the nuisance parameter has a large effect on the calculated power.

Example 2 – Finding Sample Size

Continuing with Example 1, the analysts want to determine the exact sample size necessary to achieve 90% power for both values of the nuisance parameter.

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**
Power Calculation Method	**Normal Approximation**
Power	**0.90**
Alpha	**0.05**
D0 (Non-Inferiority Difference)	**-0.05**
D1 (Actual Difference)	**0.0**
Ps (Standard Proportion)	**0.80**
Nuisance Parameter Type	**P01 (% -Trt +Std)**
Nuisance Parameter Value	**0.05 0.10**

Output

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

Numeric Results

Solve For: Sample Size
Hypotheses: H0: Pt - Ps ≤ D0 vs. H1: Pt - Ps > D0

		Differences		Proportions			
Power*	Sample Size N	Non-Inferiority D0	Actual D1	Treatment Pt	Standard Ps	Nuisance Parameter P01	Alpha
0.90026	374	-0.05	0	0.8	0.8	0.05	0.05
0.90014	699	-0.05	0	0.8	0.8	0.10	0.05

* Power was computed using the normal approximation method.

This report shows that the sample size required nearly doubles when P01 is changed from 0.05 to 0.10.

Example 3 – Validation using Liu et al. (2002)

Liu *et al*. (2002) give an example in which P01 is 0.05, P10 is 0.05, D1 is 0.00, D0 is -0.05, the significance level is 0.025, and the power is 80%. From their Table III, the sample size is 350.

In this example, the value of Ps is arbitrary. We set it at 0.50.

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**
Power Calculation Method	**Normal Approximation**
Power	**0.80**
Alpha	**0.025**
D0 (Non-Inferiority Difference)	**-0.05**
D1 (Actual Difference)	**0.0**
Ps (Standard Proportion)	**0.50**
Nuisance Parameter Type	**P01 (% -Trt +Std)**
Nuisance Parameter Value	**0.05**

Output

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

Numeric Results

Solve For: Sample Size
Hypotheses: H0: Pt - Ps ≤ D0 vs. H1: Pt - Ps > D0

		Differences		Proportions			
Power*	Sample Size N	Non-Inferiority D0	Actual D1	Treatment Pt	Standard Ps	Nuisance Parameter P01	Alpha
0.80046	350	-0.05	0	0.5	0.5	0.05	0.025

* Power was computed using the normal approximation method.

The calculated sample size of 350 matches the results of Liu et al. (2002).