Chapter 522

Two-Sample T-Tests for Non-Inferiority Allowing Unequal Variance

Introduction

This procedure computes power and sample size for *non-inferiority* tests in two-sample designs in which the outcome is a continuous normal random variable and the variances of the two groups (populations) are assumed to be unequal. 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 Two-Sample T-Tests Allowing Unequal Variance 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 two means are presented in Chow et al. (2018) pages 50-51.

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

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

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

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

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

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

$H_0: \delta \geq |M_{NI}|$ 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 treatment group standard deviation is 0.000350 gm/cm.

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.*

Two-Sample Unequal-Variance T-Test (Welch's T-Test) Statistic

Welch (1938) proposed the following test when the two variances are not assumed to be equal.

$$t_{df}^{*}=\frac{(\bar{X}_1-\bar{X}_2)-\delta_0}{s_{\bar{X}_1-\bar{X}_2}^{*}}$$

where

$$\bar{X}_k=\frac{\sum_{i=1}^{n_k}X_{ki}}{n_k},$$

$$s_k=\sqrt{\left(\frac{\sum_{i=1}^{n_k}(X_{ki}-\bar{X}_k)^2}{(n_k-1)}\right)},$$

$$s_{\bar{X}_1-\bar{X}_2}^{*}=\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},$$

$$df=\frac{{s_{\bar{X}_1-\bar{X}_2}^{*}}^{4}}{\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_2-1)}},$$

and δ_0 is the value of the difference hypothesized by the null hypothesis which depends on the magnitude and sign of M_{NI}.

The null hypothesis is rejected if the computed p-value is less than a specified level (usually 0.05). Otherwise, no conclusion can be reached.

Computing the Power

When $\sigma_1 \neq \sigma_2$, the power for Welch's unequal-variance *t*-test is calculated as follows.

1. Calculate: $\sigma_{\bar{X}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
2. Calculate: $df = \frac{\sigma_{\bar{X}}^4}{\frac{\sigma_1^4}{n_1^2(n_1-1)} + \frac{\sigma_2^4}{n_2^2(n_2-1)}}$

   which is the adjusted degrees of freedom.
3. 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.
4. Calculate the noncentrality parameter: $\lambda = \frac{\delta - \delta_0}{\sigma_{\bar{X}}}$.
5. 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, 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 treatment group standard deviation is 0.000350 gm/cm.

Following accepted procedure, the analysis will be a non-inferiority test using the *t*-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)**
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**
σ1 (Standard Deviation of Group 1)	**3**
σ2 (Standard Deviation of Group 2)	**3.5**

Output

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

Numeric Results

Solve For:	Power
Test Type:	Two-Sample Welch's Unequal-Variance T-Test
Difference:	δ = μ1 - μ2 = μT - μR
Higher Means Are:	Better
Hypotheses:	H0: δ ≤ -NIM vs. H1: δ > -NIM

	Sample Size			Non-Inferiority Margin	Mean Difference	Standard Deviation		
Power	N1	N2	N	-NIM	δ	σ1	σ2	Alpha
0.05631	10	10	20	-0.575	0	3	3.5	0.025
0.13857	50	50	100	-0.575	0	3	3.5	0.025
0.23613	100	100	200	-0.575	0	3	3.5	0.025
0.42062	200	200	400	-0.575	0	3	3.5	0.025
0.57807	300	300	600	-0.575	0	3	3.5	0.025
0.79641	500	500	1000	-0.575	0	3	3.5	0.025
0.86323	600	600	1200	-0.575	0	3	3.5	0.025
0.94149	800	800	1600	-0.575	0	3	3.5	0.025
0.11250	10	10	20	-1.150	0	3	3.5	0.025
0.41541	50	50	100	-1.150	0	3	3.5	0.025
0.69928	100	100	200	-1.150	0	3	3.5	0.025
0.94054	200	200	400	-1.150	0	3	3.5	0.025
0.99071	300	300	600	-1.150	0	3	3.5	0.025
0.99985	500	500	1000	-1.150	0	3	3.5	0.025
0.99998	600	600	1200	-1.150	0	3	3.5	0.025
1.00000	800	800	1600	-1.150	0	3	3.5	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.
σ1, σ2 The assumed standard deviations for groups 1 and 2, respectively.
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) mean (μ1) is non-inferior to the Group 2 (reference) mean (μ2), 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 unequal-variance t-test, with a Type I error rate (α) of 0.025. The standard deviation for Group 1 is assumed to be 3 and the standard deviation for Group 2 is assumed to be 3.5. To detect a difference in means of 0, with sample sizes of 10 in Group 1 and 10 in Group 2, the power is 0.05631.

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

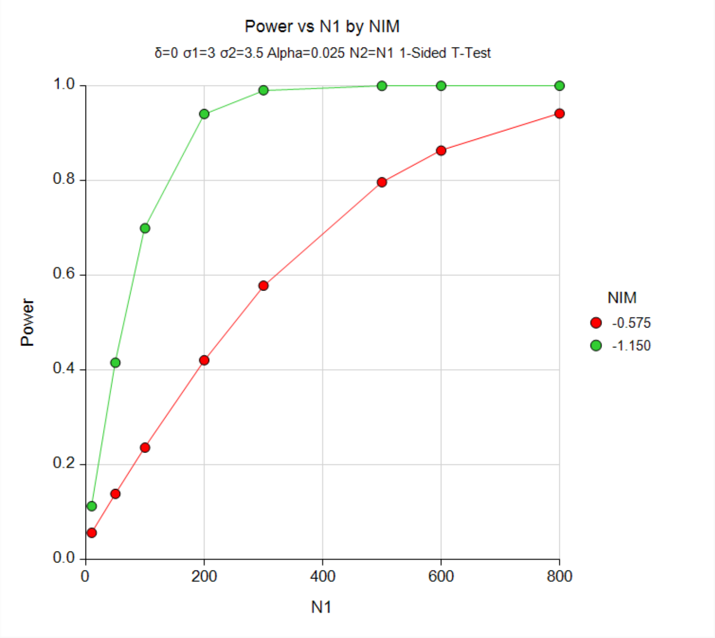
Anticipating a 20% dropout rate, 13 subjects should be enrolled in Group 1, and 13 in Group 2, to obtain final group sample sizes of 10 and 10, respectively.

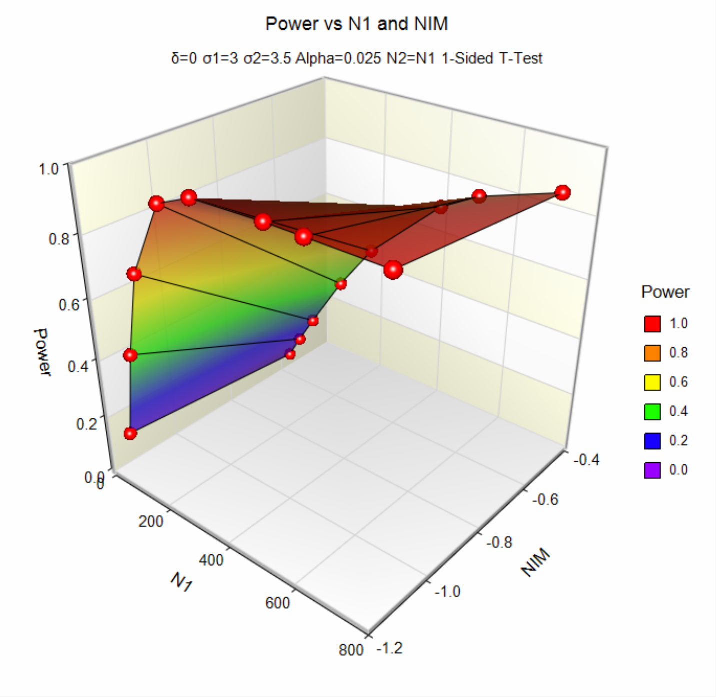
References

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 about 170 per group. However, if NIM = 0.575, the required sample size is about 675 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)**
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**
σ1 (Standard Deviation of Group 1)............... **3**
σ2 (Standard Deviation of Group 2)............... **3.5**

Output

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

Numeric Results

Solve For: Sample Size
Test Type: Two-Sample Welch's Unequal-Variance T-Test
Difference: δ = μ1 - μ2 = μT - μR
Higher Means Are: Better
Hypotheses: H0: δ ≤ -NIM vs. H1: δ > -NIM

Power		Sample Size			Non-Inferiority Margin	Mean Difference	Standard Deviation		
Target	Actual	N1	N2	N	-NIM	δ	σ1	σ2	Alpha
0.9	0.90018	676	676	1352	-0.575	0	3	3.5	0.025
0.9	0.90030	170	170	340	-1.150	0	3	3.5	0.025

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

Example 3 – Validation of Sample Size using Simulation

Example 5 of the Tests for Two Means (Simulation) procedure documentation presents a sample size calculation for a non-inferiority test of

$$H_0: \mu_1 - \mu_2 \geq 4 \quad \text{versus} \quad H_1: \mu_1 - \mu_2 < 4$$

or equivalently

$$H_0: \delta \geq 4 \quad \text{versus} \quad H_1: \delta < 4.$$

This example demonstrates an equal-variance *t*-test with standard deviation in both groups equal to 6, a NIM of 4, and an actual difference of 0. If we modify this example slightly such that the standard deviation in group 1 is 6 and the standard deviation in group 2 is 9, and we use the Welch's unequal-variance *t*-test (see Tests for Two Means (Simulation) – Example 8), then the sample size result is 78 per group with 10,000 simulations.

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	**Worse (H1: δ < NIM)**
Power	**0.90**
Alpha	**0.025**
Group Allocation	**Equal (N1 = N2)**
NIM (Non-Inferiority Margin)	**4**
δ (Actual Difference to Detect)	**0**
σ1 (Standard Deviation of Group 1)	**6**
σ2 (Standard Deviation of Group 2)	**9**

Output

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

```
Numeric Results
──────────────────────────────────────────────────────────────────────────────────────
Solve For:           Sample Size
Test Type:           Two-Sample Welch's Unequal-Variance T-Test
Difference:          δ = μ1 - μ2 = μT - μR
Higher Means Are:    Worse
Hypotheses:          H0: δ ≥ NIM   vs.   H1: δ < NIM
──────────────────────────────────────────────────────────────────────────────────────
                                                  Non-                   Standard
       Power                Sample Size       Inferiority       Mean     Deviation
───────────────────     ───────────────────       Margin   Difference   ──────────
Target       Actual      N1      N2      N         NIM            δ     σ1     σ2     Alpha
──────────────────────────────────────────────────────────────────────────────────────
0.9         0.90018      78      78    156           4            0      6      9     0.025
──────────────────────────────────────────────────────────────────────────────────────
```

The sample size of 78 per group matches the simulation result.