

Chapter 340

Multi-Arm Non-Inferiority Tests for the Difference Between Treatment and Control Means Allowing Unequal Variance

Introduction

This module computes power and sample size for multiple non-inferiority tests of treatment means versus a control mean when no assumption of equal variances for the group populations is made. This is commonly known as the Aspin-Welch test, Welch's t-test (Welch, 1937), or the Satterthwaite method. Sample size formulas for non-inferiority tests of two means are presented in Chow et al. (2018) pages 50-51.

In this design, there are k treatment groups and one control group. A mean is measured in each group. A total of k hypothesis tests are anticipated, each comparing a treatment group with the common control group using a t-test of the difference between two means.

The Bonferroni adjustment of the type I error rate may be optionally made because several comparisons are being tested using the same data. Making a multiplicity adjustment is usually recommended, but not always. In fact, Saville (1990) advocates not applying it and Machin, Campbell, Tan, and Tan (2018) include omitting it as a possibility.

Background

Whether you want to test several doses of a single treatment or several types of treatments, good research practice requires that each treatment be compared with a control. For example, a popular three-arm design consists of three groups: control, treatment A, and treatment B. Two tests are run: treatment A versus control and treatment B versus the same control. This design avoids having to obtain a second control group for treatment B. Besides the obvious efficiency in subjects, it may be easier to recruit subjects if their chances of receiving the new treatment are better than 50-50.

Technical Details

Suppose you want to compare k treatment groups with means μ_i and sample sizes N_i and one control group with mean μ_c and sample size N_c . The total sample size is $N = N_1 + N_2 + \cdots + N_k + N_c$.

Non-Inferiority Tests

A *non-inferiority test* tests that the treatment mean is not worse than the control mean by more than the small, non-inferiority margin called NIM . The actual direction of the hypothesis depends on the response variable being studied.

In the following sections, define $\delta_i = \mu_i - \mu_c$.

Case 1: High Values Better

If 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 control mean. In this case $NIM < 0$. The value of δ at which power is calculated is often set to zero.

The null and alternative hypotheses are

$$H_{0i}: \mu_i - \mu_c \leq NIM \quad \text{vs.} \quad H_{1i}: \mu_i - \mu_c > NIM$$

$$H_{0i}: \mu_i \leq \mu_c + NIM \quad \text{vs.} \quad H_{1i}: \mu_i > \mu_c + NIM$$

$$H_{0i}: \delta_i \leq NIM \quad \text{vs.} \quad H_{1i}: \delta_i > NIM$$

Case 2: High Values Worse

If 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 control mean. In this case $NIM > 0$. The value of δ at which power is calculated is often set to zero.

The null and alternative hypotheses are

$$H_{0i}: \mu_i - \mu_c \geq NIM \quad \text{vs.} \quad H_{1i}: \mu_i - \mu_c < NIM$$

$$H_{0i}: \mu_i \geq \mu_c + NIM \quad \text{vs.} \quad H_{1i}: \mu_i < \mu_c + NIM$$

$$H_{0i}: \delta_i \geq NIM \quad \text{vs.} \quad H_{1i}: \delta_i < NIM$$

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

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

A suitable Type I error probability is chosen for the test, the data are collected, and a t-statistic is generated using the formula

$$t = \frac{(\bar{x}_i - \bar{x}_c) - NIM}{\sqrt{\frac{s_i^2}{N_i} + \frac{s_c^2}{N_c}}}$$

where

$$\bar{X}_i = \frac{\sum_{j=1}^{N_i} X_{ij}}{N_i}$$

$$s_i = \frac{\sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i)^2}{N_i - 1}$$

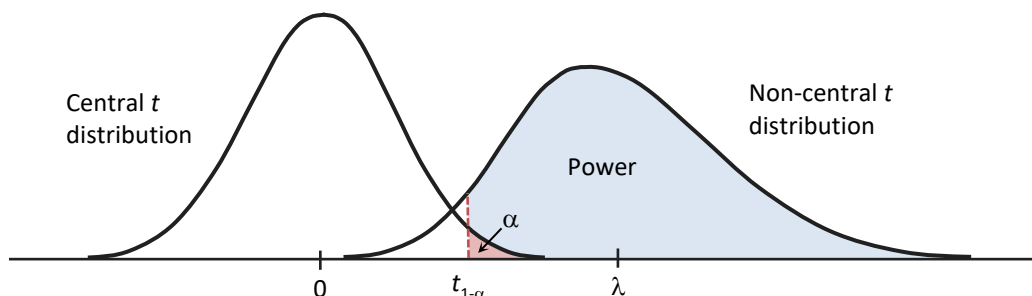
This t-statistic follows a t distribution approximately, with estimated degrees of freedom

$$df' = \frac{\left(\frac{s_i^2}{N_i} + \frac{s_c^2}{N_c}\right)^2}{\frac{1}{N_i - 1} \left(\frac{s_i^2}{N_i}\right)^2 + \frac{1}{N_c - 1} \left(\frac{s_c^2}{N_c}\right)^2}$$

Power Calculation

This section describes the procedure for computing the power from the N_i and N_c , α , the assumed μ_i and μ_c , and the assumed standard deviations, σ_i and σ_c . Two good references for these general methods are Julious (2010) and Chow, Shao, Wang, and Lohknygina (2018), although these texts do not specifically cover the Aspin-Welch-Satterthwaite t-test methods.

The figure below gives a visual representation for the calculation of power for a one-sided test.



Multi-Arm Non-Inferiority Tests for the Difference Between Treatment and Control Means Allowing Unequal Variance

If we call the assumed difference between the means, $\delta_i = \mu_i - \mu_c$, the steps for calculating the power are as follows:

1. Find $t_{1-\alpha}$ based on the central- t distribution with degrees of freedom,

$$df_i = \frac{\left(\frac{\sigma_i^2}{N_i} + \frac{\sigma_c^2}{N_c}\right)^2}{\frac{1}{N_i - 1} \left(\frac{\sigma_i^2}{N_i}\right)^2 + \frac{1}{N_c - 1} \left(\frac{\sigma_c^2}{N_c}\right)^2}.$$

2. Calculate the non-centrality parameter:

$$\lambda_i = \frac{\delta_i}{\sqrt{\frac{\sigma_i^2}{N_i} + \frac{\sigma_c^2}{N_c}}}.$$

3. Calculate the power as the probability that the test statistic t is greater than $t_{1-\alpha}$ under the non-central- t distribution with non-centrality parameter λ_i

$$Power = \Pr_{Non-central-t}(t > t_{1-\alpha} | df_i, \lambda_i).$$

The algorithms for calculating power for the opposite direction are analogous.

When solving for sample size, **PASS** uses this same power calculation formulation, but performs a binary search to determine the sample size.

Multiplicity Adjustment

Because k t -tests between treatment groups and the control group are run when analyzing the results of this study, many statisticians recommend that the Bonferroni adjustment be applied. This adjustment is easy to apply: the value of alpha that is used in the test is found by dividing the original alpha by the number of tests. For example, if the original alpha is set at 0.05 and the number of treatment (not including the control) groups is five, the individual tests will be conducted using an alpha of 0.01.

The main criticism of this procedure is that if there are many tests, the value of alpha becomes very small. To mitigate against this complaint, some statisticians recommend separating the treatment groups into those that are of primary interest and those that are of secondary interest. The Bonferroni adjustment is made by using the number of primary treatments rather than the total number of treatments.

There are some who advocate ignoring the adjustment entirely in the case of randomized clinical trials. See for example Saville (1990) and the discussion in chapter 14 of Machin, Campbell, Tan, and Tan (2018).

Size of the Control Group

Because the control group is used over and over, some advocate increasing the number of subjects in this group. The standard adjustment is to include \sqrt{k} subjects in the control group for each subject in one of the treatment groups. See Machin, Campbell, Tan, and Tan (2018, pages 231-232). Note that usually, the treatment groups all have the same size.

Example 1 – Finding the Sample Size

A parallel-group clinical trial is being designed to compare three treatment therapies against the standard therapy. Higher values of the response are desirable. Suppose the standard therapy has a mean response of 9.3. The investigators would like a sample size large enough to find statistical significance at the 0.025 level when the actual mean responses of the three treatments are all 9.3 and the power of each test is 0.80. The non-inferiority margin will be set to 20% of the means. Since higher values are desirable, the margin will be negative. The result is $NIM = -1.86$.

They want to consider standard deviations of 2.7 in the control group and 2.1 in the three treatment groups. To investigate the sensitivity of the sample sizes to the values of the standard deviations, additional runs with a set of standard deviations 20% higher than those selected and another set 20% lower will be made. These runs are made using the standard deviation multiplier option with $K = 0.8, 1.0, \text{ and } 1.2$.

Following standard procedure, the control group multiplier will be set to $\sqrt{k} = \sqrt{3} = 1.732$ since the control group is used for three comparisons in this design.

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	Sample Size
Higher Means Are	Better (H1: $\delta > NIM$)
Power of Each Test	0.80
Overall Alpha	0.025
Bonferroni Adjustment	Standard Bonferroni
Group Allocation	Enter Group Allocation Pattern, solve for group sample sizes
NIM (Non-Inferiority Margin)	-1.86
Control Mean	9.3
Control Standard Deviation.....	2.7
Control Sample Size Allocation.....	1.732
Set A Number of Groups.....	3
Set A Mean	9.3
Set A Standard Deviation.....	2.1
Set A Sample Size Allocation	1
Set B Number of Groups.....	0
Set C Number of Groups	0
Set D Number of Groups	0
More.....	Unchecked
Add sets of standard deviations with	Checked
different magnitudes, but identical ratio patterns	
K (σ Multiplier)	0.8 1 1.2

Multi-Arm Non-Inferiority Tests for the Difference Between Treatment and Control Means Allowing Unequal Variance

Output

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

Numeric Reports

Numeric Results

Solve For: [Sample Size](#)
 Group Allocation: Enter Group Allocation Pattern, solve for group sample sizes
 Test Type: Unequal-Variance T-Test
 Higher Means Are: Better
 Hypotheses: $H_0: \delta \leq \text{NIM}$ vs. $H_1: \delta > \text{NIM}$
 Number of Groups: 4
 Bonferroni Adjustment: Standard Bonferroni (Divisor = 3)

Comparison	Power		Sample Size		Mean μ_i	Difference δ_i	Non-Inferiority Margin NIM	Standard Deviation		Alpha	
	Target	Actual	N_i	Allocation				Value σ_i	Multiplier K	Overall	Bonferroni-Adjusted
Control			31	1.732	9.3			2.16	0.8		
vs A1	0.8	0.80189	18	1.000	9.3	0	-1.86	1.68	0.8	0.025	0.00833
vs A2	0.8	0.80189	18	1.000	9.3	0	-1.86	1.68	0.8	0.025	0.00833
vs A3	0.8	0.80189	18	1.000	9.3	0	-1.86	1.68	0.8	0.025	0.00833
Total			85								
Control			48	1.732	9.3			2.70	1.0		
vs A1	0.8	0.81033	28	1.000	9.3	0	-1.86	2.10	1.0	0.025	0.00833
vs A2	0.8	0.81033	28	1.000	9.3	0	-1.86	2.10	1.0	0.025	0.00833
vs A3	0.8	0.81033	28	1.000	9.3	0	-1.86	2.10	1.0	0.025	0.00833
Total			132								
Control			68	1.732	9.3			3.24	1.2		
vs A1	0.8	0.80477	39	1.000	9.3	0	-1.86	2.52	1.2	0.025	0.00833
vs A2	0.8	0.80477	39	1.000	9.3	0	-1.86	2.52	1.2	0.025	0.00833
vs A3	0.8	0.80477	39	1.000	9.3	0	-1.86	2.52	1.2	0.025	0.00833
Total			185								

Comparison	The group that is involved in the comparison between the treatment and control displayed on this report line. The comparison is made using the difference.
Target Power	The power desired. Power is probability of rejecting a false null hypothesis for this comparison. This power is of the comparison shown on this line only.
Actual Power	The power actually achieved.
N_i	The number of subjects in the i th group. The total sample size shown below the groups is equal to the sum of all individual group sample sizes.
Allocation	The group sample size allocation ratio of the i th group. The value on each row represents the relative number of subjects assigned to the group.
μ_i	The mean of the i th group at which the power is computed. The first row contains μ_c , the control group mean.
δ_i	The difference between the i th treatment mean and the control mean ($\mu_i - \mu_c$) at which the power is computed.
NIM	The margin of non-inferiority in the scale of the mean difference. $\text{NIM} < 0$.
σ_i	The standard deviation of the responses within this group.
K	The multiplier that was applied to form the group standard deviations shown on this line.
Overall Alpha	The probability of rejecting at least one of the comparisons in this experiment when each null hypothesis is true.
Bonferroni Alpha	The adjusted significance level at which each individual comparison is made.

Multi-Arm Non-Inferiority Tests for the Difference Between Treatment and Control Means Allowing Unequal Variance

Summary Statements

A parallel, 4-group design (with one control group and 3 treatment groups) will be used to test whether the mean for each treatment group is non-inferior to the control group mean, with a non-inferiority margin of -1.86 ($H_0: \delta \leq -1.86$ versus $H_1: \delta > -1.86$, $\delta = \mu_i - \mu_c$). In this study, higher means are considered to be better. The non-inferiority hypotheses will be evaluated using 3 one-sided, two-sample, Bonferroni-adjusted, unequal-variance (Welch's) t-tests, with an overall (experiment-wise) Type I error rate (α) of 0.025. The group standard deviations (beginning with the control group) are assumed to be 2.16, 1.68, 1.68, and 1.68. The control group mean is assumed to be 9.3. To detect the treatment means 9.3, 9.3, and 9.3 with at least 80% power for each test, the control group sample size needed will be 31 and the number of needed subjects for the treatment groups will be 18, 18, and 18 (totaling 85 subjects overall).

Dropout-Inflated Sample Size

Group	Dropout Rate	Sample Size Ni	Dropout- Inflated Enrollment Sample Size Ni'	Expected Number of Dropouts Di
1	20%	31	39	8
2	20%	18	23	5
3	20%	18	23	5
4	20%	18	23	5
Total		85	108	23
1	20%	48	60	12
2	20%	28	35	7
3	20%	28	35	7
4	20%	28	35	7
Total		132	165	33
1	20%	68	85	17
2	20%	39	49	10
3	20%	39	49	10
4	20%	39	49	10
Total		185	232	47

Group Lists the group numbers.

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.

Ni The evaluable sample size for each group at which power is computed (as entered by the user). If Ni subjects are evaluated out of the Ni' subjects that are enrolled in the study, the design will achieve the stated power.

Ni' The number of subjects that should be enrolled in each group in order to obtain Ni evaluable subjects, based on the assumed dropout rate. Ni' is calculated by inflating Ni using the formula $Ni' = Ni / (1 - DR)$, with Ni' 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.)

Di The expected number of dropouts in each group. $Di = Ni' - Ni$.

Dropout Summary Statements

Anticipating a 20% dropout rate, group sizes of 39, 23, 23, and 23 subjects should be enrolled to obtain final group sample sizes of 31, 18, 18, and 18 subjects.

Multi-Arm Non-Inferiority Tests for the Difference Between Treatment and Control Means Allowing Unequal Variance

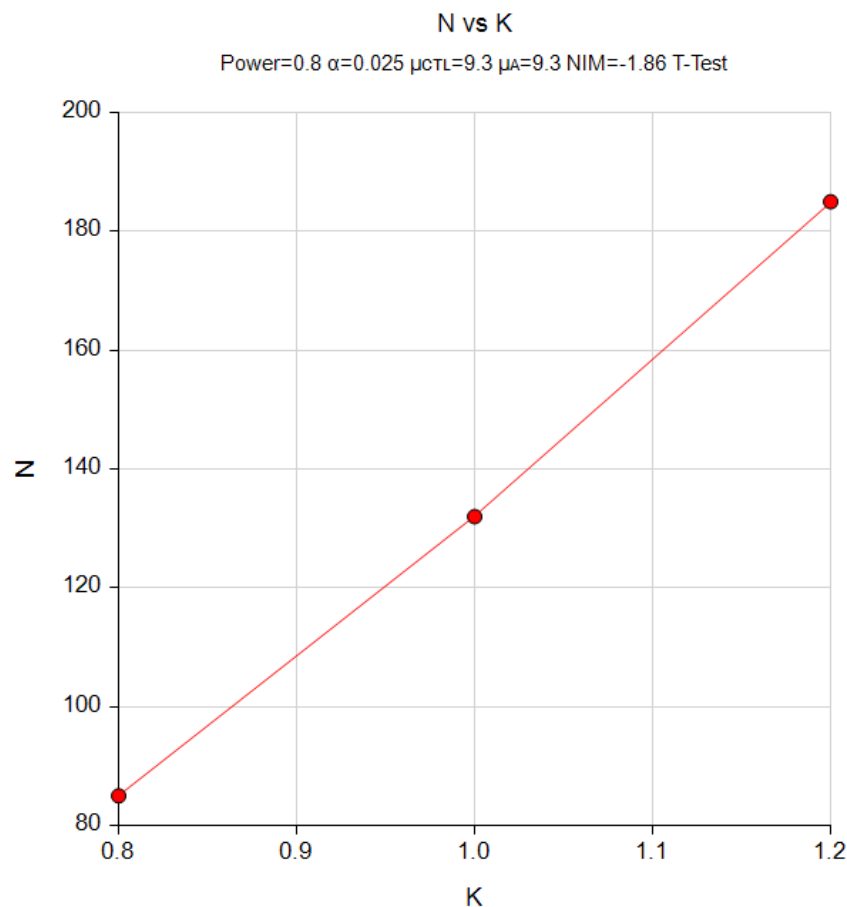
References

- Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. 2018. Sample Size Calculations in Clinical Research, 3rd Edition. Chapman & Hall/CRC. Boca Raton, FL. Pages 86-88.
- Machin, D., Campbell, M.J., Tan, S.B, and Tan, S.H. 2018. Sample Sizes for Clinical, Laboratory, and Epidemiology Studies, 4th Edition. Wiley Blackwell.
- Julious, Steven A. 2004. 'Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data.' *Statistics in Medicine*, 23:1921-1986.
- Welch, B.L. 1938. 'The significance of the difference between two means when the population variances are unequal.' *Biometrika*, 29, 350-362.
- Zar, Jerrold H. 1984. *Biostatistical Analysis* (Second Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

This report shows the numeric results of this power study. Notice that the results are shown in blocks of three rows at a time. Each block represents a single design.

Plots Section

Plots



This plot gives a visual presentation to the results in the Numeric Report. We can quickly see the impact on the sample size of changing the standard deviation magnitude.

Example 2 – Validation using a Previously Validated Procedure

We could not find a validation result in the statistical literature, so we will use a previously validated **PASS** procedure (**Two-Sample T-Tests for Non-Inferiority Allowing Unequal Variance**) to produce the results for the following example.

A parallel-group clinical trial is being designed to compare three treatment therapies against the standard therapy. Higher values of the response are desirable. Suppose the standard therapy has a mean response of 9.3. The investigators would like a sample size large enough to find statistical significance at the 0.025 level when the actual mean responses of the three treatments are all 9.3 and the power of each test is 0.80. The non-inferiority margin will be set to 20% of the means. Since higher values are desirable, the margin will be negative. The result is $NIM = -1.86$.

They want to consider standard deviations of 2.7 in the control group and 2.1 in the three treatment groups. The sample sizes of all groups will be equal.

The **Two-Sample T-Tests for Non-Inferiority Allowing Unequal Variance** procedure is set up as follows.

Design Tab

Solve For **Sample Size**
 Higher Means Are **Better (H1: $\delta > -NIM$)**
 Power **0.8**
 Alpha **0.00833** (which is Alpha / k)
 Group Allocation **Equal (N1 = N2)**
 NIM (Non-Inferiority Margin) **-1.86**
 δ (Actual Difference to Detect) **0**
 σ_1 (Standard Deviation of Group 1) **2.1**
 σ_2 (Standard Deviation of Group 2) **2.7**

This set of options generates the following report.

Numeric Results

Solve For: [Sample Size](#)
 Test Type: Two-Sample Welch's Unequal-Variance T-Test
 Difference: $\delta = \mu_1 - \mu_2 = \mu_T - \mu_R$
 Higher Means Are: Better
 Hypotheses: $H_0: \delta \leq -NIM$ vs. $H_1: \delta > -NIM$

Power		Sample Size			Non-Inferiority Margin -NIM	Mean Difference δ	Standard Deviation		Alpha
Target	Actual	N1	N2	N			σ_1	σ_2	
0.8	0.80049	37	37	74	-1.86	0	2.1	2.7	0.00833

In order to maintain a power of 80% for all three groups, it is apparent that the groups will all need to have a sample size of 37 per group. This table contains the validation values. We will now run these values through the current procedure and compare the results with these values.

Multi-Arm Non-Inferiority Tests for the Difference Between Treatment and Control Means Allowing Unequal Variance

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: $\delta > \text{NIM}$)**
 Power of Each Test **0.80**
 Overall Alpha **0.025**
 Bonferroni Adjustment **Standard Bonferroni**
 Group Allocation **Equal (Nc = N1 = N2 = ...)**
 NIM (Non-Inferiority Margin) **-1.86**
 Control Mean **9.3**
 Control Standard Deviation **2.7**
 Set A Number of Groups **3**
 Set A Mean **9.3**
 Set A Standard Deviation **2.1**
 Set B Number of Groups **0**
 Set C Number of Groups **0**
 Set D Number of Groups **0**
 More **Unchecked**
 Add sets of standard deviations with **Unchecked**
 different magnitudes, but identical
 ratio patterns

Output

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

Numeric Results

Solve For: [Sample Size](#)
 Group Allocation: Equal (Nc = N1 = N2 = ...)
 Test Type: Unequal-Variance T-Test
 Higher Means Are: Better
 Hypotheses: H0: $\delta \leq \text{NIM}$ vs. H1: $\delta > \text{NIM}$
 Number of Groups: 4
 Bonferroni Adjustment: Standard Bonferroni (Divisor = 3)

Comparison	Power		Sample Size Ni	Mean μ_i	Difference δ_i	Non- Inferiority Margin NIM	Standard Deviation σ_i	Alpha	
	Target	Actual						Overall	Bonferroni- Adjusted
Control			37	9.3			2.7		
vs A1	0.8	0.80053	37	9.3	0	-1.86	2.1	0.025	0.00833
vs A2	0.8	0.80053	37	9.3	0	-1.86	2.1	0.025	0.00833
vs A3	0.8	0.80053	37	9.3	0	-1.86	2.1	0.025	0.00833
Total			148						

As you can see, the sample sizes are all 37. This matches the sample size found in the validation run above. The procedure is validated.