

Chapter 330

Multi-Arm Equivalence Tests for the Difference Between Treatment and Control Means Assuming Equal Variance

Introduction

This module computes power and sample size for multiple equivalence tests of treatment means versus a control mean. Schuirmann's (1987) two one-sided tests (TOST) approach is used to test equivalence. Only a brief introduction to the subject will be given here. For a comprehensive discussion, refer to Chow and Liu (1999). The multiplicity is based on the results in Machin, Campbell, Tan, and Tan (2018).

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 an equivalence 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 equivalence 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 $N = N_1 + N_2 + \cdots + N_k + N_C$.

Equivalence Tests

Measurements are made on individuals that have been randomly assigned to the groups. This *parallel-groups* design may be analyzed by a set of TOST equivalence tests to show that the means of the treatment and control groups do not differ by more than a small amount, either positive or negative. To conduct an equivalence test, you must set upper and lower equivalence limits for the difference between a treatment mean and the control mean. These limits, which will be called EL and EU , establish an interval of equivalence. When the sample mean difference falls between these limits, the null hypothesis of non-equivalence is rejected and the equivalence of the two group means is concluded.

The statistical hypotheses are written as follows:

$$H_{0i}: \mu_i - \mu_C \leq EL \text{ or } \mu_i - \mu_C \geq EU \quad \text{vs.} \quad H_{1i}: EL < \mu_i - \mu_C < EU$$

or, if we define $\delta_i = \mu_i - \mu_C$,

$$H_{0i}: \delta_i \leq EL \text{ or } \delta_i \geq EU \quad \text{vs.} \quad H_{1i}: \delta_i < EU$$

where $EL < 0$ and $EU > 0$. Usually, $EL = -EU$.

Two-Sample Equal-Variance T-Test Statistic

Under the null hypothesis, this test assumes that the two groups of data are simple random samples from a single population of normally distributed values that all have the same mean and variance. This assumption implies that the data are continuous, and their distribution is symmetric. The calculation of the two one-sided test statistics uses the following equations.

$$t_L = \frac{\bar{x}_i - \bar{x}_C - EL}{\sqrt{\frac{(N_i - 1)s_i^2 + (N_C - 1)s_C^2}{N_i + N_C - 2} \left(\frac{1}{N_i} + \frac{1}{N_C} \right)}}$$

$$t_U = \frac{\bar{x}_i - \bar{x}_C - EU}{\sqrt{\frac{(N_i - 1)s_i^2 + (N_C - 1)s_C^2}{N_i + N_C - 2} \left(\frac{1}{N_i} + \frac{1}{N_C} \right)}}$$

Multi-Arm Equivalence Tests for the Difference Between Treatment and Control Means Assuming Equal Variance

where

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

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

$$df = N_i + N_c - 2$$

These t -statistics follow t distributions with $N_i + N_c - 2$ degrees of freedom. The null hypothesis is rejected if t_L and $-t_U$ are both greater than or equal to $t_{1-\alpha, df}$.

Power Calculation

The power of the equal-variance equivalence t -test is calculated as

$$\Pr(t_L \geq t_{1-\alpha, df} \text{ and } t_U \leq -t_{1-\alpha, df} | \mu_i, \mu_c, \sigma)$$

where t_L and t_U are distributed as the bivariate, noncentral t distribution with noncentrality parameters Δ_L and Δ_U given by

$$\Delta_L = \frac{\mu_i - \mu_c - EL}{\sigma \sqrt{\frac{1}{N_i} + \frac{1}{N_c}}}$$

$$\Delta_U = \frac{\mu_i - \mu_c - EU}{\sigma \sqrt{\frac{1}{N_i} + \frac{1}{N_c}}}$$

Multiplicity Adjustment

Because k 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 0.05 and the number of treatment (not including the control) groups is five, the individual tests will be conducted using an alpha of $0.05 / 5 = 0.01$.

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 the 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 determine if any or all of three treatment therapies are equivalent to the standard therapy. Suppose the standard therapy has a mean response of 9.3 with a standard deviation of 2.5. The investigators would like a sample size large enough to find statistical significance at the 0.05 level if the actual mean responses of the three treatments are 9.1, 9.3, and 9.6. The power of each test is set to 0.80. The equivalence margin is 10% of 9.3 = 0.93, so the equivalence limits are set to -0.93 and 0.93. The standard deviation ranges from 2.0 to 3.0.

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
Power of Each Test	0.8
Overall Alpha	0.05
Bonferroni Adjustment	Standard Bonferroni
Group Allocation	Enter Group Allocation Pattern, solve for group sample sizes
EU (Upper Equivalence Limit).....	0.93
EL (Lower Equivalence Limit)	-Upper Limit
Control Mean	9.3
Control Sample Size Allocation.....	1.732
Set A Number of Groups.....	1
Set A Mean	9.1
Set A Sample Size Allocation	1
Set B Number of Groups.....	1
Set B Mean	9.3
Set B Sample Size Allocation	1
Set C Number of Groups	1
Set C Mean.....	9.6
Set C Sample Size Allocation	1
Set D Number of Groups	0
More.....	Unchecked
σ (Standard Deviation).....	2 2.5 3

Multi-Arm Equivalence Tests for the Difference Between Treatment and Control Means Assuming Equal 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: T-Test
 Hypotheses: $H_0: \delta \leq EL \text{ or } \delta \geq EU$ vs. $H_1: EL < \delta < EU$
 Number of Groups: 4
 Bonferroni Adjustment: Standard Bonferroni (Divisor = 3)

Comparison	Power		Sample Size		Mean μ_i	Mean Difference δ_i	Equivalence Limits		Standard Deviation σ	Alpha	
	Target	Actual	Ni	Allocation			Lower EL	Upper EU		Overall	Bonferroni-Adjusted
Control			246	1.732	9.3				2.0		
vs A	0.8	0.90678	142	1.000	9.1	-0.2	-0.93	0.93	2.0	0.05	0.01667
vs B	0.8	0.97686	142	1.000	9.3	0.0	-0.93	0.93	2.0	0.05	0.01667
vs C	0.8	0.80279	142	1.000	9.6	0.3	-0.93	0.93	2.0	0.05	0.01667
Total			672								
Control			381	1.732	9.3				2.5		
vs A	0.8	0.90486	220	1.000	9.1	-0.2	-0.93	0.93	2.5	0.05	0.01667
vs B	0.8	0.97598	220	1.000	9.3	0.0	-0.93	0.93	2.5	0.05	0.01667
vs C	0.8	0.80008	220	1.000	9.6	0.3	-0.93	0.93	2.5	0.05	0.01667
Total			1041								
Control			547	1.732	9.3				3.0		
vs A	0.8	0.90499	316	1.000	9.1	-0.2	-0.93	0.93	3.0	0.05	0.01667
vs B	0.8	0.97598	316	1.000	9.3	0.0	-0.93	0.93	3.0	0.05	0.01667
vs C	0.8	0.80040	316	1.000	9.6	0.3	-0.93	0.93	3.0	0.05	0.01667
Total			1495								

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.
δ	This is the difference of the means given by $\mu_i - \mu_c$.
H_0	The null hypothesis is $H_0: \delta \leq EL \text{ or } \delta \geq EU$.
H_1	The alternative hypothesis is $H_1: EL < \delta < EU$.
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.
Ni	The number of subjects in the ith 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 ith group. The value on each row represents the relative number of subjects assigned to the group.
μ_i	The mean of the ith group at which the power is computed. The first row contains μ_c , the control group mean.
δ_i	The difference between the ith treatment mean and the control mean ($\mu_i - \mu_c$) at which the power is computed.
EL	The lower equivalence limit for the difference. This is the smallest negative mean difference between each treatment group and the control group that still results in the conclusion that the treatment group is equivalent to the control group.
EU	The upper equivalence limit for the difference. This is the largest positive difference mean difference between each treatment group and the control group that still results in the conclusion that the treatment group is equivalent to the control group.
σ	The standard deviation of the responses within each group.
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 Equivalence Tests for the Difference Between Treatment and Control Means Assuming Equal 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 equivalent to the control group mean, with equivalence difference bounds of -0.93 and 0.93 ($H_0: \delta \leq -0.93$ or $\delta \geq 0.93$ versus $H_1: -0.93 < \delta < 0.93$, $\delta = \mu_i - \mu_c$). Each of the 3 equivalence comparisons will be made using two one-sided, two-sample, Bonferroni-adjusted, equal-variance t-tests. The overall (experiment-wise) Type I error rate (α) is 0.05. The common standard deviation for all groups is assumed to be 2. The control group mean is assumed to be 9.3. To detect the treatment means 9.1, 9.3, and 9.6 with at least 80% power for each test, the control group sample size needed will be 246 and the number of needed subjects for the treatment groups will be 142, 142, and 142 (totaling 672 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%	246	308	62
2	20%	142	178	36
3	20%	142	178	36
4	20%	142	178	36
Total		672	842	170
1	20%	381	477	96
2	20%	220	275	55
3	20%	220	275	55
4	20%	220	275	55
Total		1041	1302	261
1	20%	547	684	137
2	20%	316	395	79
3	20%	316	395	79
4	20%	316	395	79
Total		1495	1869	374

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 308, 178, 178, and 178 subjects should be enrolled to obtain final group sample sizes of 246, 142, 142, and 142 subjects.

Multi-Arm Equivalence Tests for the Difference Between Treatment and Control Means Assuming Equal Variance

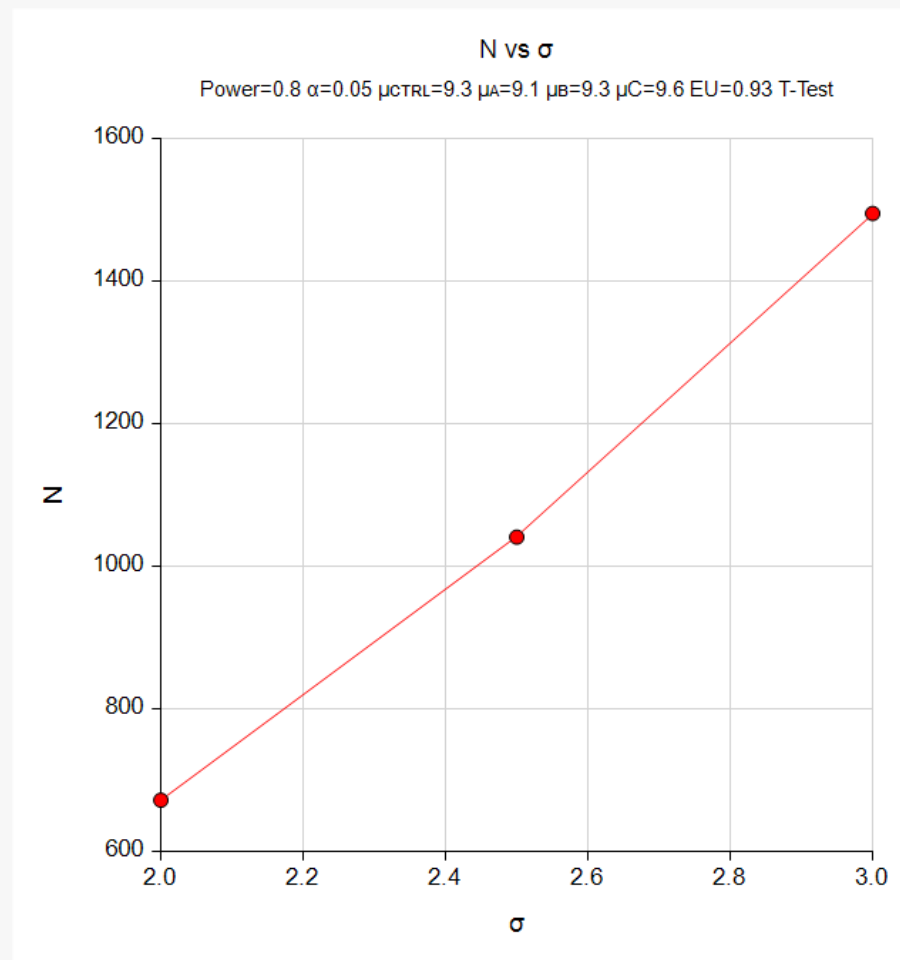
References

- Blackwelder, W.C. 1998. 'Equivalence Trials.' In Encyclopedia of Biostatistics, John Wiley and Sons. New York. Volume 2, 1367-1372.
- 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.
- Julious, Steven A. 2004. 'Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data.' Statistics in Medicine, 23:1921-1986.
- 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.

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 representation of the results in the Numeric Report. We can quickly see the impact on the sample size of varying the standard deviation.

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 Equivalence Assuming Equal Variance**) to produce the results for the following example.

A parallel-group clinical trial is being designed to determine if any or all of three treatment therapies are equivalent to the standard therapy. Suppose the standard therapy has a mean response of 9.3 with a standard deviation of 2.5. The investigators would like a sample size large enough to find statistical significance at the $0.05/3 = 0.01667$ level if the actual mean responses of the three treatments are 9.1, 9.3, and 9.6. The power of each test is set to 0.80. The equivalence margin is 10% of $9.3 = 0.93$, so the equivalence limits are set to -0.93 and 0.93. The standard deviation is set to 2.5.

The sample sizes of all groups will be equal.

The **Two-Sample T-Tests for Equivalence Assuming Equal Variance** procedure is set up as follows.

Design Tab

Solve For **Sample Size**
 Power..... **0.8**
 Alpha..... **0.01667** (which is Alpha / k)
 Group Allocation **Equal (N1 = N2)**
 EU (Upper Equivalence Limit)..... **0.93**
 EL (Lower Equivalence Limit) **-Upper Limit**
 δ (Actual Difference) **-0.2 0 0.3**
 σ (Standard Deviation)..... **2.5**

This set of options generates the following report.

Numeric Results

Solve For: **Sample Size**
 Difference: $\delta = \mu_1 - \mu_2 = \mu_T - \mu_R$
 Hypotheses: $H_0: \delta \leq EL \text{ or } \delta \geq EU$ vs. $H_1: EL < \delta < EU$
 Test Type: **Two One-Sided Equal-Variance T-Tests**

Power		Sample Size			Equivalence Limits		Actual Difference	Standard Deviation	Alpha
Target	Actual	N1	N2	N	Lower EL	Upper EU	δ	σ	
0.8	0.80195	212	212	424	-0.93	0.93	-0.2	2.5	0.01667
0.8	0.80302	170	170	340	-0.93	0.93	0.0	2.5	0.01667
0.8	0.80008	279	279	558	-0.93	0.93	0.3	2.5	0.01667

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 279. 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 Equivalence Tests for the Difference Between Treatment and Control Means Assuming Equal 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
Power of Each Test	0.8
Overall Alpha	0.05
Bonferroni Adjustment	Standard Bonferroni
Group Allocation	Equal (Nc = N1 = N2 = ...)
EU (Upper Equivalence Limit)	0.93
EL (Lower Equivalence Limit)	-Upper Limit
Control Mean	9.3
Set A Number of Groups	1
Set A Mean	9.1
Set B Number of Groups	1
Set B Mean	9.3
Set C Number of Groups	1
Set C Mean	9.6
Set D Number of Groups	0
More	Unchecked
σ (Standard Deviation)	2.5

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:	T-Test
Hypotheses:	H0: $\delta \leq EL$ or $\delta \geq EU$ vs. H1: $EL < \delta < EU$
Number of Groups:	4
Bonferroni Adjustment:	Standard Bonferroni (Divisor = 3)

Comparison	Power		Sample Size Ni	Mean μ_i	Mean Difference δ_i	Equivalence Limits		Standard Deviation σ	Alpha	
	Target	Actual				Lower EL	Upper EU		Overall	Bonferroni- Adjusted
Control			279	9.3				2.5		
vs A	0.8	0.90485	279	9.1	-0.2	-0.93	0.93	2.5	0.05	0.01667
vs B	0.8	0.97597	279	9.3	0.0	-0.93	0.93	2.5	0.05	0.01667
vs C	0.8	0.80006	279	9.6	0.3	-0.93	0.93	2.5	0.05	0.01667
Total			1116							

As you can see, the sample sizes are all 279, which match the largest sample size found in the validation run above. The procedure is validated.