Chapter 342

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

Introduction

This module computes power and sample size for multiple equivalence tests of treatment means versus a control mean when no assumption of equal variances is made. Schuirmann's (1987) two one-sided tests (TOST) approach is used to test equivalence. A modified t-test, known as the Aspin-Welch test, Welch's t-test (Welch, 1937), or the Satterthwaite method, is used for the individual tests. 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 Aspin-Welch equivalence test.

A 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 $N = N_1 + N_2 + \cdots + N_k + N_C$.

Let $\delta = \mu_i - \mu_C$.

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.

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

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

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

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

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 (usually 0.05), the data are collected, and a pair of t-statistics are generated using the formulas

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

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

where

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

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

These *t*-statistics follow a *t* distribution approximately, with modified 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

The power of the unequal-variance equivalence *t*-test procedure is calculated as

$$\Pr\left(t_L^* \geq t_{1-\alpha,df} \text{ and } t_U^* \leq -t_{1-\alpha,df} | \mu_i, \mu_C, \sigma_i, \sigma_C\right)$$

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

$$\Delta_L = \frac{\delta_i - E_L}{\sqrt{\frac{\sigma_i^2}{N_i} + \frac{\sigma_C^2}{N_C}}}$$

$$\Delta_U = \frac{\delta_i - E_U}{\sqrt{\frac{\sigma_i^2}{N_i} + \frac{\sigma_C^2}{N_C}}}$$

and

$$df = \frac{\left(\sqrt{\frac{\sigma_i^2}{N_i} + \frac{\sigma_C^2}{N_C}}\right)^4}{\frac{\sigma_i^4}{N_i^2(N_i - 1)} + \frac{\sigma_C^4}{N_C^2(N_C - 1)}}.$$

When solving for sample size, **PASS** uses this same power calculation formulation, but performs a 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 conduction 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 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 mean response of 9.3. They want to consider standard deviations of 2.7 in the control group and 3.5 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, and 1.2.

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 also 9.3. The power of each test is set to 0.80. The equivalence margin is 20% of 9.3 = 1.86, so the equivalence limits are set to -1.86 and 1.86.

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	**0.80**
Overall Alpha	**0.05**
Bonferroni Adjustment	**Standard Bonferroni**
Group Allocation	**Enter Group Allocation Pattern, solve for group sample sizes**
EU (Upper Equivalence Limit)	**1.86**
EL (Lower Equivalence Limit)	**-Upper Limit**
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	**3.5**
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 different magnitudes, but identical ratio patterns	**Checked**
K (σ Multiplier)	**0.8 1 1.2**

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
Hypotheses:	H0: δ ≤ EL or δ ≥ EU vs. H1: EL < δ < EU
Number of Groups:	4
Bonferroni Adjustment:	Standard Bonferroni (Divisor = 3)

	Power		Sample Size				Equivalence Limits		Standard Deviation		Alpha	
Comparison	Target	Actual	Ni	Allocation	Mean μi	Difference δi	Lower EL	Upper EU	Value σi	Multiplier K	Overall	Bonferroni-Adjusted
Control			64	1.732	9.3				2.16	0.8		
vs A1	0.8	0.80311	37	1.000	9.3	0	-1.86	1.86	2.80	0.8	0.05	0.01667
vs A2	0.8	0.80311	37	1.000	9.3	0	-1.86	1.86	2.80	0.8	0.05	0.01667
vs A3	0.8	0.80311	37	1.000	9.3	0	-1.86	1.86	2.80	0.8	0.05	0.01667
Total			175									
Control			99	1.732	9.3				2.70	1.0		
vs A1	0.8	0.80240	57	1.000	9.3	0	-1.86	1.86	3.50	1.0	0.05	0.01667
vs A2	0.8	0.80240	57	1.000	9.3	0	-1.86	1.86	3.50	1.0	0.05	0.01667
vs A3	0.8	0.80240	57	1.000	9.3	0	-1.86	1.86	3.50	1.0	0.05	0.01667
Total			270									
Control			140	1.732	9.3				3.24	1.2		
vs A1	0.8	0.80064	81	1.000	9.3	0	-1.86	1.86	4.20	1.2	0.05	0.01667
vs A2	0.8	0.80064	81	1.000	9.3	0	-1.86	1.86	4.20	1.2	0.05	0.01667
vs A3	0.8	0.80064	81	1.000	9.3	0	-1.86	1.86	4.20	1.2	0.05	0.01667
Total			383									

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.
Ni	Sample Size. The number of subjects in the ith group. The total sample size, N, is shown as the last row of the column.
Allocation	The group sample size allocation pattern. The value on each row represents the relative number of subjects assigned to the group.
μi	The mean of the ith group at which the power is computed. The first row contains μc, the control group mean.
δi	The difference between the ith treatment mean and the control mean (μi - μc) 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.
σi	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.

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 -1.86 and 1.86 (H0: $\delta \leq -1.86$ or $\delta \geq 1.86$ versus H1: $-1.86 < \delta < 1.86$, $\delta = \mu i - \mu c$). Each of the 3 equivalence comparisons will be made using two one-sided, two-sample, Bonferroni-adjusted, unequal-variance (Welch's) t-tests. The overall (experiment-wise) Type I error rate (α) is 0.05. The group standard deviations (beginning with the control group) are assumed to be 2.16, 2.8, 2.8, and 2.8. 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 64 and the number of needed subjects for the treatment groups will be 37, 37, and 37 (totaling 175 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%	64	80	16
2	20%	37	47	10
3	20%	37	47	10
4	20%	37	47	10
Total		175	221	46
1	20%	99	124	25
2	20%	57	72	15
3	20%	57	72	15
4	20%	57	72	15
Total		270	340	70
1	20%	140	175	35
2	20%	81	102	21
3	20%	81	102	21
4	20%	81	102	21
Total		383	481	98

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 80, 47, 47, and 47 subjects should be enrolled to obtain final group sample sizes of 64, 37, 37, and 37 subjects.

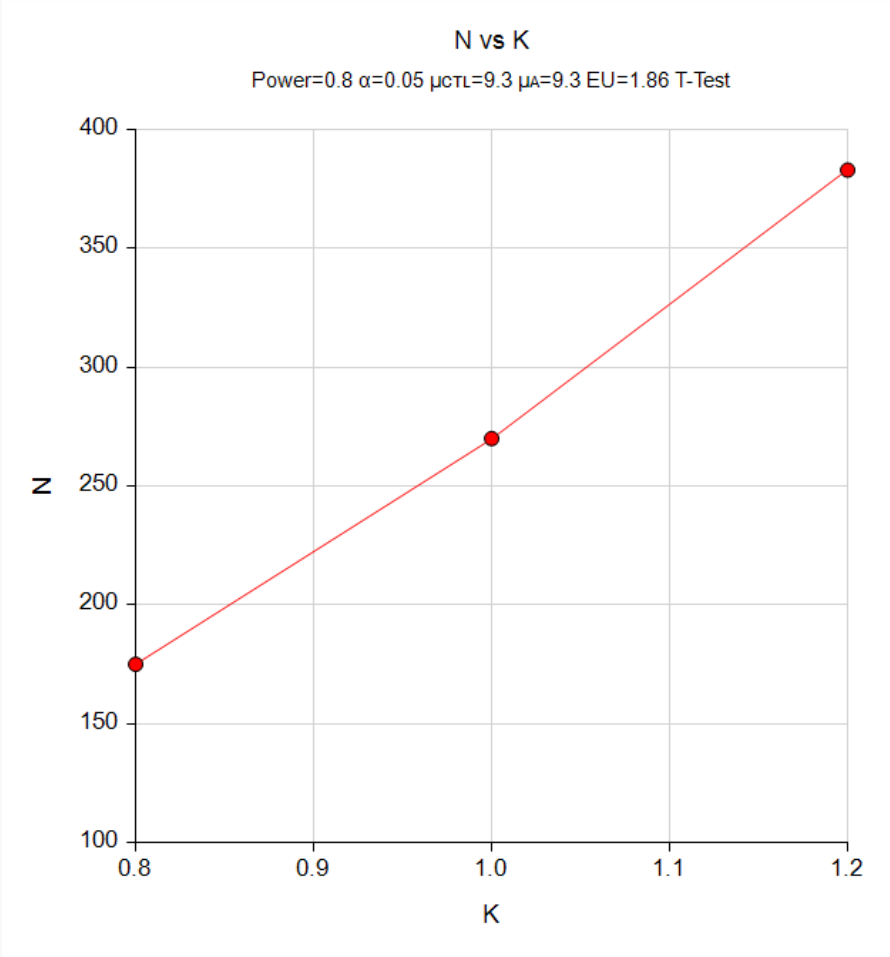
References

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.
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 of 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 Equivalence Allowing Unequal 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 mean response of 9.3. They want to consider standard deviations of 2.7 in the control group and 3.5 in the three treatment groups. 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 also 9.3. The power of each test is set to 0.80. The equivalence margin is 20% of 9.3 = 1.86, so the equivalence limits are set to -1.86 and 1.86.

The sample sizes of all groups will be equal.

The **Two-Sample T-Tests for Equivalence Allowing Unequal 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)	**1.86**
EL (Lower Equivalence Limit)	**-Upper Limit**
δ (Actual Difference)	**0**
σ1 (Standard Deviation of Group 1)	**3.5**
σ2 (Standard Deviation of Group 2)	**2.7**

This set of options generates the following report.

Numeric Results for Two One-Sided Unequal-Variance T-Tests

Solve For: Sample Size
Difference: δ = μ1 - μ2 = μT - μR
Hypotheses: H0: δ ≤ EL or δ ≥ EU vs. H1: EL < δ < EU

Target Power	Actual Power	N1	N2	N	Lower Equiv Limit EL	Upper Equiv Limit EU	δ	σ1	σ2	Alpha
0.8	0.81252	68	68	136	-1.86	1.86	0	3.5	2.7	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 68 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.

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	**0.80**
Overall Alpha	**0.05**
Bonferroni Adjustment	**Standard Bonferroni**
Group Allocation	**Equal (Nc = N1 = N2 = ...)**
EU (Upper Equivalence Limit)	**1.86**
EL (Lower Equivalence Limit)	**-Upper Limit**
Control Mean	**9.3**
Control Standard Deviation	**2.7**
Set A Number of Groups	**3**
Set A Mean	**9.3**
Set A Standard Deviation	**3.5**
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 different magnitudes, but identical ratio patterns	**Unchecked**

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
Hypotheses:	H0: δ ≤ EL or δ ≥ EU vs. H1: EL < δ < EU
Number of Groups:	4
Bonferroni Adjustment:	Standard Bonferroni (Divisor = 3)

	Power		Sample Size			Equivalence Limits			Alpha	
Comparison	Target	Actual	Ni	Mean μi	Difference δi	Lower EL	Upper EU	Standard Deviation σi	Overall	Bonferroni-Adjusted
Control			68	9.3				2.7		
vs A1	0.8	0.81249	68	9.3	0	-1.86	1.86	3.5	0.05	0.01667
vs A2	0.8	0.81249	68	9.3	0	-1.86	1.86	3.5	0.05	0.01667
vs A3	0.8	0.81249	68	9.3	0	-1.86	1.86	3.5	0.05	0.01667
Total			272							

As you can see, the sample sizes are all 68. This matches the sample size found in the validation run above. The procedure is validated.