Chapter 341

Multi-Arm Superiority by a Margin Tests for the Difference Between Treatment and Control Means Allowing Unequal Variance

Introduction

This module computes power and sample size for multiple superiority by a margin 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 superiority by a margin 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 $N = N_1 + N_2 + \cdots + N_k + N_C$.

Superiority by a Margin Tests

A *superiority by a margin test* tests that the treatment mean is better than the control mean by more than the small, superiority margin called *SM*. 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 at least a small amount above the control mean. In this case *SM* > 0. The value of δ at which power is calculated is often set to zero.

The null and alternative hypotheses are

$H_{0i}: \mu_i - \mu_C \leq SM$ vs. $H_{1i}: \mu_i - \mu_C > SM$

$H_{0i}: \mu_i \leq \mu_C + SM$ vs. $H_{1i}: \mu_i > \mu_C + SM$

$H_{0i}: \delta_i \leq SM$ vs. $H_{1i}: \delta_i > SM$

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 more than a small amount below the control mean. In this case *SM* < 0. The value of δ at which power is calculated is often set to zero.

The null and alternative hypotheses are

$H_{0i}: \mu_i - \mu_C \geq SM$ vs. $H_{1i}: \mu_i - \mu_C < SM$

$H_{0i}: \mu_i \geq \mu_C + SM$ vs. $H_{1i}: \mu_i < \mu_C + SM$

$H_{0i}: \delta_i \geq SM$ vs. $H_{1i}: \delta_i < SM$

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

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

where

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

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

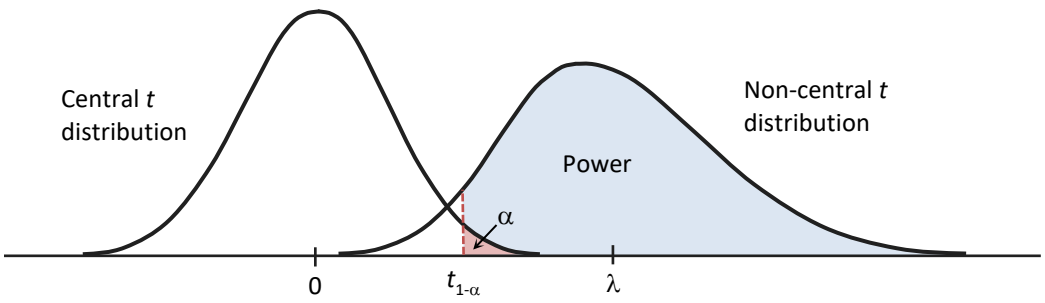
This *t*-statistic follows a *t* distribution approximately, with estimated degrees of freedom

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

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 Lokhnygina (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.



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 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 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 12.1 and the power of each test is 0.80. The superiority margin will be set to 20% of the control mean. Since higher values are desirable, the margin will be positive. The result is SM = 1.86.

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.

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: δ > SM)**
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**
SM (Superiority 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	**12.1**
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
Higher Means Are:	Better
Hypotheses:	H0: δ ≤ SM vs. H1: δ > SM
Number of Groups:	4
Bonferroni Adjustment:	Standard Bonferroni (Divisor = 3)

	Power		Sample Size				Superiority	Standard Deviation		Alpha	
Comparison	Target	Actual	Ni	Allocation	Mean μi	Difference δi	Margin SM	Value σi	Multiplier K	Overall	Bonferroni-Adjusted
Control			220	1.732	9.3			2.16	0.8		
vs A1	0.8	0.80178	127	1.000	12.1	2.8	1.86	2.80	0.8	0.025	0.00833
vs A2	0.8	0.80178	127	1.000	12.1	2.8	1.86	2.80	0.8	0.025	0.00833
vs A3	0.8	0.80178	127	1.000	12.1	2.8	1.86	2.80	0.8	0.025	0.00833
Total			601								
Control			341	1.732	9.3			2.70	1.0		
vs A1	0.8	0.80060	197	1.000	12.1	2.8	1.86	3.50	1.0	0.025	0.00833
vs A2	0.8	0.80060	197	1.000	12.1	2.8	1.86	3.50	1.0	0.025	0.00833
vs A3	0.8	0.80060	197	1.000	12.1	2.8	1.86	3.50	1.0	0.025	0.00833
Total			932								
Control			490	1.732	9.3			3.24	1.2		
vs A1	0.8	0.80074	283	1.000	12.1	2.8	1.86	4.20	1.2	0.025	0.00833
vs A2	0.8	0.80074	283	1.000	12.1	2.8	1.86	4.20	1.2	0.025	0.00833
vs A3	0.8	0.80074	283	1.000	12.1	2.8	1.86	4.20	1.2	0.025	0.00833
Total			1339								

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	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.
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.
SM	The margin of superiority in the scale of the mean difference. SM > 0.
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 superior to the control group mean by a margin, with a superiority margin of 1.86 (H0: $\delta \leq 1.86$ versus H1: $\delta > 1.86$, $\delta = \mu i - \mu c$). In this study, higher means are considered to be better. The superiority-by-a-margin 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, 2.8, 2.8, and 2.8. The control group mean is assumed to be 9.3. To detect the treatment means 12.1, 12.1, and 12.1 with at least 80% power for each test, the control group sample size needed will be 220 and the number of needed subjects for the treatment groups will be 127, 127, and 127 (totaling 601 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%	220	275	55
2	20%	127	159	32
3	20%	127	159	32
4	20%	127	159	32
Total		601	752	151
1	20%	341	427	86
2	20%	197	247	50
3	20%	197	247	50
4	20%	197	247	50
Total		932	1168	236
1	20%	490	613	123
2	20%	283	354	71
3	20%	283	354	71
4	20%	283	354	71
Total		1339	1675	336

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 275, 159, 159, and 159 subjects should be enrolled to obtain final group sample sizes of 220, 127, 127, and 127 subjects.

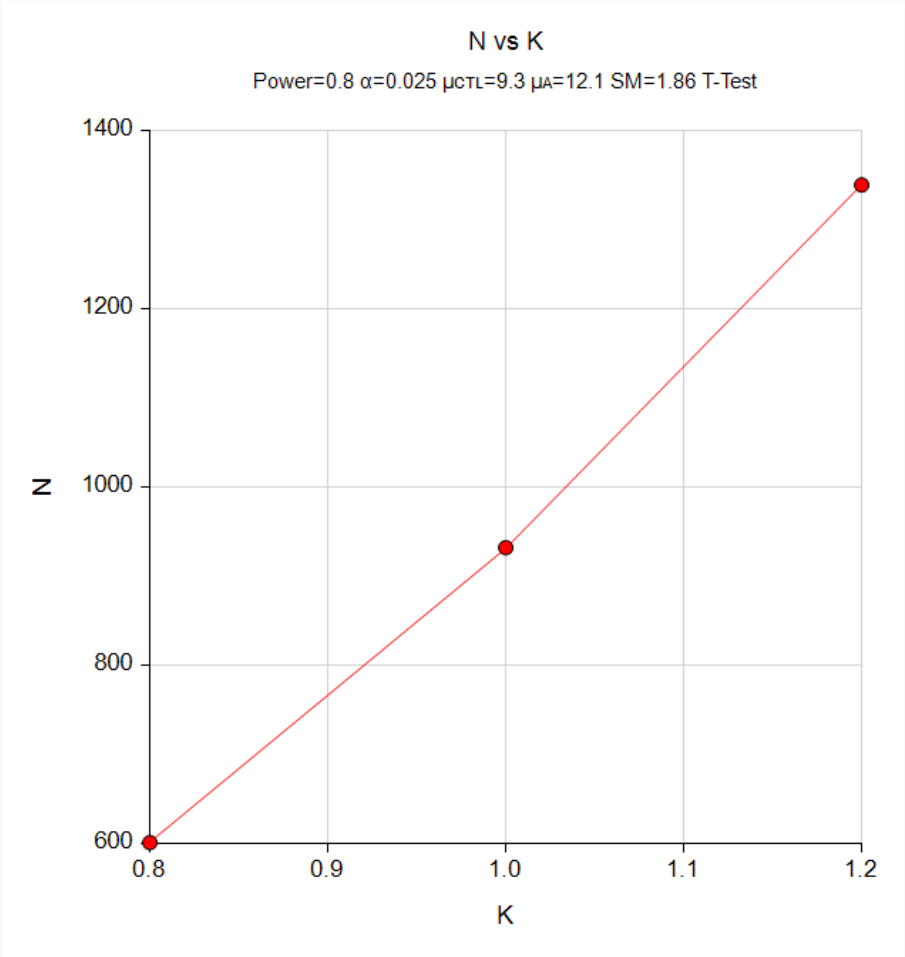
References

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 Superiority by a Margin 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 12.1 and the power of each test is 0.80. The superiority margin will be set to 20% of the control mean. Since higher values are desirable, the margin will be positive. The result is SM = 1.86.

They want to consider standard deviations of 2.7 in the control group and 3.5 in the three treatment groups. The sample sizes of all groups will be equal.

The **Two-Sample T-Tests for Superiority by a Margin Allowing Unequal Variance** procedure is set up as follows.

Design Tab

Solve For	**Sample Size**
Higher Means Are	**Better (H1: δ > SM)**
Power	**0.8**
Alpha	**0.00833** (which is Alpha / k)
Group Allocation	**Equal (N1 = N2)**
SM (Superiority Margin)	**1.86**
δ (Actual Difference to Detect)	**2.8**
σ1 (Standard Deviation of Group 1)	**2.7**
σ2 (Standard Deviation of Group 2)	**3.5**

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:	δ = μ1 - μ2 = μT - μR
Higher Means Are:	Better
Hypotheses:	H0: δ ≤ SM vs. H1: δ > SM

Power		Sample Size			Superiority Margin	Mean Difference	Standard Deviation		
Target	Actual	N1	N2	N	SM	δ	σ1	σ2	Alpha
0.8	0.80182	234	234	468	1.86	2.8	2.7	3.5	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 234 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**
Higher Means Are	**Better (H1: δ > SM)**
Power of Each Test	**0.80**
Overall Alpha	**0.025**
Bonferroni Adjustment	**Standard Bonferroni**
Group Allocation	**Equal (Nc = N1 = N2 = ...)**
SM (Superiority Margin)	**1.86**
Control Mean	**9.3**
Control Standard Deviation	**2.7**
Set A Number of Groups	**3**
Set A Mean	**12.1**
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
Higher Means Are:	Better
Hypotheses:	H0: δ ≤ SM vs. H1: δ > SM
Number of Groups:	4
Bonferroni Adjustment:	Standard Bonferroni (Divisor = 3)

	Power		Sample Size					Alpha	
Comparison	Target	Actual	Ni	Mean μi	Difference δi	Superiority Margin SM	Standard Deviation σi	Overall	Bonferroni-Adjusted
Control			234	9.3			2.7		
vs A1	0.8	0.80186	234	12.1	2.8	1.86	3.5	0.025	0.00833
vs A2	0.8	0.80186	234	12.1	2.8	1.86	3.5	0.025	0.00833
vs A3	0.8	0.80186	234	12.1	2.8	1.86	3.5	0.025	0.00833
Total			936						

As you can see, the sample sizes are all 234. This matches the sample size found in the validation run above. The procedure is validated.