Chapter 509

Superiority by a Margin Tests for Pairwise Mean Differences in a Williams Cross-Over Design

Introduction

Senn (2002) defines a *cross-over* design as one in which each subject receives all treatments and the objective is to study differences among the treatments. The name *cross-over* comes from the most common case in which there are only two treatments. In this case, each subject *crosses over* from one treatment to the other. It is assumed that there is a *washout* period between treatments during which the response returns back to its baseline value. If this does not occur, there is said to be a *carry-over* effect.

An $a \times k$ cross-over design contains *a sequences* (treatment orderings) and *k* time periods (occasions) corresponding to the *k* treatments. The design includes a washout period between responses to make certain that the effects of the first drug do not carry over to the second. Thus, the groups in this design are defined by the sequence in which the drugs are administered, not by the treatments they receive.

Cross-over designs are employed because, if the no-carryover assumption is met, treatment differences are measured within a subject rather than between subjects—making a more precise measurement. Examples of the situations that might use a cross-over design are the comparison of anti-inflammatory drugs in arthritis and the comparison of hypotensive agents in essential hypertension. In both cases, symptoms are expected to return to their usual baseline level shortly after the treatment is stopped.

The sample size calculations in the procedure are based on the formulas presented in Chow, Shao, Wang, & Lokhnygina (2018).

Advantages of Cross-Over Designs

A comparison of treatments on the same subject is expected to be more precise. The increased precision often translates into a smaller sample size. Also, patient enrollment into the study may be easier because each patient will receive both treatments. Finally, it is often more difficult to obtain a subject than to obtain a measurement.

Disadvantages of Cross-Over Designs

The statistical analysis of a cross-over experiment is more complex than a parallel-group experiment and requires additional assumptions. It may be difficult to separate the treatment effect from the period effect, the carry-over effect of the previous treatment, and the interaction between period and treatment.

The design cannot be used when the treatment (or the measurement of the response) alters the subject permanently. Hence, it should not be used to compare treatments that are intended to provide a cure.

Because subjects must be measured at least twice, it is often more difficult to keep patients enrolled in the study. It is arguably simpler to measure a subject once than to obtain their measurement twice. This is particularly true when the measurement process is painful, uncomfortable, embarrassing, or time consuming.

Technical Details

The *a* × *k* crossover design may be described as follows. Randomly assign the subjects to one of *a* sequence groups with n_1 subjects in sequence one, n_2 subjects in sequence two, and so forth up to sequence a. In order to achieve design balance, the sample sizes $n_1, n_2, \ldots, n_a$ are assumed to be equal so that $n_1 = n_2 = \cdots = n_a = n = N/a$. Sequence one is given a specific sequence of *k* treatments, sequence two is given a different sequence of the same *k* treatments, and so forth up to sequence a.

Williams Cross-Over Design

Williams cross-over designs are constructed from Latin squares as outlined in Chow and Liu (2009). If the number of treatments (*k*) is even, then Williams design results in a *k* × *k* cross-over design (i.e. with *k* sequences and *k* treatments/periods). If the number of treatments (*k*) is odd, then Williams design results in a 2*k* × *k* cross-over design (i.e. with 2*k* sequences and *k* treatments/periods). For example, a Williams design with 4 treatments would result in a 4 × 4 cross-over design and would have 4 sequences with 4 periods corresponding to the 4 treatments. On the other hand, a Williams design with 3 treatments would result in a 6 × 3 cross-over design and would have 6 sequences with 3 periods corresponding to the 3 treatments.

Define y_{ijl} as the continuous response from subject *j* (*j* = 1, …, *n*) in sequence *i* (*i* = 1, …, *a*) given treatment *l* (*l* = 1, …, *k*). The observations taken from the same subject may be correlated with one another.

Further define the paired differences between treatments *u* and *v* for each subject within each sequence as

$$d_{ij}(u,v) = y_{iju} - y_{ijv}$$

and the overall true difference as

$$\delta = \mu_u - \mu_v.$$

The overall difference can be estimated as

$$\hat{\delta} = \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n} d_{ij}(u,v).$$

The estimated difference is asymptotically normally distributed with variance σ_d^2, which can be estimated as

$$\hat{\sigma}_d^2 = \frac{1}{a(n-1)}\sum_{i=1}^{a}\sum_{j=1}^{n}\left(d_{ij}(u,v) - \bar{d}_{i\cdot}(u,v)\right)^2,$$

where

$$\bar{d}_{i\cdot}(u,v) = \frac{1}{n}\sum_{j=1}^{n} d_{ij}(u,v).$$

The standard deviation, then, is

$$SD = \sigma_d = \sqrt{\sigma_d^2}$$

with estimate

$$\widehat{SD} = \hat{\sigma}_d = \sqrt{\hat{\sigma}_d^2}.$$

Superiority by a Margin Test Statistics

Higher Means Better

When higher means are better, the null and alternative hypotheses for a one-sided superiority test are

$$H_0{:}\,\mu_u - \mu_v \le D_0 \;\text{ vs }\; H_A{:}\,\mu_u - \mu_v > D_0$$

or equivalently

$$H_0{:}\,\delta \le D_0 \;\text{ vs }\; H_A{:}\,\delta > D_0$$

where D_0 is the superiority bound (i.e. the smallest difference $(\mu_u - \mu_v)$ for which treatment *u* will be considered superior to treatment *v*). When higher means are better, D_0 should be greater than zero.

The power and sample size calculations are based on the test statistic

$$t = \frac{\hat{\delta} - D_0}{\frac{\hat{\sigma}_d}{\sqrt{an}}}$$

which follows a central *T* distribution with $a(n-1)$ degrees of freedom under the null hypothesis. The null hypothesis is rejected in favor of the alternative at level α if

$$\frac{\hat{\delta} - D_0}{\frac{\hat{\sigma}_d}{\sqrt{an}}} > t_{1-\alpha,\,a(n-1)}$$

where $t_{1-\alpha,\,a(n-1)}$ is the upper $1-\alpha$ percentile of a central *T* distribution with $a(n-1)$ degrees of freedom.

Higher Means Worse

When higher means are worse, the null and alternative hypotheses for a one-sided superiority test are

$$H_0{:}\,\mu_u - \mu_v \ge D_0 \;\text{ vs }\; H_A{:}\,\mu_u - \mu_v < D_0$$

or equivalently

$$H_0{:}\,\delta \ge D_0 \;\text{ vs }\; H_A{:}\,\delta < D_0$$

where D_0 is the superiority bound (i.e. the largest difference $(\mu_u - \mu_v)$ for which treatment *u* will be considered superior to treatment *v*). When higher means are worse, D_0 should be less than zero.

The power and sample size calculations are based on the test statistic

$$t = \frac{\hat{\delta} - D_0}{\frac{\hat{\sigma}_d}{\sqrt{an}}}$$

which follows a central T distribution with $a(n-1)$ degrees of freedom under the null hypothesis. The null hypothesis is rejected in favor of the alternative at level α if

$$\frac{\hat{\delta} - D_0}{\frac{\hat{\sigma}_d}{\sqrt{an}}} < t_{\alpha,\, a(n-1)}$$

where $t_{\alpha,\, a(n-1)}$ is the lower α percentile of a central T distribution with $a(n-1)$ degrees of freedom.

Bonferroni Adjustment for Multiple Tests

In a design with k treatments, there are $k(k - 1)/2$ possible pairwise (u, v) comparison tests. To protect the overall alpha level, the individual test alpha level if often divided by the number of tests performed. This is known as the Bonferroni adjustment for multiple comparisons. When this adjustment is used in hypothesis testing, the individual test alpha value of $\alpha/(k(k - 1)/2)$ is substituted for α in the formulas above.

Superiority by a Margin Power Calculations

Higher Means Better

According to Chow, Shao, Wang, & Lokhnygina (2018) page 65, the power for the one-sided superiority test of $H_0{:}\, \delta \leq D_0$ versus $H_A{:}\, \delta > D_0$ is

$$1 - T_{a(n-1)}\left(t_{1-\alpha,\, a(n-1)} \middle| \frac{\delta_1 - D_0}{\frac{\sigma_d}{\sqrt{an}}}\right)$$

where $T_{df}(X|NCP)$ is the non-central T distribution function with df degrees of freedom and non-centrality parameter NCP evaluated at X, δ_1 is the actual value of the minimum difference under the alternative hypothesis, and $t_{1-\alpha,\, a(n-1)}$ is the upper $1-\alpha$ percentile of a central T distribution with $a(n-1)$ degrees of freedom. The sample size is determined using a binary search of possible values for n.

Higher Means Worse

Derived from Chow, Shao, Wang, & Lokhnygina (2018) page 65, the power for the one-sided superiority test of $H_0{:}\, \delta \geq D_0$ versus $H_A{:}\, \delta < D_0$ is

$$1 - T_{a(n-1)}\left(t_{1-\alpha,\, a(n-1)} \middle| \frac{D_0 - \delta_1}{\frac{\sigma_d}{\sqrt{an}}}\right)$$

where $T_{df}(X|NCP)$ is the non-central T distribution function with df degrees of freedom and non-centrality parameter NCP evaluated at X, δ_1 is the actual value of the minimum difference under the alternative hypothesis, and $t_{1-\alpha,\, a(n-1)}$ is the upper $1-\alpha$ percentile of a central T distribution with $a(n-1)$ degrees of freedom. The sample size is determined using a binary search of possible values for n.

Bonferroni Adjustment for Multiple Tests

In a design with k treatments, there are $k(k - 1)/2$ possible pairwise (u, v) comparison tests. To protect the overall alpha level, the individual test alpha level if often divided by the number of tests performed. This is known as the Bonferroni adjustment for multiple comparisons. When this adjustment is used in power calculations, the individual test alpha value of $\alpha/(k(k - 1)/2)$ is substituted for α in the formulas above.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Design tab. For more information about the options of other tabs, go to the Procedure Window chapter.

Design Tab

The Design tab contains most of the parameters and options that you will be concerned with.

Solve For

Solve For

This option specifies the parameter to be calculated from the values of the other parameters. Under most conditions, you would select either *Power* or *Sample Size*.

Select *Sample Size* when you want to determine the sample size needed to achieve a given power and alpha level.

Select *Power* when you want to calculate the power of an experiment that has already been run.

Select *Effect Size (D1)* when you want to calculate the minimum effect size that can be detected for a particular scenario.

Test

Higher Means Are

Use this option to specify the direction of the test.

If Higher Means are “Better”, the alternative hypothesis is H1: $\mu_u - \mu_v > D0$.

If Higher Means are “Worse”, the alternative hypothesis is H1: $\mu_u - \mu_v < D0$.

Power and Alpha

Power

This option specifies one or more values for power. Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of equal means when in fact the means are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Adjust Alpha for Multiple Tests

Check this box to adjust the alpha-level for each individual test to maintain the overall experiment-wise error rate of Alpha.

The adjustment is made using the Bonferroni method where the overall Alpha is divided by the number of tests. The total number of tests is equal to k(k-1)/2, where k is the number of treatments.

Sample Size / Treatments

k (Number of Treatments)

This is the number of treatments given to each subject in each sequence.

Number of Sequences

If k is even, the number of sequences (a) in Williams design is equal to k, resulting in a k x k cross-over design.

If k is odd, the number of sequences (a) in Williams design is equal to 2k, resulting in a 2k x k cross-over design.

Number of Tests

The total number of tests is equal to k(k-1)/2.

Range

$k \geq 2$.

n (Sample Size per Sequence)

This is the sample size of each sequence in the Williams cross-over design. The individual sequence sample sizes are assumed to be equal such that the total sample size is equal to

$$N = an$$

where *a* is the number of sequences. If the number of treatments (k) is even, the number of sequences (a) in Williams design is equal to k, resulting in a k x k cross-over design. If k is odd, the number of sequences (a) in Williams design is equal to 2k, resulting in a 2k x k cross-over design.

You can enter a single value such as *50* or a list of values using the syntax *50 100 150 200 250* or *50 to 250 by 50*.

Effect Size – Difference

D0 (Superiority Difference)

Specify the superiority difference.

When higher means are “Better”, the superiority difference is the smallest treatment difference (μ_u - μ_v) for which treatment u will be considered superior to treatment v.

When higher means are “Worse”, the superiority difference is the largest treatment difference (μ_u - μ_v) for which treatment u will be considered superior to treatment v.

You can enter a single value such as *1* or a series of values such as *1 2 3* or *1 to 3 by 1* in the range D0 ≠ D1. When higher means are “Better”, D0 should be greater than 0. When higher means are “Worse”, D0 should be less than 0.

D1 (Minimum Difference|H1)

Enter a value for the minimum treatment difference to detect under the alternative hypothesis, H1. The power calculations assume that this is the actual value of the difference.

$$D1 = \text{Minimum of } (\mu_u - \mu_v)|H1 \text{ for } u, v = 1, ..., k \text{ with } u \neq v$$

You can enter a single value such as *3* or a series of values such as *3 4 5* or *3 to 5 by 1* in the range D1 ≠ D0. When higher means are “Better”, D1 should be greater than D0. When higher means are “Worse”, D1 should be less than D0.

Effect Size – Standard Deviation of Paired Differences

Standard Deviation (SD)

Enter a value for the standard deviation of the paired differences, SD.

Estimating SD using Previous Cross-Over Data

The standard deviation may be estimated using cell counts from a previous cross-over study with n subjects per sequence as described on pages 63 and 64 of Chow, Shao, Wang, & Lokhnygina (2018).

Assume that y_ijl is the continuous treatment response for the jth subject (j = 1 to n) in the ith sequence (i = 1, ..., a) given the lth treatment (l = 1, ..., k). Note that we assume that there are equal numbers of subjects in each sequence such that n_1 = n_2 = ... = n_a = n.

Define

d_ij_(u,v) = y_iju - y_ijv

dbar_i._(u,v) = (1/n)Σ_j[d_ij_(u,v)]

The formula for SD is then

SD = √[($\Sigma_i\Sigma_j$(d_ij_(u,v) - dbar_i._(u,v))^2)/(a(n-1))].

Example 1 – Power Analysis

Suppose you want to consider the power of a balanced Williams cross-over design with 3 groups and a continuous endpoint where the test is computed based on the difference for sequence sample sizes between 30 and 100. The actual minimum difference is 1.5, the superiority difference is 1, and the estimated standard deviation of the paired differences is 3.5. The overall significance level is 0.05 with individual test alpha adjusted for 3 tests.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Superiority by a Margin Tests for Pairwise Mean Differences in a Williams Cross-Over Design** procedure window by expanding **Means**, then **Cross-Over (Williams) Design**, then clicking on **Superiority by a Margin**, and then clicking on **Superiority by a Margin Tests for Pairwise Mean Differences in a Williams Cross-Over Design**. You may then make the appropriate entries as listed below, or open **Example 1** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Higher Means Are	**Better**
Alpha	**0.05**
Adjust Alpha for Multiple Tests	**Checked**
k (Number of Treatments)	**3**
n (Sample Size per Sequence)	**30 to 100 by 10**
D0 (Superiority Difference)	**1**
D1 (Minimum Difference|H1)	**1.5**
Standard Deviation (SD)	**3.5**

Annotated Output

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

Numeric Results

```
Numeric Results for a Superiority by a Margin T-Test in a 6x3 Williams Cross-Over Design
H0: μu - μv ≤ D0 vs. H1: μu - μv > D0 for u, v = 1, ..., 3 with u ≠ v.
Number of Possible Tests = 3

               Sequence        Total
                 Sample       Sample  Superiority    Minimum    Standard                  Individual
                   Size         Size   Difference Difference   Deviation      Overall           Test
Power                 n            N           D0         D1          SD       Alpha*         Alpha*
0.41142              30          180        1.000      1.500       3.500        0.050          0.017
0.52964              40          240        1.000      1.500       3.500        0.050          0.017
0.63186              50          300        1.000      1.500       3.500        0.050          0.017
0.71695              60          360        1.000      1.500       3.500        0.050          0.017
0.78572              70          420        1.000      1.500       3.500        0.050          0.017
0.83997              80          480        1.000      1.500       3.500        0.050          0.017
0.88191              90          540        1.000      1.500       3.500        0.050          0.017
0.91380             100          600        1.000      1.500       3.500        0.050          0.017

* Alpha was adjusted for 3 tests using the Bonferroni method. Power was calculated using Individual Test Alpha.

References
Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. 2018. Sample Size Calculations in Clinical Research, Third
   Edition. Chapman & Hall/CRC. Boca Raton, Florida.
```

Report Definitions
Power is the probability of rejecting a false null hypothesis. It should be close to one.
n is the sample size in each sequence.
N is the total sample size from all 6 sequences combined. The sample is divided equally among sequences.
D0 is the superiority difference used to specify the hypothesis test.
D1 is the minimum treatment difference to detect at which power is calculated. D1 = Minimum of $(\mu_u - \mu_v)|H1$ for u, v = 1, ..., k with u ≠ v.
SD is the standard deviation of paired differences. This is estimated from a previous study.
Alpha is the probability of rejecting a true null hypothesis. It should be small.

Summary Statements

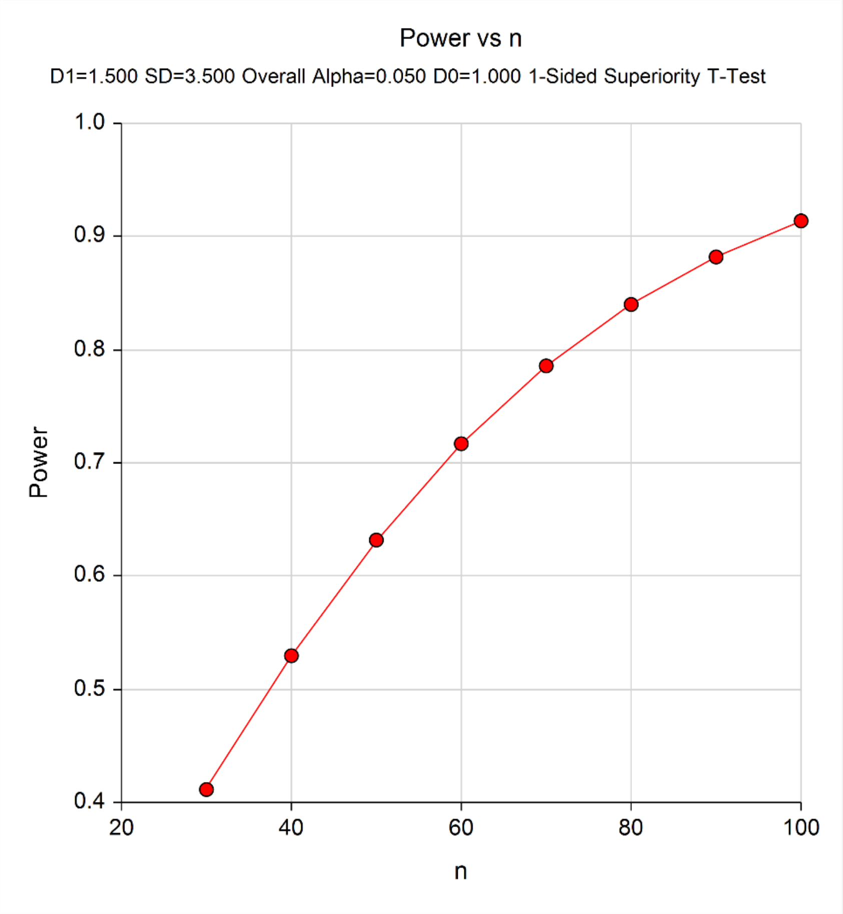
For a 6x3 Williams Cross-Over Design, a sample size of 30 in each sequence for a total of 180 achieves 41.142% power to detect a difference of 1.500 using a one-sided superiority by a margin T-Test against a bound of 1.000 with an overall significance level of 0.050 and individual test Bonferroni-adjusted significance level of 0.017 when the standard deviation of paired differences is 3.500.

Dropout-Inflated Sample Size

Group	Dropout Rate	Sample Size Ni	Dropout- Inflated Enrollment Sample Size Ni'	Expected Number of Dropouts Di
1 - 6	20%	30	38	8
Total		180	228	48
1 - 6	20%	40	50	10
Total		240	300	60
1 - 6	20%	50	63	13
Total		300	378	78
1 - 6	20%	60	75	15
Total		360	450	90
1 - 6	20%	70	88	18
Total		420	528	108
1 - 6	20%	80	100	20
Total		480	600	120
1 - 6	20%	90	113	23
Total		540	678	138
1 - 6	20%	100	125	25
Total		600	750	150

Definitions
Group lists the group numbers.
Dropout Rate (DR) is 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”).
Ni is 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' is the number of subjects that should be enrolled in each group in order to end up with 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., and Wang, H. (2008) pages 39-40.)
Di is the expected number of dropouts in each group. Di = Ni' - Ni.

Charts Section



This report shows the values of each of the parameters, one scenario per row. This plot shows the relationship between sample size and power. We see that a sample size of about 70 per sequence is required to detect a minimum difference of 1.5 with 80% power when the superiority bound is 1.

Example 2 – Calculating Sample Size (Validation using Hand Calculations)

In this example, we’ll find the sample size required in a 6 × 3 Williams cross-over design (k = 3) to detect a difference of 1.2 with 80% power in a superiority test with a margin of 1 with a significance level of 0.05 when the standard deviation of paired differences is 1.5. We’ll make no adjustment for multiple testing in this example. We’ll also validate this procedure by computing power values manually.

The power for per-sequence sample sizes of 58 and 59 calculated by hand using the power formula referenced earlier is

$$\text{Power} = 1 - T_{a(n-1)}\left(t_{1-\alpha,\, a(n-1)} \middle| \frac{\delta_1 - D_0}{\frac{\sigma_d}{\sqrt{an}}}\right)$$

$$\text{Power}_{(n=58)} = 1 - T_{342}\left(1.649321 \middle| \frac{1.2 - 1}{\frac{1.5}{\sqrt{6 \times 58}}}\right)$$

$$= 0.798851$$

$$\text{Power}_{(n=59)} = 1 - T_{348}\left(1.649244 \middle| \frac{1.2 - 1}{\frac{1.5}{\sqrt{6 \times 59}}}\right)$$

$$= 0.804807$$

These results indicate that the minimum required sample size per group is 59, since it is the smallest sample size that achieves the desired 80% power.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Superiority by a Margin Tests for Pairwise Mean Differences in a Williams Cross-Over Design** procedure window by expanding **Means**, then **Cross-Over (Williams) Design**, then clicking on **Superiority by a Margin**, and then clicking on **Superiority by a Margin Tests for Pairwise Mean Differences in a Williams Cross-Over Design**. You may then make the appropriate entries as listed below, or open **Example 2** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Higher Means Are	**Better**
Power	**0.80**
Alpha	**0.05**
Adjust Alpha for Multiple Tests	**Unchecked**
k (Number of Treatments)	**3**
D0 (Superiority Difference)	**1**
D1 (Minimum Difference|H1)	**1.2**
Standard Deviation (SD)	**1.5**

Output

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

Numeric Results

```
Numeric Results for a Superiority by a Margin T-Test in a 6x3 Williams Cross-Over Design ──────────
H0: μu - μv ≤ D0 vs. H1: μu - μv > D0 for u, v = 1, ..., 3 with u ≠ v.
Number of Possible Tests = 3

              Sequence          Total
                Sample         Sample  Superiority    Minimum   Standard
                  Size           Size   Difference Difference  Deviation
Power                n              N           D0         D1         SD      Alpha*
0.80481             59            354        1.000      1.200      1.500       0.050

* Alpha was not adjusted for multiple tests.
```

The result from **PASS** is match our hand calculations exactly.