Chapter 185

Equivalence Tests for Pairwise Proportion 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 $a \times k$ 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.

The statistical model employed by this procedure and given in Chow, Shao, Wang, & Lokhnygina (2018) assumes that there are no sequence, period, or cross-over effects. The statistical model that incorporates these effects is complex for binary data.

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 $k \times k$ 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 $2k \times k$ cross-over design (i.e. with $2k$ 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 binary response from subject j ($j = 1, \ldots, n$) in sequence i ($i = 1, \ldots, a$) given treatment l ($l = 1, \ldots, k$). Assume that the responses y_{ijl} are independent and randomly distributed with $P(y_{ijl} = 1) = P_l$, which implies that there are no sequence, period, or cross-over effects. 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

$$\delta = P_u - P_v.$$

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

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

Equivalence Test Statistics

The null and alternative hypotheses for an equivalence test are

$$H_0: P_u - P_v \le D_{0L} \text{ or } P_u - P_v \ge D_{0U} \quad \text{vs} \quad H_A: D_{0L} < P_u - P_v < D_{0U}$$

or equivalently

$$H_0: \delta \le D_{0L} \text{ or } \delta \ge D_{0U} \quad \text{vs} \quad H_A: D_{0L} < \delta < D_{0U}$$

where D_{0L} and D_{0U} are the lower and upper equivalence bounds, respectively (i.e. the smallest and largest differences ($P_u - P_v$) for which treatment *u* and treatment *v* will be considered equivalent).

The power and sample size calculations are based on the two one-sided test (TOST) statistics

$$Z_L = \frac{\hat{\delta} - D_{0L}}{\frac{\hat{\sigma}_d}{\sqrt{an}}} \quad \text{and} \quad Z_U = \frac{\hat{\delta} - D_{0U}}{\frac{\hat{\sigma}_d}{\sqrt{an}}}$$

which are each asymptotically distributed as standard normal under the null hypothesis. The null hypothesis is rejected in favor of the alternative at level α using the TOST procedure if

$$Z_L > Z_{1-\alpha} \quad \text{and} \quad Z_U < Z_\alpha$$

where $Z_{1-\alpha}$ is the upper $1-\alpha$ percentile and Z_α is the lower α percentile of the standard normal distribution.

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.

Equivalence Power Calculations

Derived from Chow, Shao, Wang, & Lokhnygina (2018) page 90, the power for an equivalence test of $H_0: \delta \leq D_{0L}$ or $\delta \geq D_{0U}$ versus $H_A: D_{0L} < \delta < D_{0U}$ is given as

$$\Phi\left(\frac{D_{0U} - \delta_1}{\frac{\sigma_d}{\sqrt{an}}} - Z_{1-\alpha}\right) - \Phi\left(\frac{D_{0L} - \delta_1}{\frac{\sigma_d}{\sqrt{an}}} + Z_{1-\alpha}\right)$$

where $\Phi()$ is the standard normal distribution function, δ_1 is the actual value of the difference under the alternative hypothesis, and $Z_{1-\alpha}$ is the upper $1-\alpha$ percentile of the standard normal distribution. 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 design.

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 – Equivalence Differences

D0.U (Upper Equivalence Difference)

Specify the upper equivalence bound for the difference. This value along with the lower equivalence difference (D0.L) is used to setup the hypothesis test. This value represents the largest difference (Pu - Pv) for which treatment u and treatment v will be considered equivalent.

You can enter a single value such as *0.5* or a series of values such as *0.5 0.6 0.7* or *0.5 to 0.7 by 0.1* in the range $0 < D0.U < 1$ and $D0.U > D1$.

D0.L (Lower Equivalence Difference)

Specify the lower equivalence bound for the difference. This value along with the upper equivalence difference (D0.U) is used to setup the hypothesis test. This value represents the smallest difference (Pu - Pv) for which treatment u and treatment v will be considered equivalent.

For symmetric bounds, enter *-D0.U*. This is the default. You can also enter a single value such as *-0.5* or a series of values such as *-0.7 -0.6 -0.5* or *-0.7 to -0.5 by 0.1* in the range -1 < D0.L < 0 and D0.L < D1. Note that if you enter values for D0.L other than *-D0.U*, they are used in pairs with the values of D0.U. Thus, the first values of D0.U and D0.L are used, then the second values of each are used, and so on.

Effect Size – Actual Difference

D1 (Minimum Difference|H1)

Enter a value for the actual minimum difference (Pu - Pv) at which power is calculated.

You can enter a single value such as *0* or a series of values such as *0 0.1 0.2* or *0 to 0.2 by 0.1* in the range -1 < D1 < 1 and D0.L < D1 < D0.U.

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 88 and 89 of Chow, Shao, Wang, & Lokhnygina (2018).

Assume that y_ijl is the binary 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

$$d\_ij\_(u,v) = y\_iju - y\_ijv$$

$$dbar\_i.\_(u,v) = (1/n)\Sigma_j[d\_ij\_(u,v)]$$

The formula for SD is then

$$SD = \sqrt{[(\Sigma_i\Sigma_j(d\_ij\_(u,v) - dbar\_i.\_(u,v))^2)/(a(n-1))]}.$$

Example 1 – Power Analysis

Suppose you want to consider the power of an equivalence test of the hypotheses $H_0: \delta \leq -0.1$ or $\delta \geq 0.1$ versus $H_A: -0.1 < \delta < 0.1$ in a balanced Williams cross-over design with 3 groups and a binary endpoint where the test is computed based on the difference for sequence sample sizes between 50 and 300. Let's assume that the actual difference is 0 and the estimated standard deviation of the paired differences is 1. 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 **Equivalence Tests for Pairwise Proportion Differences in a Williams Cross-Over Design** procedure window by expanding **Proportions**, then **Cross-Over (Williams) Design**, then clicking on **Equivalence**, and then clicking on **Equivalence Tests for Pairwise Proportion 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**
Alpha	**0.05**
Adjust Alpha for Multiple Tests	**Checked**
k (Number of Treatments)	**3**
n (Sample Size per Sequence)	**50 to 300 by 50**
D0.U (Upper Equivalence Difference)	**0.1**
D0.L (Lower Equivalence Difference)	**-D0.U**
D1 (Minimum Difference|H1)	**0**
Standard Deviation (SD)	**1**

Annotated Output

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

Numeric Results

Numeric Results for an Equivalence Test in a 6x3 Williams Cross-Over Design

H0: $P_u - P_v \leq D0.L$ or $P_u - P_v \geq D0.U$ vs. H1: $D0.L < P_u - P_v < D0.U$ for u, v = 1, ..., 3 with $u \neq v$.
Number of Possible Tests = 3

Power	Sequence Sample Size n	Total Sample Size N	Lower Equiv. Difference D0.L	Upper Equiv. Difference D0.U	Minimum Difference D1	Standard Deviation SD	Overall Alpha*	Individual Test Alpha*
0.00000	50	300	-0.100	0.100	0.000	1.000	0.050	0.017
0.25213	100	600	-0.100	0.100	0.000	1.000	0.050	0.017
0.61677	150	900	-0.100	0.100	0.000	1.000	0.050	0.017
0.81847	200	1200	-0.100	0.100	0.000	1.000	0.050	0.017
0.91900	250	1500	-0.100	0.100	0.000	1.000	0.050	0.017
0.96554	300	1800	-0.100	0.100	0.000	1.000	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.

Equivalence Tests for Pairwise Proportion Differences in a Williams Cross-Over Design

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.L is the lower equivalence difference used to specify the hypothesis test.
D0.U is the upper equivalence difference used to specify the hypothesis test.
D1 is the minimum treatment difference to detect at which power is calculated. D1 = Minimum of $(P_u - P_v)|H1$ for u, v = 1, ..., k with $u \neq v$.
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 50 in each sequence for a total of 300 achieves 0.000% power to detect a difference of 0.000 using an equivalence test with lower and upper equivalence bounds of -0.100 and 0.100, respectively, 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 1.000.

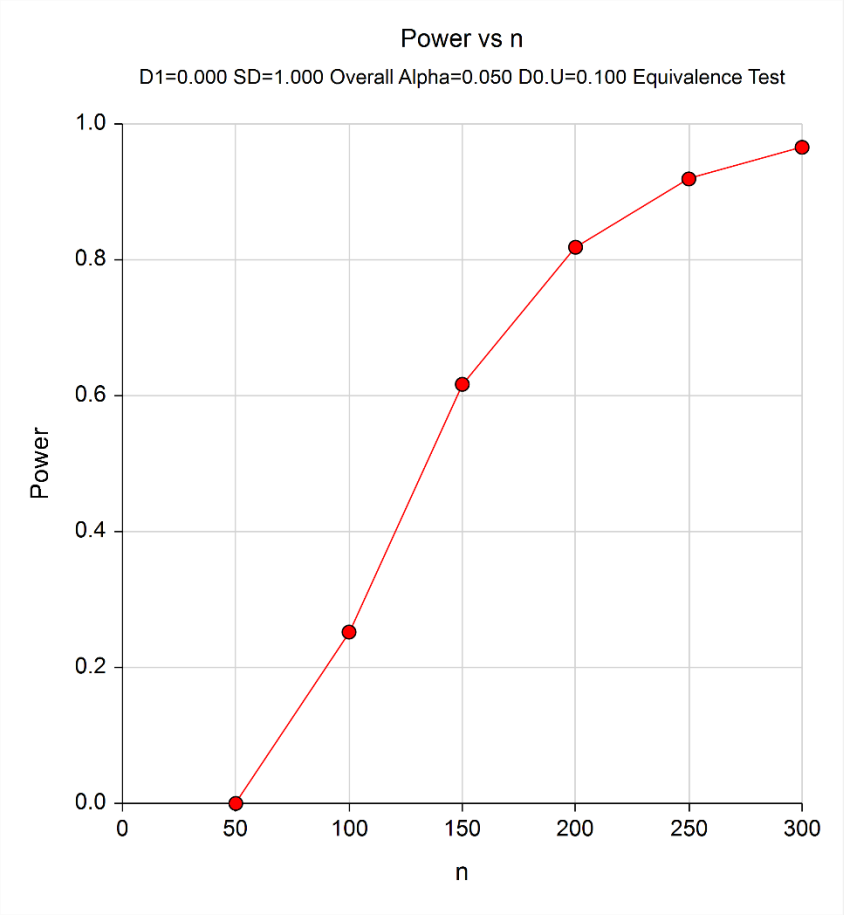
Dropout-Inflated Sample Size

Group	Dropout Rate	Sample Size Ni	Dropout-Inflated Enrollment Sample Size Ni'	Expected Number of Dropouts Di
1 - 6	20%	50	63	13
Total		300	378	78
1 - 6	20%	100	125	25
Total		600	750	150
1 - 6	20%	150	188	38
Total		900	1128	228
1 - 6	20%	200	250	50
Total		1200	1500	300
1 - 6	20%	250	313	63
Total		1500	1878	378
1 - 6	20%	300	375	75
Total		1800	2250	450

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 just under 200 per sequence is required to detect a minimum difference of 0 with 80% power when the equivalence bounds are -0.1 and 0.1.

Example 2 – Calculating Sample Size (Validation using Chow, Shao, Wang, & Lokhnygina (2018) and Hand Calculations)

On page 92, Chow, Shao, Wang, & Lokhnygina (2018) presents an example of finding the sample size required in a 6 × 3 Williams cross-over design (k = 3) to detect a difference of 0.2 with 80% power in an equivalence test with a margin of 0.2 and a significance level of 0.05 when the standard deviation of paired differences is 0.75. They compute the required sample size to be 80 per sequence. Note that there is no adjustment for multiple testing in this example. Further note that this sample size is based on a very conservative estimate of power not used by **PASS**. **PASS** uses the complete power formula based on the two one-sided tests themselves for its calculations. We'll demonstrate in this example that the calculation of sample size by **PASS** is correct based on the complete power formula.

The power for per-sequence sample sizes of 57 and 58 calculated by hand using the full power formula used by **PASS** is

$$\text{Power} = \Phi\left(\frac{D_{0U} - \delta_1}{\frac{\sigma_d}{\sqrt{an}}} - Z_{1-\alpha}\right) - \Phi\left(\frac{D_{0L} - \delta_1}{\frac{\sigma_d}{\sqrt{an}}} + Z_{1-\alpha}\right)$$

$$\text{Power}_{(n=57)} = \left(\frac{0.3 - 0.2}{\frac{0.75}{\sqrt{6 \times 57}}} - 1.644854\right) - \Phi\left(\frac{-0.3 - 0.2}{\frac{0.75}{\sqrt{6 \times 57}}} + 1.644854\right)$$

$$= 0.794152$$

$$\text{Power}_{(n=58)} = \left(\frac{0.3 - 0.2}{\frac{0.75}{\sqrt{6 \times 58}}} - 1.644854\right) - \Phi\left(\frac{-0.3 - 0.2}{\frac{0.75}{\sqrt{6 \times 58}}} + 1.644854\right)$$

$$= 0.800231$$

These results indicate that the minimum required sample size per group is 58, since it is the smallest sample size that achieves the desired 80% power.

The power formula that Chow, Shao, Wang, & Lokhnygina (2018) uses on page 90 to arrive at a "conservative" sample size of 80 per sequence on page 92 is much more conservative with power estimate of

$$\text{Power} = 2\Phi\left(\frac{D_{0U} - |\delta_1|}{\frac{\sigma_d}{\sqrt{an}}} - Z_{1-\alpha}\right) - 1$$

$$\text{Power}_{(n=58)} = 2\Phi\left(\frac{0.3 - |0.2|}{\frac{0.75}{\sqrt{6 \times 58}}} - 1.644854\right) - 1$$

$$= 0.600462$$

This estimate of power is overly conservative when $\delta_1 \neq 0$ (see Chow, Shao, Wang, & Lokhnygina (2018) page 44).

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Pairwise Proportion Differences in a Williams Cross-Over Design** procedure window by expanding **Proportions**, then **Cross-Over (Williams) Design**, then clicking on **Equivalence**, and then clicking on **Equivalence Tests for Pairwise Proportion 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**
Power	**0.80**
Alpha	**0.05**
Adjust Alpha for Multiple Tests	**Unchecked**
k (Number of Treatments)	**3**
D0.U (Upper Equivalence Difference)	**0.3**
D0.L (Lower Equivalence Difference)	**-D0.U**
D1 (Minimum Difference|H1)	**0.2**
Standard Deviation (SD)	**0.75**

Output

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

Numeric Results

Numeric Results for an Equivalence Test in a 6x3 Williams Cross-Over Design

H0: $P_u - P_v \le D0.L$ or $P_u - P_v \ge D0.U$ vs. H1: $D0.L < P_u - P_v < D0.U$ for u, v = 1, ..., 3 with $u \ne v$.

Number of Possible Tests = 3

Power	Sequence Sample Size n	Total Sample Size N	Lower Equiv. Difference D0.L	Upper Equiv. Difference D0.U	Minimum Difference D1	Standard Deviation SD	Alpha*
0.80023	58	348	-0.300	0.300	0.200	0.750	0.050

* Alpha was not adjusted for multiple tests.

The results from **PASS** match our hand calculations above exactly. As a side note, a sample size of 80 achieves 89.908% power using the complete power formula.