

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, & Lohknygina (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, \dots, n_a are assumed to be equal so that $n_1 = n_2 = \dots = n_a = n = N/a$. 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, & Lohknygina (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 ($j = 1, \dots, n$) in sequence i ($i = 1, \dots, a$) given treatment l ($l = 1, \dots, k$). 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

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

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

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

$$H_0: P_u - P_v \leq D_{0L} \text{ or } P_u - P_v \geq D_{0U} \quad \text{vs.} \quad H_A: D_{0L} < P_u - P_v < D_{0U}$$

or equivalently

$$H_0: \delta \leq D_{0L} \text{ or } \delta \geq D_{0U} \quad \text{vs.} \quad H_A: D_{0L} < \delta < D_{0U}$$

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 is 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 is 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.

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

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	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

Output

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

Numeric Results

Solve For: [Power](#)
 Design: 6x3 Williams Cross-Over Design
 Hypotheses: $H_0: P_u - P_v \leq D_{0.L}$ or $P_u - P_v \geq D_{0.U}$ vs. $H_1: D_{0.L} < P_u - P_v < D_{0.U}$ for $u, v = 1, \dots, 3$ with $u \neq v$.
 Number of Possible Tests: 3

Power	Sample Size		Difference		Minimum D1	Standard Deviation SD	Alpha*	
	Sequence n	Total N	Equivalence				Overall	Individual Test
			Lower D0.L	Upper D0.U				
0.00000	50	300	-0.1	0.1	0	1	0.05	0.017
0.25213	100	600	-0.1	0.1	0	1	0.05	0.017
0.61677	150	900	-0.1	0.1	0	1	0.05	0.017
0.81847	200	1200	-0.1	0.1	0	1	0.05	0.017
0.91900	250	1500	-0.1	0.1	0	1	0.05	0.017
0.96554	300	1800	-0.1	0.1	0	1	0.05	0.017

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

- Power The probability of rejecting a false null hypothesis when the alternative hypothesis is true.
- n The sample size in each sequence.
- N The total sample size from all 6 sequences combined. The sample is divided equally among sequences.
- D0.L The lower equivalence difference used to specify the hypothesis test.
- D0.U The upper equivalence difference used to specify the hypothesis test.
- D1 The minimum treatment difference to detect at which power is calculated. $D1 = \text{Minimum of } (P_u - P_v) | H_1 \text{ for } u, v = 1, \dots, k \text{ with } u \neq v$.
- SD The standard deviation of paired differences. This is estimated from a previous study.
- Alpha The probability of rejecting a true null hypothesis.

Summary Statements

A Williams cross-over design with 3 treatments and 6 sequences will be used to test whether each proportion is equivalent to the others, with equivalence difference bounds of -0.1 and 0.1 ($H_0: P_u - P_v \leq -0.1$ or $P_u - P_v \geq 0.1$ versus $H_1: -0.1 < P_u - P_v < 0.1$ for $u, v = 1, \dots, 3$ with $u \neq v$). Each comparison will be made using two one-sided proportion difference Z-tests (for one equivalence test), with a Bonferroni-adjusted individual test Type I error rate of 0.017 (based on 3 comparisons) to give an overall (familywise) Type I error rate (α) of 0.05. The standard deviation of paired differences is assumed to be 1. To detect a proportion difference of 0 with a sample size of 50 in each sequence (totaling 300 subjects), the power is 0.

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

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

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 Lohknygina, Y. (2018) pages 32-33.)

Di The expected number of dropouts in each group. $Di = Ni' - Ni$.

Dropout Summary Statements

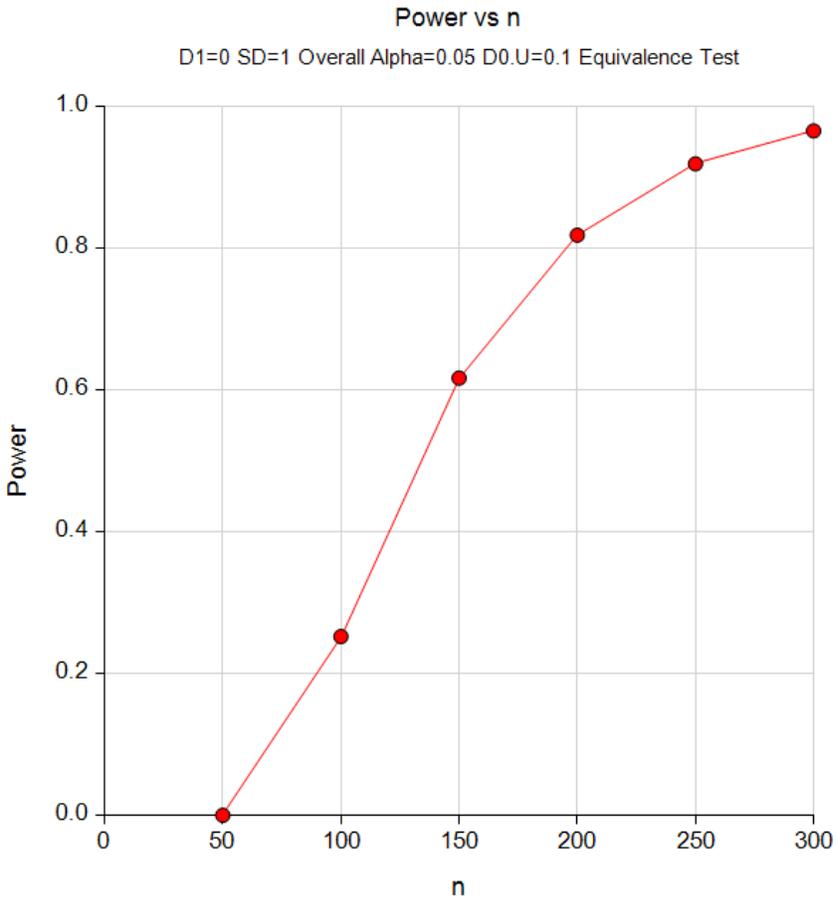
Anticipating a 20% dropout rate, group sizes of 63, 63, 63, 63, 63, and 63 subjects should be enrolled to obtain final group sample sizes of 50, 50, 50, 50, 50, and 50 subjects.

References

Chow, S.C., Shao, J., Wang, H., and Lohknygina, Y. 2018. Sample Size Calculations in Clinical Research, Third Edition. Taylor & Francis/CRC. Boca Raton, Florida.

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

Plots



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

$$\begin{aligned} \text{Power} &= \Phi\left(\frac{D_{0U} - \delta_1}{\frac{\sigma_d}{\sqrt{an}}} - Z_{1-\alpha}\right) - \Phi\left(\frac{D_{0L} - \delta_1}{\frac{\sigma_d}{\sqrt{an}}} + Z_{1-\alpha}\right) \\ \text{Power}_{(n=57)} &= \left(\frac{0.3 - 0.2}{\frac{0.75}{\sqrt{6 \times 57}}} - 1.644854\right) - \Phi\left(\frac{-0.3 - 0.2}{\frac{0.75}{\sqrt{6 \times 57}}} + 1.644854\right) \\ &= 0.794152 \\ \text{Power}_{(n=58)} &= \left(\frac{0.3 - 0.2}{\frac{0.75}{\sqrt{6 \times 58}}} - 1.644854\right) - \Phi\left(\frac{-0.3 - 0.2}{\frac{0.75}{\sqrt{6 \times 58}}} + 1.644854\right) \\ &= 0.800231 \end{aligned}$$

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

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

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

$$\begin{aligned} \text{Power}_{(n=58)} &= 2\Phi\left(\frac{\frac{0.3 - |0.2|}{0.75}}{\sqrt{6 \times 58}} - 1.644854\right) - 1 \\ &= 0.600462 \end{aligned}$$

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

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.....	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

Solve For: [Sample Size](#)
 Design: 6x3 Williams Cross-Over Design
 Hypotheses: $H_0: P_u - P_v \leq D_{0.L}$ or $P_u - P_v \geq D_{0.U}$ vs. $H_1: D_{0.L} < P_u - P_v < D_{0.U}$ for $u, v = 1, \dots, 3$ with $u \neq v$.
 Number of Possible Tests: 3

Power	Sample Size		Difference			Standard Deviation SD	Alpha*
	Sequence n	Total N	Equivalence		Minimum D1		
			Lower D0.L	Upper D0.U			
0.80023	58	348	-0.3	0.3	0.2	0.75	0.05

* 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.