Chapter 182

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* × *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, …, 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* × *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 binary response from subject *j* (*j* = 1, …, *n*) in sequence *i* (*i* = 1, …, *a*) given treatment *l* (*l* = 1, …, *k*). Assume that the responses y_{ijl} are independent and randomly distributed with $P\left(y_{ijl} = 1\right) = P_l$, 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

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

Test Statistic

For a two-sided test of the hypotheses

$$H_0\colon P_u - P_v = 0 \quad \text{vs.} \quad H_A\colon P_u - P_v \neq 0$$

or equivalently

$$H_0\colon \delta = 0 \quad \text{vs.} \quad H_A\colon \delta \neq 0$$

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

$$Z = \frac{\hat{\delta}}{\frac{\hat{\sigma}_d}{\sqrt{an}}}$$

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

$$\frac{|\hat{\delta}|}{\frac{\hat{\sigma}_d}{\sqrt{an}}} > Z_{1-\alpha/2}$$

where $Z_{1-\alpha/2}$ is the upper $1-\alpha/2$ percentile of the standard normal distribution. One-sided tests reject the null hypothesis at level α if

$$\frac{|\hat{\delta}|}{\frac{\hat{\sigma}_d}{\sqrt{an}}} > Z_{1-\alpha}$$

where $Z_{1-\alpha}$ is the upper $1-\alpha$ 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.

Power Calculation

According to Chow, Shao, Wang, & Lokhnygina (2018) page 89, the power for the two-sided test of $H_0{:}\,\delta = 0$ versus $H_A{:}\,\delta \neq 0$ is

$$\Phi\left(\frac{|\delta_1|}{\frac{\sigma_d}{\sqrt{an}}} - Z_{1-\alpha/2}\right)$$

where $\Phi()$ is the standard normal distribution function, δ_1 is a value of the difference under the alternative hypothesis, and $Z_{1-\alpha/2}$ is the upper $1 - \alpha/2$ percentile of the standard normal distribution. The sample size calculation formula for a two-sided test is

$$n = \text{Ceiling}\left\{\frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \sigma_d^2}{a\delta_1^2}\right\}.$$

The power for a one-sided test is

$$\Phi\left(\frac{|\delta_1|}{\frac{\sigma_d}{\sqrt{an}}} - Z_{1-\alpha}\right)$$

where $Z_{1-\alpha}$ is the upper $1 - \alpha$ percentile of the standard normal distribution. The sample size calculation formula for a one-sided test is

$$n = \text{Ceiling}\left\{\frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2 \sigma_d^2}{a\delta_1^2}\right\}.$$

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 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 30 and 100. The minimum difference to detect is 0.2 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
Alternative Hypothesis ....................................Two-Sided
Alpha......................................................0.05
Adjust Alpha for Multiple Tests ...........................Checked
k (Number of Treatments)...................................3
n (Sample Size per Sequence)...............................30 to 100 by 10
D1 (Minimum Difference|H1) ................................0.2
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
Alternative Hypothesis:     Two-Sided
Hypotheses:                 H0: Pu - Pv = 0   vs.   H1: Pu - Pv ≠ 0 for u, v = 1, ..., 3 with u ≠ v.
Number of Possible Tests:   3
─────────────────────────────────────────────────────────────────────────────────────
               Sample Size
           ───────────────────   Minimum     Standard              Alpha*
            Sequence    Total   Difference   Deviation   ─────────────────────────
Power              n        N           D1          SD   Overall   Individual Test
─────────────────────────────────────────────────────────────────────────────────────
0.61382           30      180          0.2           1      0.05             0.017
0.75941           40      240          0.2           1      0.05             0.017
0.85772           50      300          0.2           1      0.05             0.017
0.91936           60      360          0.2           1      0.05             0.017
0.95588           70      420          0.2           1      0.05             0.017
0.97658           80      480          0.2           1      0.05             0.017
0.98789           90      540          0.2           1      0.05             0.017
0.99388          100      600          0.2           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.
D1 The minimum treatment difference to detect under the alternative hypothesis. D1 = Minimum of $(P_u - P_v)|H1$ for u, v = 1, ..., k with u ≠ 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 different from the others (H0: $P_u - P_v = 0$ versus H1: $P_u - P_v \neq 0$ for u, v = 1, ..., 3 with u ≠ v). Each comparison will be made using a two-sided proportion difference Z-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 0.2 with a sample size of 30 in each sequence (totaling 180 subjects), the power is 0.61382.

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

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.

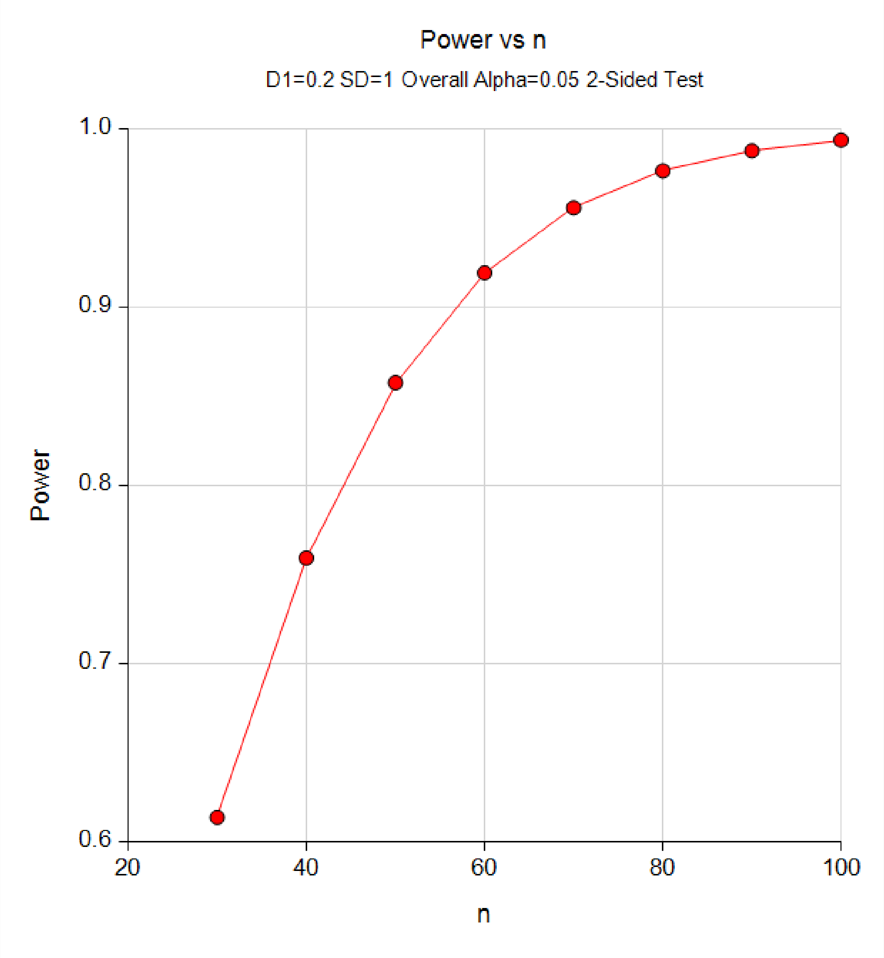
Dropout Summary Statements

Anticipating a 20% dropout rate, group sizes of 38, 38, 38, 38, 38, and 38 subjects should be enrolled to obtain final group sample sizes of 30, 30, 30, 30, 30, and 30 subjects.

References

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

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 over 40 per sequence is required to detect a minimum difference of 0.2 with 80% power.

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

On page 91, 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 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 19 per sequence. Note that there is no adjustment for multiple testing in this example.

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**
Alternative Hypothesis	**Two-Sided**
Power	**0.80**
Alpha	**0.05**
Adjust Alpha for Multiple Tests	**Unchecked**
k (Number of Treatments)	**3**
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
Alternative Hypothesis:    Two-Sided
Hypotheses:                H0: Pu - Pv = 0   vs.   H1: Pu - Pv ≠ 0 for u, v = 1, ..., 3 with u ≠ v.
Number of Possible Tests:  3
───────────────────────────────────────────────────────────────────────────────
              Sample Size
        ─────────────────        Minimum     Standard
          Sequence   Total    Difference    Deviation
Power            n       N            D1           SD    Alpha*
───────────────────────────────────────────────────────────────
0.81253         19     114           0.2         0.75      0.05
───────────────────────────────────────────────────────────────────────────────
* Alpha was not adjusted for multiple tests.
```

The result from **PASS** matches the result from Chow, Shao, Wang, & Lokhnygina (2018) exactly.