Chapter 173

# Equivalence Tests for the Odds Ratio of Two Proportions in a 2x2 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.

A 2×2 cross-over design contains two *sequences* (treatment orderings) and two time periods (occasions). One sequence receives treatment A followed by treatment B. The other sequence receives B and then A. 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. Indeed, higher-order cross-over designs have been used in which the same treatment is used on both occasions.

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 Lui (2016).

### 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 2×2 crossover design may be described as follows. Randomly assign the subjects to one of two sequence groups so that there are $n_1$ subjects in sequence one and $n_2$ subjects in sequence two. In order to achieve design balance, the sample sizes $n_1$ and $n_2$ are assumed to be equal so that $n_1 = n_2 = n = N/2$.

Sequence one is given the control (A) followed by the treatment (B). Sequence two is given the treatment (B) followed by the control (A).

The design can be analyzed using a simple *z*-test if we ignore period and sequence effects or using a more complex random effects logistic regression model that adjusts for period and sequence effects. The sample size calculations herein ignore period and sequence effects. Julious (2010) suggests on page 175 that the bias due to ignoring period effects if a period-adjusted analysis is planned is not great and that sample size calculations that ignore period effects are adequate.

## Cross-Over Design

The discussions that follow summarize the results in Lui (2016). Consider a 2×2 cross-over design and let $x_{ij}^{(g)}$ represent the binary response (0 or 1) from the $j^{th}$ subject, $j$ = 1, ..., $n_g$, in the $i^{th}$ period ($i$ = 1, 2), in sequence $g$ ($g$ = 1, 2). Let $n_{rc}^{(g)}$ represent the number of subjects among $n_g$ subjects in sequence $g$ with the response vector ($x_{1j} = r$, $x_{2j} = c$). We can then summarize the results in terms of counts from a cross-over design with the following table for sequences 1 and 2 as

**SEQUENCE 1 (Control → Treatment)**

| | | Period 2 (Treatment) | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Period 1 (Control) | Yes | $n_{11}^{(1)}$ | $n_{10}^{(1)}$ | $n_{1\cdot}^{(1)}$ |
| | No | $n_{01}^{(1)}$ | $n_{00}^{(1)}$ | $n_{0\cdot}^{(1)}$ |
| | Total | $n_{\cdot 1}^{(1)}$ | $n_{\cdot 0}^{(1)}$ | $n_1$ |

**SEQUENCE 2 (Treatment → Control)**

| | | Period 2 (Control) | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Period 1 (Treatment) | Yes | $n_{11}^{(2)}$ | $n_{10}^{(2)}$ | $n_{1\cdot}^{(2)}$ |
| | No | $n_{01}^{(2)}$ | $n_{00}^{(2)}$ | $n_{0\cdot}^{(2)}$ |
| | Total | $n_{\cdot 1}^{(2)}$ | $n_{\cdot 0}^{(2)}$ | $n_2$ |

In terms of proportions, the 2×2 cross-over design tables can be summarized as

**SEQUENCE 1 (Control → Treatment)**

| | | Period 2 (Treatment) | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Period 1 (Control) | Yes | $P_{11}^{(1)}$ | $P_{10}^{(1)}$ | $P_{1\cdot}^{(1)}$ |
| | No | $P_{01}^{(1)}$ | $P_{00}^{(1)}$ | $P_{0\cdot}^{(1)}$ |
| | Total | $P_{\cdot 1}^{(1)}$ | $P_{\cdot 0}^{(1)}$ | 1 |

**SEQUENCE 2 (Treatment → Control)**

| | | Period 2 (Control) | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Period 1 (Treatment) | Yes | $P_{11}^{(2)}$ | $P_{10}^{(2)}$ | $P_{1\cdot}^{(2)}$ |
| | No | $P_{01}^{(2)}$ | $P_{00}^{(2)}$ | $P_{0\cdot}^{(2)}$ |
| | Total | $P_{\cdot 1}^{(2)}$ | $P_{\cdot 0}^{(2)}$ | 1 |

with the individual proportions estimated as

$$\hat{P}_{rc}^{(g)} = \frac{n_{rc}^{(g)}}{n_g}.$$

Lui (2016) indicates on pages 32-42 that the odds ratio for the treatment versus the control ($O_T/O_C$) is defined for a 2×2 cross-over design as

$$OR = \sqrt{\frac{P_{01}^{(1)} P_{10}^{(2)}}{P_{10}^{(1)} P_{01}^{(2)}}}$$

with estimate

$$\widehat{OR} = \sqrt{\frac{\hat{P}_{01}^{(1)} \hat{P}_{10}^{(2)}}{\hat{P}_{10}^{(1)} \hat{P}_{01}^{(2)}}}.$$

The estimated log odds ratio, $\log\left(\widehat{OR}\right)$, has asymptotic variance $\sigma^2/n$ with

$$\sigma^2 = \frac{1}{4}\left(\frac{1}{P_{01}^{(1)}} + \frac{1}{P_{10}^{(1)}} + \frac{1}{P_{01}^{(2)}} + \frac{1}{P_{10}^{(2)}}\right)$$

which can be estimated as

$$\hat{\sigma}^2 = \frac{1}{4}\left(\frac{1}{\hat{P}_{01}^{(1)}} + \frac{1}{\hat{P}_{10}^{(1)}} + \frac{1}{\hat{P}_{01}^{(2)}} + \frac{1}{\hat{P}_{10}^{(2)}}\right).$$

The standard deviation, then, is

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

with estimate

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

## Equivalence Test Statistics

The null and alternative hypotheses for an equivalence test are

$$H_0: OR \leq OR_{0L} \quad \text{or} \quad OR \geq OR_{0U} \qquad \text{vs.} \qquad H_A: OR_{0L} < OR < OR_{0U}$$

where $OR_{0L}$ and $OR_{0U}$ are the lower and upper equivalence bounds, respectively (i.e., the smallest and largest odds ratios (Ot/Oc) for which the treatment and control will be considered equivalent).

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

$$Z_L = \frac{\log\left(\widehat{OR}\right) - \log(OR_{0L})}{\frac{\widehat{SD}}{\sqrt{n}}} \quad \text{and} \quad Z_U = \frac{\log\left(\widehat{OR}\right) - \log(OR_{0U})}{\frac{\widehat{SD}}{\sqrt{n}}}$$

which are each asymptotically distributed as standard normal under the null hypothesis. The null hypothesis is rejected in favor of the alternative at level $\alpha$ 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 $Z_\alpha$ is the lower $\alpha$ percentile of the standard normal distribution.

## Equivalence Power Calculation

On page 43 of Lui (2016), the power for an equivalence test of $H_0: OR \leq OR_{0L}$ or $OR \geq OR_{0U}$ versus $H_A: OR_{0L} < OR < OR_{0U}$ is given as

$$\Phi\left(\frac{\log(OR_{0U}) - \log(OR_1)}{\frac{SD}{\sqrt{n}}} - Z_{1-\alpha}\right) - \Phi\left(\frac{\log(OR_{0L}) - \log(OR_1)}{\frac{SD}{\sqrt{n}}} + Z_{1-\alpha}\right)$$

where $\Phi()$ is the standard normal distribution function, $OR_1$ is the actual value of the odds ratio 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*.

# Example 1 – Power Analysis

Suppose you want to consider the power of an equivalence test of the hypotheses $H_0: OR \leq 0.667$ or $OR \geq 1.5$ versus $H_A: 0.667 < OR < 1.5$ in a balanced cross-over design with a binary endpoint where the test is computed based on the odds ratio for sequence sample sizes between 100 and 300. The equivalence bounds in this example are log-scale symmetric since 0.667 = 1/1.5. Let's assume that the actual odds ratio is 1 and the estimated standard deviation of the log odds ratio is 2.5. The significance level is 0.05.

## 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** |
| n (Sample Size per Sequence) | **100 to 300 by 50** |
| OR0.U (Upper Equivalence Odds Ratio) | **1.5** |
| OR0.L (Lower Equivalence Odds Ratio) | **1/OR0.U** |
| OR1 (Actual Odds Ratio) | **1** |
| Estimation Method | **Enter SD Directly** |
| Standard Deviation (SD) | **2.5** |

## Output

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

**Numeric Results**

Solve For: Power
Hypotheses: H0: OR ≤ OR0.L or OR ≥ OR0.U vs. H1: OR0.L < OR < OR0.U

| | Sample Size | | Odds Ratio Equivalence Limits | | | | |
|---|---|---|---|---|---|---|---|
| Power | Sequence n | Total N | Lower OR0.L | Upper OR0.U | Actual OR1 | Standard Deviation SD | Alpha |
| 0.00000 | 100 | 200 | 0.667 | 1.5 | 1 | 2.5 | 0.05 |
| 0.26728 | 150 | 300 | 0.667 | 1.5 | 1 | 2.5 | 0.05 |
| 0.48353 | 200 | 400 | 0.667 | 1.5 | 1 | 2.5 | 0.05 |
| 0.64218 | 250 | 500 | 0.667 | 1.5 | 1 | 2.5 | 0.05 |
| 0.75569 | 300 | 600 | 0.667 | 1.5 | 1 | 2.5 | 0.05 |

| | |
|---|---|
| Power | The probability of rejecting a false null hypothesis when the alternative hypothesis is true. |
| n | The sample size in each sequence (or group). |
| N | The total combined sample size from both sequences. |
| OR0.L | The lower equivalence odds ratio used to specify the hypothesis test. |
| OR0.U | The upper equivalence odds ratio used to specify the hypothesis test. |
| OR1 | The actual odds ratio at which power is calculated. |
| SD | The user-entered standard deviation. This is estimated from a previous study. |
| Alpha | The probability of rejecting a true null hypothesis. |

### Summary Statements

A 2×2 cross-over design will be used to test whether the treatment proportion is equivalent to the standard proportion, with equivalence odds ratio (OR = Ot / Oc) bounds of 0.667 and 1.5 (H0: OR ≤ 0.667 or OR ≥ 1.5 versus H1: 0.667 < OR < 1.5). The comparison will be made using two one-sided log odds ratio Z-tests, with an overall Type I error rate (α) of 0.05. The standard deviation of the log odds ratio is assumed to be 2.5. To detect an odds ratio of 1 with a sample size of 100 in each sequence (totaling 200 subjects), the power is 0.

### Dropout-Inflated Sample Size

| | Sample Size | | Dropout-Inflated Enrollment Sample Size | | Expected Number of Dropouts | |
|---|---|---|---|---|---|---|
| **Dropout Rate** | **n** | **N** | **n'** | **N'** | **d** | **D** |
| 20% | 100 | 200 | 125 | 250 | 25 | 50 |
| 20% | 150 | 300 | 188 | 376 | 38 | 76 |
| 20% | 200 | 400 | 250 | 500 | 50 | 100 |
| 20% | 250 | 500 | 313 | 626 | 63 | 126 |
| 20% | 300 | 600 | 375 | 750 | 75 | 150 |

| | |
|---|---|
| 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. |
| n and N | The evaluable group and total sample sizes, respectively, at which power is computed (as entered by the user). If n subjects from each group are evaluated out of the n' subjects that are enrolled in the study, the design will achieve the stated power. N = 2n. |
| n' and N' | The number of subjects that should be enrolled in the study in order to obtain n and N evaluable subjects, based on the assumed dropout rate. n' is calculated by inflating n using the formula n' = n / (1 - DR), with n' 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.). N' = 2n'. |
| d and D | The expected number of group and total dropouts, respectively. d = n' - n and D = 2d. |

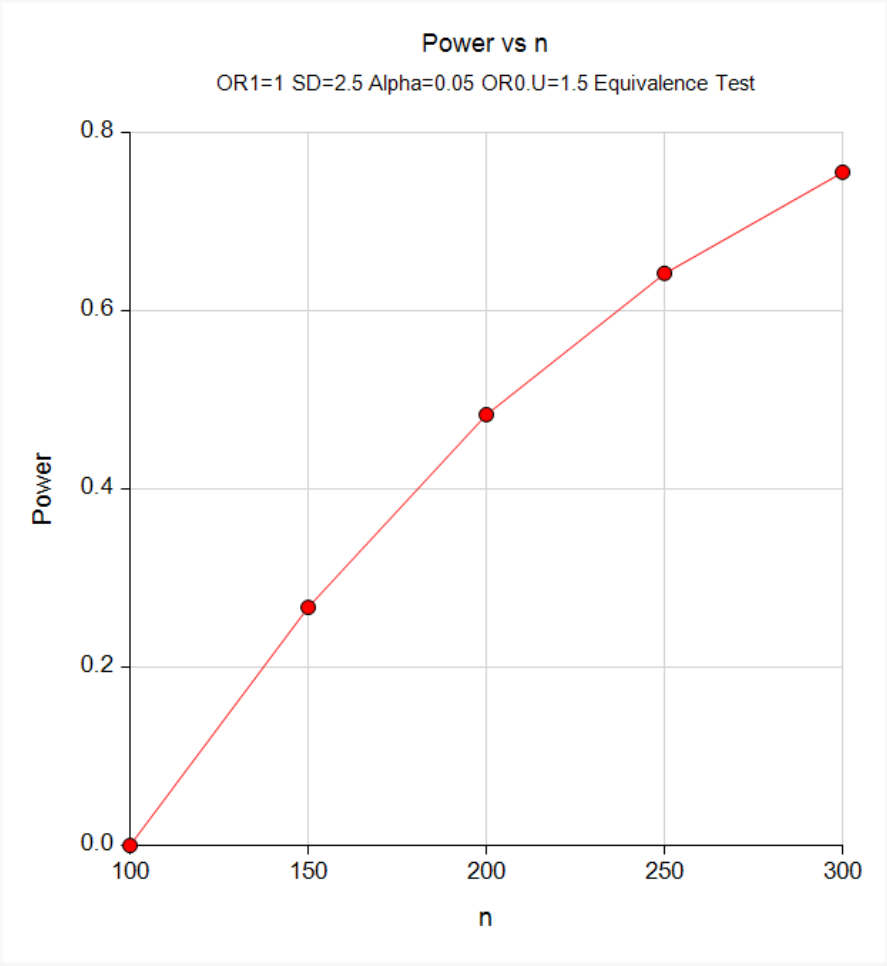
### Dropout Summary Statements

Anticipating a 20% dropout rate, 125 subjects should be enrolled in each group to obtain final sample sizes of 100 subjects per group.

### References

Lui, Kung-Jong. 2016. Crossover Designs: Testing, Estimation, and Sample Size. John Wiley & Sons Ltd. Chichester, West Sussex, England.

**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 greater than 300 per sequence is required for 80% power if the actual odds ratio is 1.

# Example 2 – Calculating Sample Size when Estimating the Standard Deviation from a Previous Study (Validation using Hand Calculations)

This example demonstrates how to calculate the sample size when estimating the standard deviation of the log odds ratio from data in a previous study using the method in Lui (2016) on page 42. In this example we'll find the sample size required to detect an odds ratio of 1 with 80% power at a significance level of 0.05 in an equivalence test of the hypotheses $H_0: OR \leq 0.667$ or $OR \geq 1.5$ versus $H_A: 0.667 < OR < 1.5$. The SD is estimated from discordant proportions in a previous study.

Table 3.2 of Lui (2016) on page 36 presents the following results from 279 subjects a simple 2x2 cross-over trial comparing two inhalation devices, A and B.

**SEQUENCE 1 (Control (A) → Treatment (B))**

| | | Period 2 (B) | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Period 1 (A) | Yes | 26 | 41 | 67 |
| | No | 15 | 57 | 72 |
| | Total | 41 | 98 | 139 |

**SEQUENCE 2 (Treatment (B) → Control (A))**

| | | Period 2 (A) | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Period 1 (B) | Yes | 38 | 16 | 54 |
| | No | 32 | 54 | 86 |
| | Total | 70 | 70 | 140 |

The discordant proportions are estimated as

$$\hat{P}_{01}^{(1)} = \frac{15}{139} = 0.1079$$

$$\hat{P}_{10}^{(1)} = \frac{41}{139} = 0.2950$$

$$\hat{P}_{01}^{(2)} = \frac{32}{140} = 0.2286$$

$$\hat{P}_{10}^{(2)} = \frac{16}{140} = 0.1143$$

Using the method of Lui (2016) on page 42, SD is estimated as

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

$$= \sqrt{\frac{1}{4}\left(\frac{1}{\hat{P}_{01}^{(1)}} + \frac{1}{\hat{P}_{10}^{(1)}} + \frac{1}{\hat{P}_{01}^{(2)}} + \frac{1}{\hat{P}_{10}^{(2)}}\right)}$$

$$= \sqrt{\frac{1}{4}\left(\frac{1}{0.1079} + \frac{1}{0.2950} + \frac{1}{0.2286} + \frac{1}{0.1143}\right)}$$

$$= \mathbf{2.5388}$$

**PASS** will calculate this SD value for you automatically when you input the discordant cell proportions.

Since there is no example given for this calculation in the book, we'll validate this procedure using hand calculations. The power for per-sequence sample sizes of 335 and 336 calculated by hand using the equation on Lui (2016) page 43 is

$$\text{Power} = \Phi\left(\frac{\log(OR_{0U}) - \log(OR_1)}{\frac{SD}{\sqrt{n}}} - Z_{1-\alpha}\right) - \Phi\left(\frac{\log(OR_{0L}) - \log(OR_1)}{\frac{SD}{\sqrt{n}}} + Z_{1-\alpha}\right)$$

$$\text{Power}_{(n=335)} = \Phi\left(\frac{\log(1.5) - \log(1)}{\frac{2.5388}{\sqrt{335}}} - 1.644854\right) - \Phi\left(\frac{\log\left(\frac{1}{1.5}\right) - \log(1)}{\frac{2.5388}{\sqrt{335}}} + 1.644854\right)$$

$$= 0.798846$$

$$\text{Power}_{(n=336)} = \Phi\left(\frac{\log(1.5) - \log(1)}{\frac{2.5388}{\sqrt{336}}} - 1.644854\right) - \Phi\left(\frac{\log\left(\frac{1}{1.5}\right) - \log(1)}{\frac{2.5388}{\sqrt{336}}} + 1.644854\right)$$

$$= 0.800379$$

These results indicate that the minimum required sample size per group is 336, since it is the smallest sample size that achieves the desired 80% power.

## 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** |
| OR0.U (Upper Equivalence Odds Ratio) | **1.5** |
| OR0.L (Lower Equivalence Odds Ratio) | **1/OR0.U** |
| OR1 (Actual Odds Ratio) | **1** |
| Estimation Method | **Use Estimated Discordant Cell Proportions** |
| p01(1) | **0.1079** |
| p10(1) | **0.2950** |
| p01(2) | **0.2286** |
| p10(2) | **0.1143** |

## Output

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

**Numeric Results**

Solve For: Sample Size
Hypotheses: H0: OR ≤ OR0.L or OR ≥ OR0.U vs. H1: OR0.L < OR < OR0.U

| | Sample Size | | Odds Ratio | | | | |
|---|---|---|---|---|---|---|---|
| | | | Equivalence Limits | | | | |
| Power | Sequence n | Total N | Lower OR0.L | Upper OR0.U | Actual OR1 | Standard Deviation SD* | Alpha |
| 0.8004 | 336 | 672 | 0.667 | 1.5 | 1 | 2.539 | 0.05 |

* SD Estimated using Previously Estimated Discordant Cell Proportions:
p01(1) = 0.1079, p10(1) = 0.295, p01(2) = 0.2286, p10(2) = 0.1143

This report indicates that the estimated standard deviation using the method of Lui (2016) is 2.539 and the required sample size is 336 per sequence with 80.04% power. The discordant cell proportions are also listed. These values for SD, *n*, and power computed by **PASS** match our hand calculations above exactly.