Chapter 540

Equivalence Tests for the Difference of Two Means in a Higher-Order Cross-Over Design

Introduction

This procedure calculates power and sample size of statistical tests of equivalence of two means of higher-order cross-over designs when the analysis uses a t-test or equivalent. The parameter of interest is the difference of the two means. Schuirmann's (1987) two one-sided tests (TOST) approach is used to test equivalence. Only a brief introduction to the subject will be given here. For a comprehensive discussion on the subject, refer to Chen, Chow, and Li (1997). The designs covered in this chapter are analyzed using what is called the 'additive model' in Chen et al (1997). The 'multiplicative model' is covered in the procedure that uses ratios.

Measurements are made on individuals that have been randomly assigned to one of several treatment sequences. This *cross-over* design may be analyzed by a TOST equivalence test to show that the two means do not differ by more than a small amount, called the margin of equivalence.

Cross-Over Designs

Senn (2002) defines a *cross-over* design as one in which each subject receives all treatments at least once and the object 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 another. It is assumed that there is a *washout* period between treatments during which the response returns to its baseline value. If this does not occur, there is said to be a *carryover* effect.

A 2x2 cross-over design refers to two treatments (periods) and two *sequences* (treatment orderings). One sequence of treatments is treatment A followed by treatment B. The other sequence is B and then A. The design includes a washout period between responses to make certain that the effects of the first drug do no carryover to the second. Thus, the groups of subjects in this design are defined by the sequence in which the two treatments are administered, not by the treatments they receive.

Higher-Order Cross-Over Designs

Chen et al. (1997) present the results for four cross-over designs that are more complicated than the 2x2 design. Assume that the two treatments are labeled A and B. The available designs are defined by the order and number of times the two treatments are administered.

Balaam's Design

Balaam's design has four sequences with two treatments each. It is popular because it allows the intrasubject variabilities to be estimated. The design is

Sequence	Period 1	Period 2
1	A	A
2	B	B
3	A	B
4	B	A

Two-Sequence Dual Design

This design has two sequences with three periods each. It is popular because it allows the intrasubject variabilities to be estimated. The design is

Sequence	Period 1	Period 2	Period 3
1	A	B	B
2	B	A	A

Four-Period Design with Two Sequences

This design has two sequences of four periods each. The design is

Sequence	Period 1	Period 2	Period 3	Period 4
1	A	B	B	A
2	B	A	A	B

Four-Period Design with Four Sequences

This design has four sequences of four periods each. The design is

Sequence	Period 1	Period 2	Period 3	Period 4
1	A	A	B	B
2	B	B	A	A
1	A	B	B	A
2	B	A	A	B

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 may be easier to obtain because each patient will receive both treatments.

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. In a cross-over experiment, it may be difficult to separate the treatment effect from the time effect and the carry-over effect of the previous treatment.

These cross-over designs cannot be used when the treatment (or the measurement of the response) alters the subject permanently. Hence, it cannot be used to compare treatments that are intended to provide a cure.

Because subjects must be measured at least twice, it may be more difficult to keep patients enrolled in the study. This is particularly true when the measurement process is painful, uncomfortable, embarrassing, or time consuming.

Outline of an Equivalence Test

PASS follows the *two one-sided tests* approach described by Schuirmann (1987) and Phillips (1990). It will be convenient to adopt the following specialized notation for the discussion of these tests.

Parameter	PASS Input/Output	Interpretation
μ_T	Not used	*Treatment mean*. This is the treatment mean.
μ_R	Not used	*Reference mean*. This is the mean of a reference population.
ε	|E|	*Margin of equivalence.* This is a tolerance value that defines the maximum difference that is not of practical importance. This may be thought of as the largest change from the baseline that is considered to be trivial. The absolute value is shown to emphasize that this is a magnitude. The sign of the value will be determined by the specific design that is being used.
δ	D	*True difference*. This is the value of $\mu_T - \mu_R$, the difference between the treatment and reference means. This is the value at which the power is calculated.

Note that the actual values of μ_T and μ_R are not needed. Only their difference is needed for power and sample size calculations.

The null hypothesis of non-equivalence is

$$H_0: \delta \leq \varepsilon_L \quad \text{or} \quad \delta \geq \varepsilon_U, \quad \text{where } \varepsilon_L < 0, \varepsilon_U > 0.$$

The alternative hypothesis of equivalence is

$$H_1: \varepsilon_L < \delta < \varepsilon_U$$

Test Statistics

The analysis for assessing equivalence using higher-order cross-over designs is discussed in detail in Chapter 9 of Chow and Liu (1999). Unfortunately, their presentation is too lengthy to give here. Their method involves the computation of an analysis of variance to estimate the error variance. It also describes the construction of confidence limits for appropriate contrasts. These confidence limits can then be compared to the equivalence limits to test for equivalence. We refer you to their book for details.

Power Calculation

The power is given by

$$Power(\delta) = T_V\left(\frac{\varepsilon_U - \delta}{\sigma_W\sqrt{b/n}} - t_{V,1-\alpha}\right) - T_V\left(t_{V,1-\alpha} - \frac{\delta - \varepsilon_L}{\sigma_W\sqrt{b/n}}\right)$$

where T represents the cumulative t distribution, V and b depend on the design, σ_W is the square root of the within mean square error from the ANOVA table using data in the original scale used to analyze the cross-over design, and n is the average number of subjects per sequence. The constants V and b depend on the design as follows.

Design Type	**Parameters (*V*,*b*)**
Balaam's Design	V = 4n – 3, b = 2.
Two-Sequence Dual Design	V = 4n – 4, b = 3/4.
Four-Period Design with Two Sequences	V = 6n – 5, b = 11/20.
Four-Period Design with Four Sequences	V = 12n – 5, b = 1/4.

The presentation of Chen et al (1997) uses the following, different parameterization.

$$Power(\theta) = T_V\left(\frac{\nabla_U - \theta}{CV\sqrt{b/n}} - t_{V,1-\alpha}\right) - T_V\left(t_{V,1-\alpha} - \frac{\theta - \nabla_L}{CV\sqrt{b/n}}\right)$$

where $\theta = \frac{\mu_T - \mu_R}{\mu_R}$, $CV = \frac{\sigma_W}{\mu_R}$, $\nabla_L = \frac{\varepsilon_L}{\mu_R}$, and $\nabla_U = \frac{\varepsilon_U}{\mu_R}$. This parameterization has the advantage that the variables are scaled by the reference mean, so all you need to know is their relative magnitudes rather than their absolute values. It turns out that you can use either parameterization as input, as long as you are consistent.

Example 1 – Finding Power

A two-sequence, dual cross-over design is to be used to compare the impact of two drugs on diastolic blood pressure. The average diastolic blood pressure after administration of the reference drug is 96 mmHg. Researchers believe this average may drop to 92 mmHg with the use of a new drug. The within mean square error found from similar studies is 324. Its square root is 18.

Following FDA guidelines, the researchers want to show that the diastolic blood pressure is within 20% of the diastolic blood pressure of the reference drug. Thus, the equivalence limits of the mean difference of the two drugs are -19.2 and 19.2. They decide to calculate the power for a range of sample sizes between 4 and 40. 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**
Design Type	**3x2 (Three-Period, Two-Sequence Dual: ABB|BAA)**
Alpha	**0.05**
N (Total Sample Size)	**4 6 8 10 12 14 16 18 20 30 40**
EU (Upper Equivalence Limit)	**19.2**
EL (Lower Equivalence Limit)	**-Upper Limit**
D (Difference, μT - μR)	**-4**
Specify σ as σw or σb and ρ	**σw (Within Std Dev)**
σw (Within Std Dev)	**18**

Output

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

Numeric Reports

Numeric Results

Solve For:	Power
Design Type:	Three-Period, Two-Sequence Dual
Treatment Sequences:	ABB | BAA
Difference:	D = μT - μR = Treatment Mean - Reference Mean

Power	Sample Size N	Equivalence Limits Lower EL	Equivalence Limits Upper EU	Actual Difference D	Within Standard Deviation σw	Alpha
0.0000	4	-19.2	19.2	-4	18	0.05
0.1878	6	-19.2	19.2	-4	18	0.05
0.4375	8	-19.2	19.2	-4	18	0.05
0.5985	10	-19.2	19.2	-4	18	0.05
0.7082	12	-19.2	19.2	-4	18	0.05
0.7855	14	-19.2	19.2	-4	18	0.05
0.8411	16	-19.2	19.2	-4	18	0.05
0.8818	18	-19.2	19.2	-4	18	0.05
0.9119	20	-19.2	19.2	-4	18	0.05
0.9800	30	-19.2	19.2	-4	18	0.05
0.9957	40	-19.2	19.2	-4	18	0.05

Power	The probability of rejecting non-equivalence when the means are equivalent.
N	The total number of subjects. They are divided evenly among all sequences.
EU and EL	The upper and lower limits of the maximum allowable difference that results in equivalence.
μT	The treatment mean. It is usually associated with the letter "A" in the design.
μR	The reference mean. It is usually associated with the letter "B" in the design.
D	The difference between the means at which the power is computed. D = μT - μR.
σw	The square root of the within mean square error from the ANOVA table.
Alpha	The probability of falsely rejecting H0 (falsely concluding superiority).

Summary Statements

A three-period, two-sequence dual cross-over design (ABB | BAA) will be used to test whether the treatment mean (μ_T) is equivalent to the reference mean (μ_R), with mean difference equivalence limits of -19.2 and 19.2 (H0: D ≤ -19.2 or D ≥ 19.2 versus H1: -19.2 < D < 19.2, D = μ_T - μ_R). The comparison will be made using two one-sided t-tests, with an overall Type I error rate (α) of 0.05. The within-subject standard deviation is assumed to be 18. To detect a difference in means (μ_T - μ_R) of -4, with a total sample size of 4 (allocated equally to the 2 sequences), the power is 0.

Dropout-Inflated Sample Size

Dropout Rate	Sample Size N	Dropout-Inflated Enrollment Sample Size N'	Expected Number of Dropouts D
20%	4	5	1
20%	6	8	2
20%	8	10	2
20%	10	13	3
20%	12	15	3
20%	14	18	4
20%	16	20	4
20%	18	23	5
20%	20	25	5
20%	30	38	8
20%	40	50	10

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	The evaluable sample size at which power is computed (as entered by the user). If N subjects are evaluated out of the N' subjects that are enrolled in the study, the design will achieve the stated power.
N'	The total number of subjects that should be enrolled in the study in order to obtain 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.)
D	The expected number of dropouts. D = N' - N.

Dropout Summary Statements

Anticipating a 20% dropout rate, 5 subjects should be enrolled to obtain a final sample size of 4 subjects.

References

Chow, S.C. and Liu, J.P. 1999. Design and Analysis of Bioavailability and Bioequivalence Studies. Marcel Dekker. New York

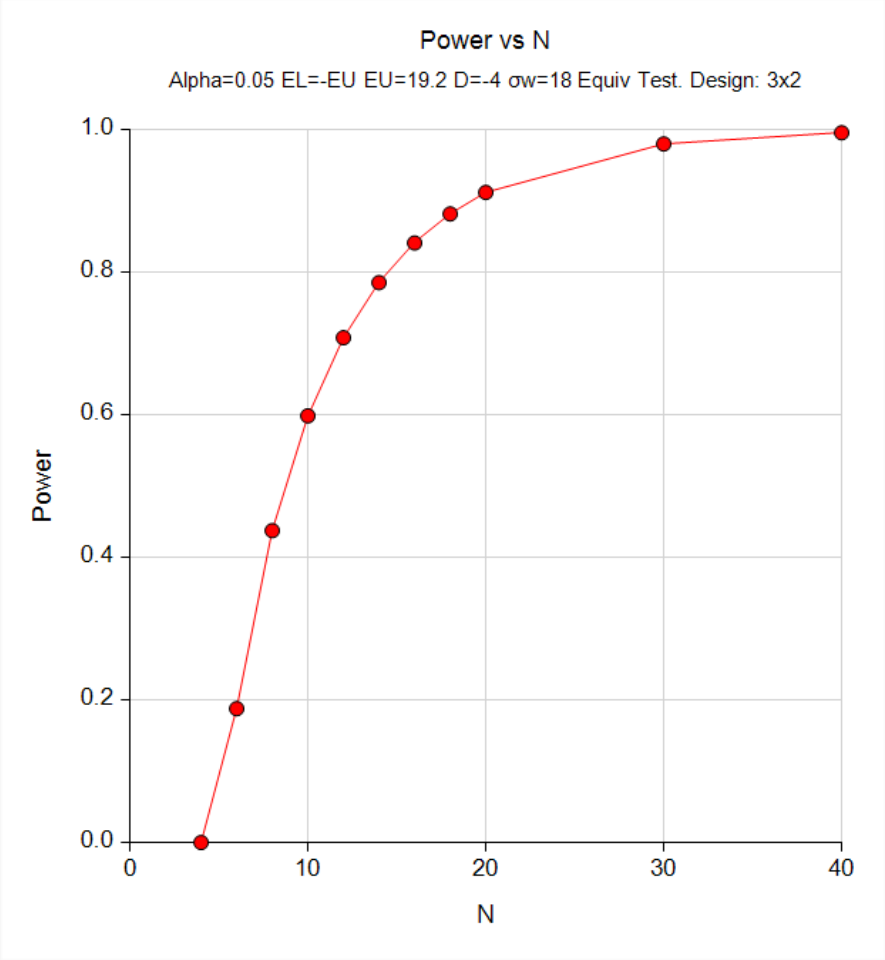
Chow, S.C., Shao, J., and Wang, H. 2003. Sample Size Calculations in Clinical Research. Marcel Dekker. New York.

Chen, K.W.; Chow, S.C.; and Li, G. 1997. 'A Note on Sample Size Determination for Bioequivalence Studies with Higher-Order Crossover Designs.' Journal of Pharmacokinetics and Biopharmaceutics, Volume 25, No. 6, pages 753-765.

This report shows the power for the indicated scenarios.

Plots Section

Plots



This plot shows the power versus the sample size.

Example 2 – Finding Sample Size

Continuing with Example 1, the researchers want to find the exact sample size needed to achieve both 80% and 90% 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 (Exact)**
Design Type	**3x2 (Three-Period, Two-Sequence Dual: ABB|BAA)**
Power	**0.80 0.90**
Alpha	**0.05**
EU (Upper Equivalence Limit)	**19.2**
EL (Lower Equivalence Limit)	**-Upper Limit**
D (Difference)	**-4**
Specify σ as σw or σb and ρ	**σw (Within Std Dev)**
σw (Within Std Dev)	**18**

Output

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

Numeric Results

Solve For:	Sample Size (Exact)
Design Type:	Three-Period, Two-Sequence Dual
Treatment Sequences:	ABB | BAA
Difference:	D = μT - μR = Treatment Mean - Reference Mean

Power	Sample Size N	Equivalence Limits Lower EL	Equivalence Limits Upper EU	Actual Difference D	Within Standard Deviation σw	Alpha
0.8155	15	-19.2	19.2	-4	18	0.05
0.9119	20	-19.2	19.2	-4	18	0.05

Twenty subjects are needed to achieve at least 90% power and fifteen subjects are needed to achieve at least 80% power.

Example 3 – Validation using Chen et al. (1997)

Chen et al. (1997) page 757 present a table of sample sizes for various parameter values. In this table, the treatment mean, standard deviation, and equivalence limits are all specified as percentages of the reference mean. We will reproduce the seventeenth line of the table in which the square root of the within mean square error is 10%, the equivalence limits are 20%, the difference between the means is 0%, 5%, 10%, and 15%, the power is 90%, and the significance level is 0.05. Chen reports total sample sizes of 24, 36, 72, and 276. We will now setup this example in **PASS**.

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 3** 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 (Equal Per Sequence)**
Design Type	**2x4 (Balaam: AA|BB|AB|BA)**
Power	**0.90**
Alpha	**0.05**
EU (Upper Equivalence Limit)	**0.2**
EL (Lower Equivalence Limit)	**-Upper Limit**
D (Difference)	**0 0.05 0.10 0.15**
Specify σ as σw or σb and ρ	**σw (Within Std Dev)**
σw (Within Std Dev)	**0.1**

Output

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

Numeric Results

Solve For:	Sample Size (Equal Per Sequence)
Design Type:	Two-Period, Four-Sequence (Balaam)
Treatment Sequences:	AA | BB | AB | BA
Difference:	D = μT - μR = Treatment Mean - Reference Mean

		Equivalence Limits				
Power	Sample Size N	Lower EL	Upper EU	Actual Difference D	Within Standard Deviation σw	Alpha
0.9041	24	-0.2	0.2	0.00	0.1	0.05
0.9266	36	-0.2	0.2	0.05	0.1	0.05
0.9065	72	-0.2	0.2	0.10	0.1	0.05
0.9003	276	-0.2	0.2	0.15	0.1	0.05

PASS obtains the same samples sizes as Chen et al. (1997).