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

$$H_0 : \delta \le \varepsilon_L \text{ or } \delta \ge \varepsilon_U \text{ where } \varepsilon_L < 0, \varepsilon_U > 0.$$

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 <u>in the original scale</u> 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 $\theta = \dfrac{\mu_T - \mu_R}{\mu_R}$, $CV = \dfrac{\sigma_W}{\mu_R}$, $\nabla_L = \dfrac{\varepsilon_L}{\mu_R}$, and $\nabla_U = \dfrac{\varepsilon_U}{\mu_R}$. 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.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Designa tab. For more information about the options of other tabs, go to the Procedure Window chapter.

Design Tab

The Design tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Solve For

Solve For

This option specifies the parameter to be solved for from the other parameters. Under most situations, you will select either *Power* for a power analysis or *Sample Size* for sample size determination.

Select *Sample Size* when you want to calculate the sample size needed to achieve a given power and alpha level. Note that there are two choices for finding *Sample Size*. Select *Sample Size (Equal Per Sequence)* when you want the design to have an equal number of subjects per sequence. Select *Sample Size (Exact)* when you want to find the exact sample size even though the number of subjects cannot be dividing equally among the sequences.

Select *Power* when you want to calculate the power of an experiment.

Test

Design Type

Specify which cross-over design that you want to analyze. These designs allow you to compare two means: a treatment (A) and a reference (B). The designs assume that each subject is measured two or more times (or periods) with ample time in between to wash-out the effect of the treatment.

A sequence is an ordering of the how A and B are applied. For example, 'ABA' means that the subjects are measured three times: first treatment A, then B, then A again. Other possible sequences are BAB, AAB, and BBA. Each design includes several sequences.

The design tries to balance the experiment so that each treatment occurs an equal number of times and after each other treatment an equal number of times.

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.

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.

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*.

Sample Size

N (Total Sample Size)

This option specifies one or more values of the sample size, the number of individuals in the study (total subjects in all sequences). This value must be an integer greater than one.

You may enter a list of values using the syntax *50,100,150,200,250* or *50 to 250 by 50*.

Effect Size – Equivalence Limits

|EU| (Upper Equivalence Limit)

This value gives the upper limit of equivalence. Differences outside EL and EU are not considered equivalent, while differences between them are.

Note that EL must be less than zero and EU must be greater than zero. Also, D, EL, and EU must satisfy EL<D<EU. Finally, the scale of these numbers must match the scale of Sw.

Alternatively, you can enter a value for $\nabla_U = \dfrac{E_U}{\mu_R}$ here as long as you enter similarly scaled values for the other

effect size parameters.

-|EL| (Lower Equivalence Limit)

This value gives lower limit on equivalence. Differences outside EL and EU are not considered equivalent, while differences between them are.

If you want symmetric limits, enter -UPPER LIMIT for EL to force EL = -|EU|.

Note that EL must be less than zero and EU must be greater than zero. Also, D, EL, and EU must satisfy EL<D<EU. Finally, the scale of these numbers must match the scale of Sw.

Alternatively, you can enter a value for $\nabla_L = \dfrac{E_L}{\mu_R}$ here as long as you enter similarly scaled values for the other effect size parameters.

Effect Size – Mean Difference

D (Difference, μT - μR)

This is the difference, $\delta = \mu_T - \mu_R$, between the two means at which the power is to be computed. Often this value is set to zero, but it can be non-zero as long as it is between the equivalence limits, EL and EU. That is, D, EL, and EU must satisfy EL < D < EU. The scale of these numbers must match the scale of σw.

Alternative, you can enter $\theta = \dfrac{\mu_T - \mu_R}{\mu_R}$ as long as you enter similarly scaled values for the other effect size parameters.

Effect Size – Standard Deviation

Specify σ as σw or σb and ρ

This option specifies how you will enter the within standard deviation, σw. You have two choices:

σw

Specify σw directly.

σb and ρ

Specify σb and ρ separately. The value of σw is calculated using the formula: σw = σb √(1 - ρ).

σw (Within Std Dev)

Specify the value(s) of the standard deviation σw. This value can be estimated using σw = √(WMSE) where WMSE is the within mean square error from the ANOVA table used to analyze the Cross-Over design.

If you were to compute the treatment difference for each subject, this would be the standard deviation of those differences. These values must be positive.

A list of values may be entered.

You can press the **σ button** to load the Standard Deviation Estimator window.

σb (Between Std Dev)

This is the between subject standard deviation of the response variable (Y) at a particular time point. It is assumed to be the same for all time points. This represents the variability from subject to subject that occurs when the subjects are treated identically.

As a standard deviation, the number(s) must be greater than zero.

You can enter a list of values separated by blanks or commas, in which case, a separate analysis will be calculated for each value.

You can press the σ button and select 'Covariance Matrix' to obtain help on estimating the standard deviation from an ANOVA table.

ρ (Correlation)

Specify the correlation between two observations made on the same subject. Negative values are not permitted, so it must be between 0 and 1. When no previous information is available, try values between 0.2 and 0.6.

You can press the **σ button** and select 'Covariance Matrix' to obtain help on estimating the standard deviation from an ANOVA table.

You can enter a single value or a list of values.

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

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 the Difference of Two Means in a Higher-Order Cross-Over Design** procedure window by expanding **Means**, then **Cross-Over (Higher-Order) Design**, then clicking on **Equivalence**, and then clicking on **Equivalence Tests for the Difference of Two Means in a Higher-Order 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**
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)	**-4**
Specify σ as σw or σb and ρ...................	**σw (Within Std Dev)**
σw (Within Std Dev)................................	**18**

Annotated Output

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

Numeric Results

Numeric Results for an Equivalence Test of the Mean Difference in a Three-Period, Two-Sequence Dual Design
Treatment Sequences: ABB | BAA

Power	Sample Size N	Lower Equiv Limit EL	Upper Equiv Limit EU	Mean Difference D (μT-μR)	Within Standard Deviation σw	Alpha
0.0000	4	-19.200	19.200	-4.000	18.0000	0.050
0.1878	6	-19.200	19.200	-4.000	18.0000	0.050
0.4375	8	-19.200	19.200	-4.000	18.0000	0.050
0.5985	10	-19.200	19.200	-4.000	18.0000	0.050
0.7082	12	-19.200	19.200	-4.000	18.0000	0.050
0.7855	14	-19.200	19.200	-4.000	18.0000	0.050
0.8411	16	-19.200	19.200	-4.000	18.0000	0.050
0.8818	18	-19.200	19.200	-4.000	18.0000	0.050
0.9119	20	-19.200	19.200	-4.000	18.0000	0.050
0.9800	30	-19.200	19.200	-4.000	18.0000	0.050
0.9957	40	-19.200	19.200	-4.000	18.0000	0.050

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.; 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.

Report Definitions
Power is the probability of rejecting non-equivalence when the means are equivalent.
N is the total number of subjects. They are divided evenly among all sequences.
EU & EL are the upper & lower limits of the maximum allowable difference that results in equivalence.
μT is the treatment mean. It is usually associated with the letter 'A' in the design.
μR is the reference mean. It is usually associated with the letter 'B' in the design.
D ($\mu T - \mu R$) is the difference between the means at which the power is computed.
σw is the square root of the within mean square error from the ANOVA table.
Alpha is the probability of falsely rejecting H0 (falsely concluding superiority).

Summary Statements
In an equivalence test of means using two one-sided tests on data drawn from a three-period,
two-sequence dual cross-over design, a total sample size of 4 achieves 0% power at a 5%
significance level when the true difference between the means is -4.000, the equivalence limits
are -19.200 and 19.200, and the within-subject standard deviation is 18.0000.

This report shows the power for the indicated scenarios.

Plots Section
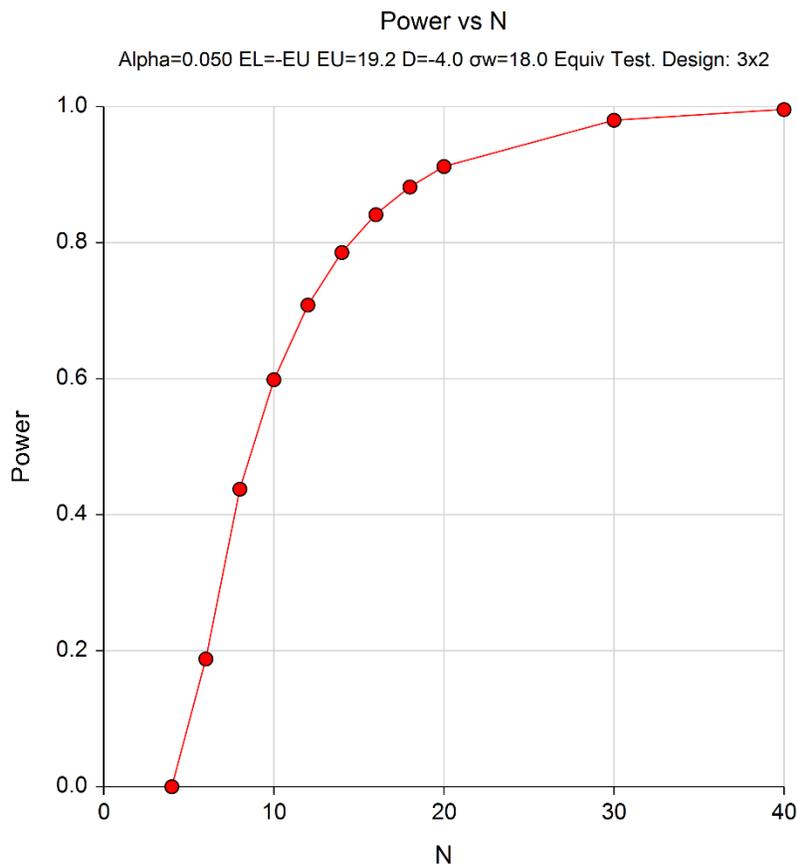
This plot shows the power versus the sample size.

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

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 the Difference of Two Means in a Higher-Order Cross-Over Design** procedure window by expanding **Means**, then **Cross-Over (Higher-Order) Design**, then clicking on **Equivalence**, and then clicking on **Equivalence Tests for the Difference of Two Means in a Higher-Order 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 (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

Numeric Results for an Equivalence Test of the Mean Difference in a Three-Period, Two-Sequence Dual Design
Treatment Sequences: ABB | BAA

Power	Sample Size N	Lower Equiv Limit EL	Upper Equiv Limit EU	Mean Difference D (μT-μR)	Within Standard Deviation σw	Alpha
0.8155	15	-19.200	19.200	-4.000	18.0000	0.050
0.9119	20	-19.200	19.200	-4.000	18.0000	0.050

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

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 the Difference of Two Means in a Higher-Order Cross-Over Design** procedure window by expanding **Means**, then **Cross-Over (Higher-Order) Design**, then clicking on **Equivalence**, and then clicking on **Equivalence Tests for the Difference of Two Means in a Higher-Order Cross-Over Design**. You may then make the appropriate entries as listed below, or open **Example 3** by going to the **File** menu and choosing **Open Example Template**.

Option   **Value**

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

Numeric Results for an Equivalence Test of the Mean Difference in a Two-Period, Four-Sequence (Balaam) Design
Treatment Sequences: AA | BB | AB | BA

Power	Sample Size N	Lower Equiv Limit EL	Upper Equiv Limit EU	Mean Difference D (μT-μR)	Within Standard Deviation σw	Alpha
0.9041	24	-0.200	0.200	0.000	0.1000	0.050
0.9266	36	-0.200	0.200	0.050	0.1000	0.050
0.9065	72	-0.200	0.200	0.100	0.1000	0.050
0.9003	276	-0.200	0.200	0.150	0.1000	0.050

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