

## Chapter 520

# Equivalence Tests for the Difference Between Two Means in a 2x2 Cross-Over Design

---

### Introduction

This procedure calculates power and sample size of statistical tests of equivalence of the means of a 2x2 cross-over design which is analyzed with a t-test. 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 Chow and Liu (1999) and Julious (2004).

Measurements are made on individuals that have been randomly assigned to one of two sequences. The first sequence receives the treatment followed by the reference (AB). The second sequence receives the reference followed by the treatment (BA). 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.

The definition of equivalence has been refined in recent years using the concepts of prescribability and switchability. *Prescribability* refers to ability of a physician to prescribe either of two drugs at the beginning of the treatment. However, once prescribed, no other drug can be substituted for it. *Switchability* refers to the ability of a patient to switch from one drug to another during treatment without adverse effects. Prescribability is associated with equivalence of location and variability. Switchability is associated with the concept of individual equivalence. This procedure analyzes average equivalence. Thus, it partially analyzes prescribability. It does not address equivalence of variability or switchability.

---

### Cross-Over Designs

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 *carryover* effect.

A 2x2 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

## Equivalence Tests for the Difference Between Two Means in a 2x2 Cross-Over Design

by the treatments they receive. Indeed, higher-order cross-over designs have been used in which the same treatment is used at 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 of these cases, symptoms are expected to return to their usual baseline level shortly after the treatment is stopped.

---

### 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.

---

### 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.

<u>Parameter</u>	<u>PASS Input/Output</u>	<u>Interpretation</u>
$\mu_T$	Not used	<i>Treatment mean.</i> This is the treatment mean.
$\mu_R$	Not used	<i>Reference mean.</i> This is the mean of a reference population.
$\varepsilon$	E	<i>Margin of equivalence.</i> This is a tolerance value that defines the maximum change 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.
$\delta$	D	<i>True difference.</i> 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  $\mu_T$  and  $\mu_R$  are not needed. Only their difference is needed for power and sample size calculations.

## Equivalence Tests for the Difference Between Two Means in a 2x2 Cross-Over Design

The null hypothesis of non-equivalence is

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

and the alternative hypothesis of equivalence is

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

### Test Statistics

This section describes the test statistic that is used to perform the hypothesis test.

#### T-Test

A t-test is used to analyze the data. The test assumes that the data are a simple random sample from a population of normally-distributed values that have the same variance. This assumption implies that the differences are continuous and normal. The calculation of the two, one-sided t-tests proceeds as follow

$$T_L = \frac{(\bar{x}_T - \bar{x}_R) - \varepsilon_L}{\hat{\sigma}_w \sqrt{\frac{2}{N}}} \text{ and } T_U = \frac{(\bar{x}_T - \bar{x}_R) - \varepsilon_U}{\hat{\sigma}_w \sqrt{\frac{2}{N}}}$$

where  $\hat{\sigma}_w^2$  is the within mean square error from the appropriate ANOVA table.

The significance of each test statistic is determined by computing the p-value. If this p-value is less than a specified level (usually 0.05), the null hypothesis is rejected.

If prior studies used a t-test rather than an ANOVA to analyze the data, you may not have a direct estimate of  $\sigma_w^2$ . Instead, you will have an estimate of the variance of the period differences from the t-test,  $\hat{\sigma}_d^2$ . These variances are functionally related by  $\sigma_w^2 = 2\sigma_d^2$ . Either variance can be entered.

### Power Calculation

The power of this test is given by

$$\Pr(T_L \geq t_{1-\alpha, N-2} \text{ and } T_U \leq -t_{1-\alpha, N-2})$$

where  $T_L$  and  $T_U$  are distributed as the bivariate, noncentral  $t$  distribution with noncentrality parameters  $\Delta_L$  and  $\Delta_U$  given by

$$\Delta_L = \frac{\delta - \varepsilon_L}{\sigma_w \sqrt{\frac{2}{N}}}$$

$$\Delta_U = \frac{\delta - \varepsilon_U}{\sigma_w \sqrt{\frac{2}{N}}}$$

## Procedure Options

This section describes the options that are specific to this procedure. These are located on the Design 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.

Select *Power* when you want to calculate the power of an experiment that has already been run.

---

#### 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. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of unequal means when in fact the means are equal.

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. In this procedure, a type-I error occurs when you reject the null hypothesis of unequal means when in fact the means are unequal.

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 both sequences). This value must be an integer greater than one.

When N is even, it is split evenly between the two sequences. When N is odd, the first sequence has one more subject than the second sequence.

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

## Equivalence Tests for the Difference Between Two Means in a 2x2 Cross-Over Design

---

### Effect Size – Equivalence Limits

#### |EU| (Upper Equivalence Limit)

This value gives upper limit on equivalence. Differences outside EL and EU are not considered equivalent. Differences between them are considered equivalent.

Note that  $EL < 0$  and  $EU > 0$ . Also, you must have  $EL < D < EU$ .

#### -|EL| (Lower Equivalence Limit)

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

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

Note that  $EL < 0$  and  $EU > 0$ . Also, you must have  $EL < D < EU$ . Finally, the scale of these numbers must match the scale of S.

---

### Effect Size – True Mean Difference

#### D (True Difference)

This is the true difference 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.

---

### Effect Size – Standard Deviation

#### Specify S as Sw or Sd

Specify the form of the standard deviation that is entered in the box below.

- **Sw**  
Specify S as the square root of the within mean square error from a repeated measures ANOVA. This is the most common method since cross-over designs are usually analyzed using ANOVA.
- **Sd**  
Specify S as the standard deviation of the individual treatment differences computed for each subject. This option is used when you have previous studies that produced this value.

#### S (Value of Sw or Sd)

Specify the value(s) of the standard deviation S. The interpretation of this value depends on the entry in *Specify S as Sw or Sd* above. If  $S = Sw$  is selected, this is the value of Sw which is  $SQR(WMSE)$  where WMSE is the within mean square error from the ANOVA table used to analyze the Cross-Over design. If  $S = Sd$  is selected, this is the value of Sd which is the standard deviation of the period differences—pooled from both sequences.

These values must be positive. A list of values may be entered.

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

## Equivalence Tests for the Difference Between Two Means in a 2x2 Cross-Over Design

### Example 1 – Finding Power

A 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 known to be 96 mmHg. Researchers believe this average may drop to 92 mmHg with the use of a new drug. The within mean square error of similar studies is 324. Its square root is 18.

Following FDA guidelines, the researchers want to show that the diastolic blood pressure with the new drug is within 20% of the diastolic blood pressure with 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 6 and 100. 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 Between Two Means in a 2x2 Cross-Over Design** procedure window by expanding **Means**, then **Cross-Over (2x2) Design**, then clicking on **Equivalence**, and then clicking on **Equivalence Tests for the Difference Between Two Means in a 2x2 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**.

<u>Option</u>	<u>Value</u>
<b>Design Tab</b>	
Find .....	<b>Power</b>
Alpha .....	<b>0.05</b>
N (Total Sample Size) .....	<b>6 10 16 20 40 60 80 100</b>
EU  (Upper Equivalence Limit) .....	<b>19.2</b>
- EL  (Lower Equivalence Limit) .....	<b>-Upper Limit</b>
D (True Difference) .....	<b>-4</b>
Specify S as .....	<b>Sw</b>
S .....	<b>18</b>

### Annotated Output

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

### Numeric Results

Numeric Results for Testing Equivalence Using a Cross-Over Design

Power	Total Sample Size (N)	Lower Equiv. Limit	Upper Equiv. Limit	True Difference	Standard Deviation Sw	Alpha	Beta
0.1470	6	-19.20	19.20	-4.00	18.00	0.0500	0.8530
0.3873	10	-19.20	19.20	-4.00	18.00	0.0500	0.6127
0.6997	16	-19.20	19.20	-4.00	18.00	0.0500	0.3003
0.8104	20	-19.20	19.20	-4.00	18.00	0.0500	0.1896
0.9804	40	-19.20	19.20	-4.00	18.00	0.0500	0.0196
0.9983	60	-19.20	19.20	-4.00	18.00	0.0500	0.0017
0.9999	80	-19.20	19.20	-4.00	18.00	0.0500	0.0001
1.0000	100	-19.20	19.20	-4.00	18.00	0.0500	0.0000

## Equivalence Tests for the Difference Between Two Means in a 2x2 Cross-Over Design

**Report Definitions**

Power is the probability of rejecting non-equivalence when the means are equivalent.

N is the total number of subjects split between both sequences.

EU & EL are the maximum allowable differences that still result in equivalence.

D is the difference between the means at which the power is computed.

Sw is the square root of the within mean square error from the ANOVA table.

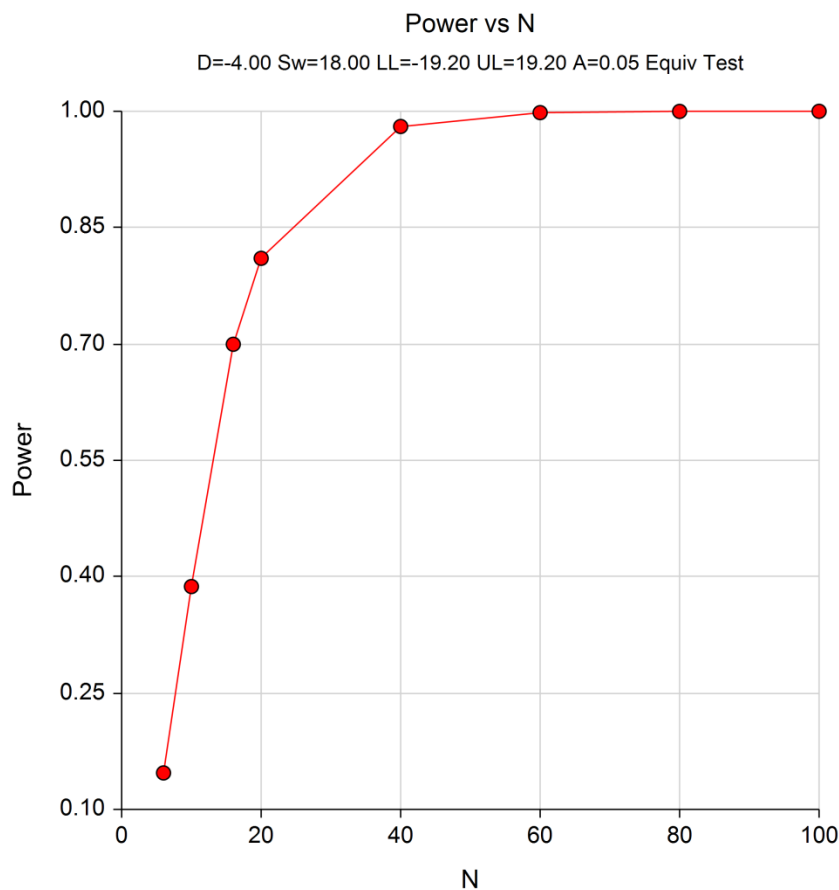
Alpha is the probability of rejecting non-equivalence when the means are non-equivalent.

Beta is the probability of accepting non-equivalence when the means are equivalent.

**Summary Statements**

In an equivalence test of means using two one-sided tests on data from a two-period cross-over design, a total sample size of 6 achieves 15% power at a 5% significance level when the true difference between the means is -4.00, the square root of the within mean square error is 18.00, and the equivalence limits are -19.20 and 19.20.

This report shows the power for the indicated scenarios. Note that if they want 90% power, they will require a sample of around 30 subjects.

**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% power 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 Between Two Means in a 2x2 Cross-Over Design** procedure window by expanding **Means**, then **Cross-Over (2x2) Design**, then clicking on **Equivalence**, and then clicking on **Equivalence Tests for the Difference Between Two Means in a 2x2 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**.

<u>Option</u>	<u>Value</u>
<b>Design Tab</b>	
Solve For .....	<b>Sample size</b>
Power.....	<b>0.8 0.9</b>
Alpha.....	<b>0.05</b>
EU  (Upper Equivalence Limit) .....	<b>19.2</b>
- EL  (Lower Equivalence Limit).....	<b>-Upper Limit</b>
D (True Difference) .....	<b>-4</b>
Specify S as.....	<b>Sw</b>
S.....	<b>18</b>

### Output

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

#### Numeric Results

Numeric Results for Testing Equivalence Using a Cross-Over Design

	Total Sample Size (N)	Lower Equiv. Limit	Upper Equiv. Limit	True Difference	Standard Deviation Sw	Alpha	Beta
Power	20	-19.20	19.20	-4.00	18.00	0.0500	0.1896
	26	-19.20	19.20	-4.00	18.00	0.0500	0.0968

We note that 20 subjects are needed to achieve 80% power and 26 subjects are needed to achieve 90% power.



## Example 3 – Validation using Phillips (1990)

Phillips (1990) page 142 presents 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 second line of the table in which the square root of the within mean square error is 20%; the equivalence limits are 20%; the treatment mean is 100%, 95%, 90%, and 85%; the power is 70%; and the significance level is 0.05. Phillips reports total sample size as 16, 20, 40, and 152 corresponding to the four treatment mean percentages. 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 Between Two Means in a 2x2 Cross-Over Design** procedure window by expanding **Means**, then **Cross-Over (2x2) Design**, then clicking on **Equivalence**, and then clicking on **Equivalence Tests for the Difference Between Two Means in a 2x2 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**.

<u>Option</u>	<u>Value</u>
<b>Design Tab</b>	
Solve For .....	<b>Sample Size</b>
Power.....	<b>0.7</b>
Alpha.....	<b>0.05</b>
EU  (Upper Equivalence Limit) .....	<b>20</b>
- EL  (Lower Equivalence Limit).....	<b>-Upper Limit</b>
D (True Difference) .....	<b>0 -5 -10 -15</b>
Specify S as.....	<b>Sw</b>
S.....	<b>20</b>

### Output

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

### Numeric Results

Numeric Results for Testing Equivalence Using a Cross-Over Design

Power	Total Sample Size (N)	Lower Equiv. Limit	Upper Equiv. Limit	True Difference	Standard Deviation Sw	Alpha	Beta
0.7001	152	-20.00	20.00	-15.00	20.00	0.0500	0.2999
0.7092	40	-20.00	20.00	-10.00	20.00	0.0500	0.2908
0.7221	20	-20.00	20.00	-5.00	20.00	0.0500	0.2779
0.7031	16	-20.00	20.00	0.00	20.00	0.0500	0.2969

Note that PASS obtains the same samples sizes as Phillips (1990).

## Example 4 – Validation using Machin et al. (1997)

Machin *et al.* (1997) page 107 present an example of determining the sample size for a cross-over design in which the reference mean is 35.03, the treatment mean is 35.03, the standard deviation, entered as the square root of the within mean square error, is 40% of the reference mean, the limits are plus or minus 20% of the reference mean, the power is 80%, and the significance level is 0.10. Machin *et al.* calculate the total sample size to be 54.

When the parameters are given as percentages of the reference mean, it is easy enough to calculate the exact amounts by applying those percentages. However, the percentages can all be entered directly as long as all parameters (EU, EL, D, and Sw) are specified as percentages.

### 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 Between Two Means in a 2x2 Cross-Over Design** procedure window by expanding **Means**, then **Cross-Over (2x2) Design**, then clicking on **Equivalence**, and then clicking on **Equivalence Tests for the Difference Between Two Means in a 2x2 Cross-Over Design**. You may then make the appropriate entries as listed below, or open **Example 4** by going to the **File** menu and choosing **Open Example Template**.

<u>Option</u>	<u>Value</u>
<b>Design Tab</b>	
Solve For .....	<b>Sample Size</b>
Power.....	<b>0.8</b>
Alpha.....	<b>0.10</b>
EU  (Upper Equivalence Limit) .....	<b>20</b>
- EL  (Lower Equivalence Limit).....	<b>-Upper Limit</b>
D (True Difference) .....	<b>0</b>
Specify S as.....	<b>Sw</b>
S.....	<b>40</b>

### Output

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

### Numeric Results

Numeric Results for Testing Equivalence Using a Cross-Over Design							
	Total Sample Size (N)	Lower Equiv. Limit	Upper Equiv. Limit	True Difference	Standard Deviation Sw	Alpha	Beta
Power	54	-20.00	20.00	0.00	40.00	0.1000	0.1950

Note that PASS also has obtained a sample size of 54.

## Example 5 – Validation using Chow and Liu (1999)

Chow and Liu (1999) page 153 present an example of determining the sample size for a cross-over design in which the reference mean is 82.559, the treatment mean is 82.559, the standard deviation, entered as the square root of the within mean square error, is 15.66%, the limits are plus or minus 20%, the power is 80%, and the significance level is 0.05. They calculate a sample size of 12. PASS calculates a sample size of 13. To see why PASS has increased the sample size by one, we will evaluate the power at sample sizes of 10, 12, 13, 14, and 16.

### 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 Between Two Means in a 2x2 Cross-Over Design** procedure window by expanding **Means**, then **Cross-Over (2x2) Design**, then clicking on **Equivalence**, and then clicking on **Equivalence Tests for the Difference Between Two Means in a 2x2 Cross-Over Design**. You may then make the appropriate entries as listed below, or open **Example 5** by going to the **File** menu and choosing **Open Example Template**.

<u>Option</u>	<u>Value</u>
<b>Design Tab</b>	
Solve For .....	<b>Power</b>
Alpha.....	<b>0.05</b>
N (Total Sample Size).....	<b>10 12 13 14 16</b>
EU  (Upper Equivalence Limit) .....	<b>20</b>
- EL  (Lower Equivalence Limit).....	<b>-Upper Limit</b>
D (True Difference).....	<b>0</b>
Specify S as.....	<b>Sw</b>
S.....	<b>15.66</b>

### Output

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

### Numeric Results

Numeric Results for Testing Equivalence Using a Cross-Over Design

	Total Sample Size (N)	Lower Equiv. Limit	Upper Equiv. Limit	True Difference	Standard Deviation Sw	Alpha	Beta
Power	10	-20.00	20.00	0.00	15.66	0.0500	0.3357
	12	-20.00	20.00	0.00	15.66	0.0500	0.2068
	13	-20.00	20.00	0.00	15.66	0.0500	0.1637
	14	-20.00	20.00	0.00	15.66	0.0500	0.1248
	16	-20.00	20.00	0.00	15.66	0.0500	0.0742

The power for  $N = 12$  is 0.7932. The power for  $N = 13$  is 0.8363. Hence, to achieve better than 80% power, a sample size of 13 is necessary. However, 0.7932 is sufficiently close to 0.800 to make  $N = 12$  a reasonable choice (as Chow and Liu did).

## Example 6 – Validation using Senn (1993)

Senn (1993) page 217 presents an example of determining the sample size for a cross-over design in which the reference mean is equal to the treatment mean, the standard deviation, entered as the square root of the within mean square error, is 45, the equivalence limits are plus or minus 30, the power is 80%, and the significance level is 0.05. He calculates a sample size of 40.

### 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 Between Two Means in a 2x2 Cross-Over Design** procedure window by expanding **Means**, then **Cross-Over (2x2) Design**, then clicking on **Equivalence**, and then clicking on **Equivalence Tests for the Difference Between Two Means in a 2x2 Cross-Over Design**. You may then make the appropriate entries as listed below, or open **Example 6** by going to the **File** menu and choosing **Open Example Template**.

<u>Option</u>	<u>Value</u>
<b>Design Tab</b>	
Solve For .....	<b>Sample Size</b>
Power.....	<b>0.8</b>
Alpha.....	<b>0.05</b>
EU  (Upper Equivalence Limit) .....	<b>30</b>
- EL  (Lower Equivalence Limit).....	<b>-Upper Limit</b>
D (True Difference).....	<b>0</b>
Specify S as.....	<b>Sw</b>
S.....	<b>45</b>

### Output

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

### Numeric Results

Numeric Results for Testing Equivalence Using a Cross-Over Design							
Power	Total Sample Size (N)	Lower Equiv. Limit	Upper Equiv. Limit	True Difference	Standard Deviation Sw	Alpha	Beta
0.8004	40	-30.00	30.00	0.00	45.00	0.0500	0.1996

PASS also calculates a sample size of 40.