Chapter 177

Equivalence Tests for the Difference 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 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.

© NCSS, LLC. All Rights Reserved

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 $n_1 = n_2 = N/2$.

Sequence one is given treatment A followed by treatment B. Sequence two is given treatment B followed by treatment 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 Chow, Shao, Wang, & Lokhnygina (2018). Consider a 2×2 cross-over design and let x_{ijk} represent the binary response (0 or 1) from the j^{th} subject (j = 1, ..., n_i) in the i^{th} sequence (i = 1, 2) given the k^{th} treatment (k = 1, 2). Here we assume that the sample sizes are equal in both sequences such that $n_1 = n_2 = n$. If replicates are taken from each subject (as in a 2×2*m* replicated cross-over design) then $x_{ijk} = \bar{x}_{ijk.} = \frac{1}{m}\left(\sum_{l=1}^{m} x_{ijkl}\right)$, where x_{ijkl} represents the l^{th} binary response replicate (l = 1, ..., m) from the j^{th} subject (j = 1, ..., n_i) in the i^{th} sequence (i = 1, 2) given the k^{th} treatment (k = 1, 2). The observations taken from the same subject may be correlated with one another. If we assume no sequence and period effects, then we can state that $P\left(x_{ijk} = 1\right) = P_k$. Further define the paired differences of treatment – control for each subject within each sequence as

$$d_{ij} = x_{ijT} - x_{ijC}$$

$$= x_{ij1} - x_{ij2}$$

and the overall treatment – control difference as

$$\delta = P_T - P_C$$

$$= P_1 - P_2.$$

The overall difference can be estimated as

$$\hat{\delta} = \frac{1}{2n}\sum_{i=1}^{2}\sum_{j=1}^{n} d_{ij}.$$

The estimated difference is asymptotically normally distributed with variance σ_d^2, which can be estimated as

$$\hat{\sigma}_d^2 = \frac{1}{2(n-1)}\sum_{i=1}^{2}\sum_{j=1}^{n}\left(d_{ij} - \bar{d}_{i.}\right)^2,$$

where

$$\bar{d}_{i.} = \frac{1}{n}\sum_{j=1}^{n} d_{ij}.$$

The standard deviation, then, is

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

with estimate

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

© NCSS, LLC. All Rights Reserved.

Equivalence Test Statistics

The null and alternative hypotheses for an equivalence test are

$$H_0: P_T - P_C \leq D_{0L} \quad \text{or} \quad P_T - P_C \geq D_{0U} \qquad \text{vs.} \qquad H_A: D_{0L} < P_T - P_C < D_{0U}$$

or equivalently

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

where D_{0L} and D_{0U} are the lower and upper equivalence bounds, respectively (i.e., the smallest and largest differences ($P_T - P_C$) 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

$$Z_L = \frac{\hat{\delta} - D_{0L}}{\frac{\hat{\sigma}_d}{\sqrt{2n}}} \quad \text{and} \quad Z_U = \frac{\hat{\delta} - D_{0U}}{\frac{\hat{\sigma}_d}{\sqrt{2n}}}$$

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

Equivalence Power Calculation

Derived from Chow, Shao, Wang, & Lokhnygina (2018) page 85, the power for an equivalence test of $H_0: \delta \leq D_{0L}$ or $\delta \geq D_{0U}$ versus $H_A: D_{0L} < \delta < D_{0U}$ is given as

$$\Phi\left(\frac{D_{0U} - \delta_1}{\frac{\sigma_d}{\sqrt{2n}}} - Z_{1-\alpha}\right) - \Phi\left(\frac{D_{0L} - \delta_1}{\frac{\sigma_d}{\sqrt{2n}}} + Z_{1-\alpha}\right)$$

where $\Phi()$ is the standard normal distribution function, δ_1 is the actual value of the difference 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*.

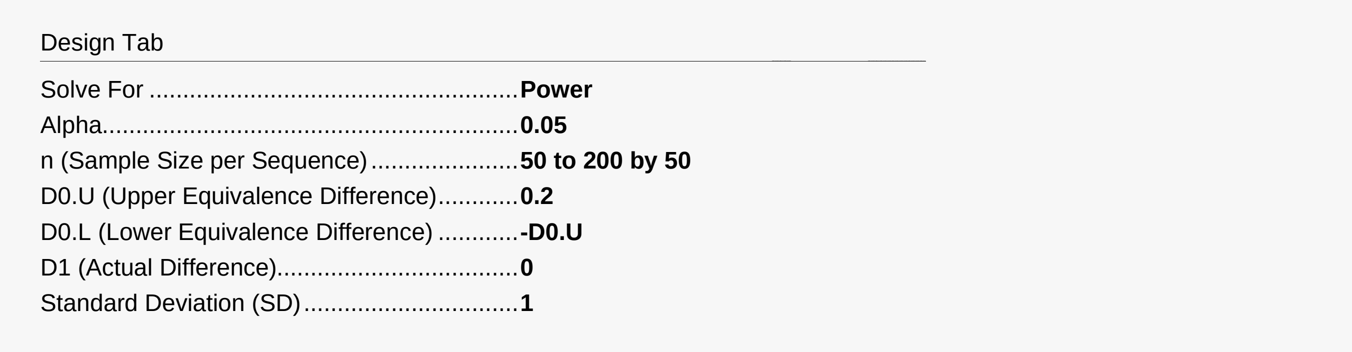
© NCSS, LLC. All Rights Reserved.

Example 1 – Power Analysis

Suppose you want to consider the power of an equivalence test of the hypotheses $H_0: \delta \leq -0.2 \text{ or } \delta \geq 0.2$ versus $H_A: -0.2 < \delta < 0.2$ in a balanced cross-over design with a binary endpoint where the test is computed based on the difference for sequence sample sizes between 50 and 200. Let's assume that the actual difference is 0 and the estimated standard deviation of the paired differences is 1. The significance level is 0.05.

Setup

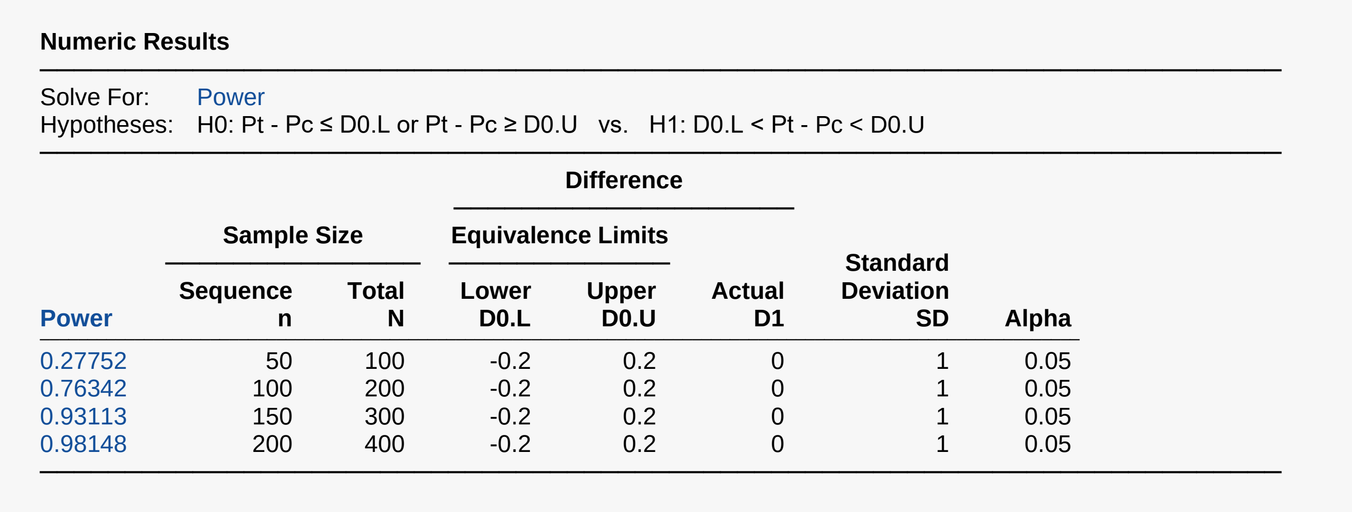
If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.



```
Design Tab
Solve For ......................................................Power
Alpha...........................................................0.05
n (Sample Size per Sequence).....................50 to 200 by 50
D0.U (Upper Equivalence Difference)............0.2
D0.L (Lower Equivalence Difference) ............-D0.U
D1 (Actual Difference)...................................0
Standard Deviation (SD)...............................1
```

Output

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



```
Numeric Results

Solve For:    Power
Hypotheses:   H0: Pt - Pc ≤ D0.L or Pt - Pc ≥ D0.U   vs.   H1: D0.L < Pt - Pc < D0.U

                                            Difference
                                     ---------------------------
               Sample Size            Equivalence Limits
           -------------------       -------------------                Standard
           Sequence      Total        Lower      Upper      Actual     Deviation
Power             n          N         D0.L       D0.U          D1            SD     Alpha
0.27752          50        100         -0.2        0.2           0             1      0.05
0.76342         100        200         -0.2        0.2           0             1      0.05
0.93113         150        300         -0.2        0.2           0             1      0.05
0.98148         200        400         -0.2        0.2           0             1      0.05
```

© NCSS, LLC. All Rights Reserved.

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 sample size from both sequences. The sample is divided equally among sequences.
D0.L The lower equivalence difference used to specify the hypothesis test.
D0.U The upper equivalence difference used to specify the hypothesis test.
D1 The actual difference at which power is calculated.
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 2×2 cross-over design will be used to test whether the treatment proportion is equivalent to the standard proportion, with equivalence difference bounds of -0.2 and 0.2 (H0: Pt - Pc ≤ -0.2 or Pt - Pc ≥ 0.2 versus H1: -0.2 < Pt - Pc < 0.2). The comparison will be made using two one-sided Z-tests, with an overall Type I error rate (α) of 0.05. The standard deviation of paired differences is assumed to be 1. To detect a proportion difference (Pt - Pc) of 0 with a sample size of 50 in each sequence (totaling 100 subjects), the power is 0.27752.

Dropout-Inflated Sample Size

	Sample Size		Dropout-Inflated Enrollment Sample Size		Expected Number of Dropouts	
Dropout Rate	n	N	n'	N'	d	D
20%	50	100	63	126	13	26
20%	100	200	125	250	25	50
20%	150	300	188	376	38	76
20%	200	400	250	500	50	100

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.

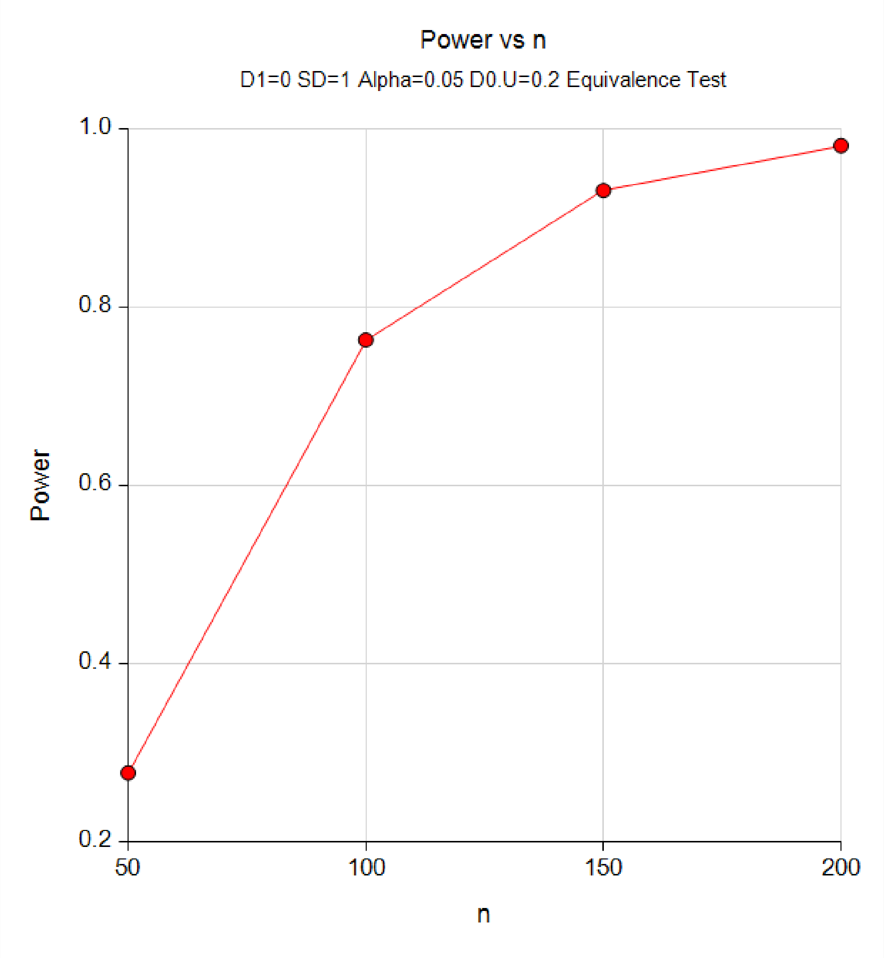
Dropout Summary Statements

Anticipating a 20% dropout rate, 63 subjects should be enrolled in each group to obtain final sample sizes of 50 subjects per group.

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 100 per sequence is required for 80% power.

© NCSS, LLC. All Rights Reserved.

Example 2 – Calculating Sample Size when Estimating the Standard Deviation from a Previous Study

This example demonstrates how to calculate the sample size when estimating the standard deviation of the paired differences from data in a previous study using the method in Chow, Shao, Wang, & Lokhnygina (2018) on pages 82 and 83. In this example we'll find the sample size required to detect a difference of 0 with 90% power in a test of the hypotheses $H_0 : \delta \leq -0.2$ or $\delta \geq 0.2$ versus $H_A : -0.2 < \delta < 0.2$ at a significance level of 0.05, with the standard deviation estimated using cell counts from a previous 2x2 cross-over study with equal sample size per sequence.

Assume that the following results were previously obtained from 280 subjects in a simple 2x2 cross-over trial comparing two inhalation devices, A and B. These results are similar to Table 3.2 of Lui (2016) on page 36 with a slight adjustment to sequence 1 that makes the sample sizes equal per sequence.

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

		Period 2 (B)		
		Yes	No	Total
Period 1 (A)	Yes	27	41	67
	No	15	57	72
	Total	42	98	140

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 paired differences of treatment – control for each subject within each sequence are

$$d_{ij} = x_{ijT} - x_{ijC}$$

$$= x_{ijB} - x_{ijA}$$

If x_{ijk} is a binary variable, then d_{ij} can take on the values 0, 1, and -1. If we summarize the results in sequence 1, then there are 15 subjects with $d_{ij} = 1$ (B = "Yes", A = "No"), 41 subjects with $d_{ij} = -1$ (B = "No", A = "Yes"), and 27 + 57 = 84 subjects with $d_{ij} = 0$ (B = "Yes", A = "Yes" or B = "No", A = "No"). The average paired difference, then, for sequence 1 is

$$\bar{d}_{1.} = \frac{1}{n}\sum_{j=1}^{n} d_{1j}$$

$$= \frac{15(1) + 41(-1) + 84(0)}{140}$$

$$= \frac{15 - 41}{140}$$

$$= -0.1857.$$

© NCSS, LLC. All Rights Reserved.

Similarly, the average paired difference from sequence 2 is

$$\bar{d}_{2.} = \frac{1}{n}\sum_{j=1}^{n} d_{2j}$$

$$= \frac{16(1) + 32(-1) + 92(0)}{140}$$

$$= \frac{16 - 32}{140}$$

$$= -0.1143.$$

The estimated overall treatment – control difference is

$$\hat{\delta} = \frac{1}{2n}\sum_{i=1}^{2}\sum_{j=1}^{n} d_{ij}$$

$$= \frac{\bar{d}_{1.} + \bar{d}_{2.}}{2}$$

$$= \frac{-0.1857 - 0.1143}{2}$$

$$= -0.15.$$

The estimated variance of paired differences, then, is

$$\hat{\sigma}_d^2 = \frac{1}{2(n-1)}\sum_{i=1}^{2}\sum_{j=1}^{n}\left(d_{ij} - \bar{d}_{i.}\right)^2$$

$$= \frac{15(1+0.1857)^2 + 41(-1+0.1857)^2 + 84(0+0.1857)^2 + 16(1+0.1143)^2 + 32(-1+0.1143)^2 + 92(0+0.1143)^2}{2(139)}$$

$$= 0.3502$$

with estimated standard deviation

$$\widehat{SD} = \hat{\sigma}_d$$

$$= \sqrt{\hat{\sigma}_d^2}$$

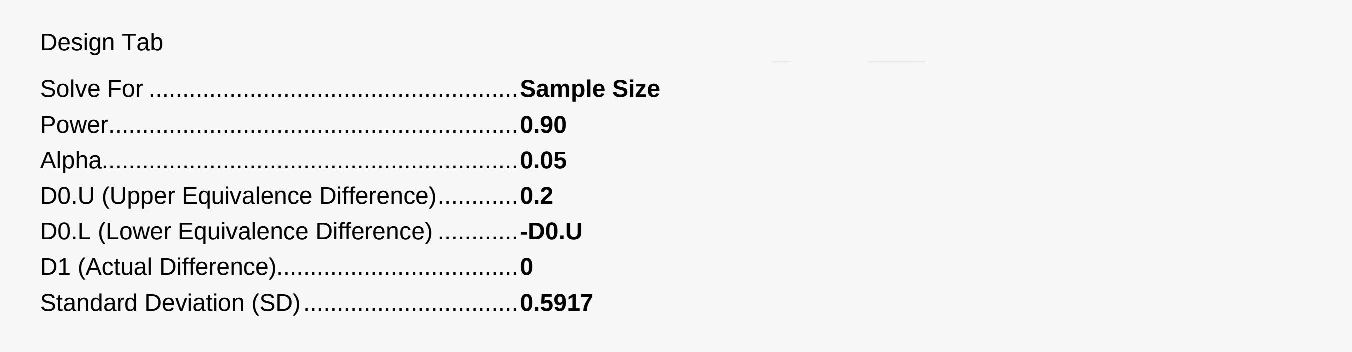
$$= \sqrt{0.3502}$$

$$= 0.5917.$$

© NCSS, LLC. All Rights Reserved.

Setup

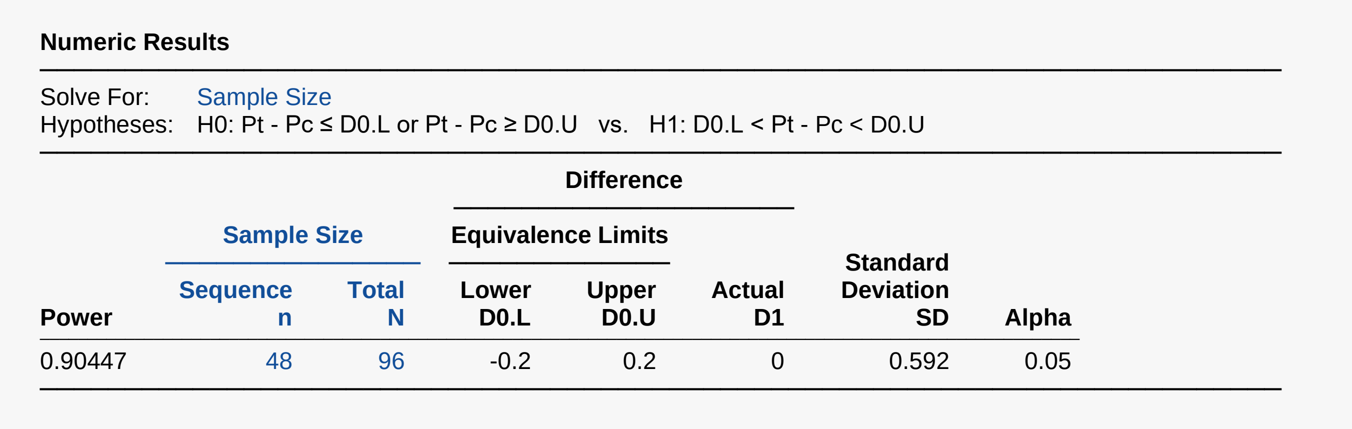
If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.



```
Design Tab
Solve For ......................................................Sample Size
Power...........................................................0.90
Alpha...........................................................0.05
D0.U (Upper Equivalence Difference)............0.2
D0.L (Lower Equivalence Difference) ............-D0.U
D1 (Actual Difference)...................................0
Standard Deviation (SD)...............................0.5917
```

Output

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



```
Numeric Results

Solve For:     Sample Size
Hypotheses:    H0: Pt - Pc ≤ D0.L or Pt - Pc ≥ D0.U   vs.   H1: D0.L < Pt - Pc < D0.U

                                                  Difference
                   Sample Size          Equivalence Limits
                                                                                    Standard
            Sequence       Total     Lower       Upper      Actual     Deviation
Power                 n           N      D0.L        D0.U           D1              SD        Alpha
0.90447            48          96        -0.2          0.2             0           0.592        0.05
```

This report indicates that the required sample size for 90% power is 48 per sequence for a total of 96.

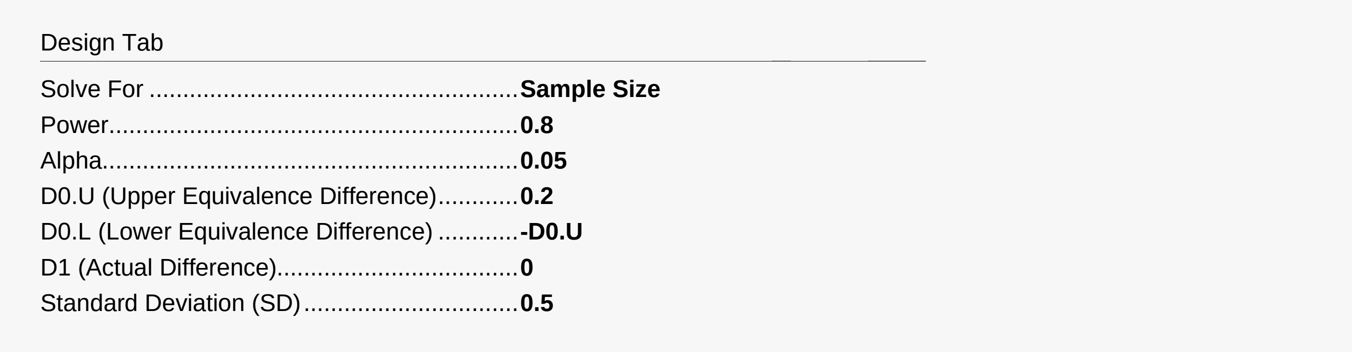
© NCSS, LLC. All Rights Reserved.

Example 3 – Calculating Sample Size (Validation using Chow, Shao, Wang, & Lokhnygina (2018))

On page 86, Chow, Shao, Wang, & Lokhnygina (2018) presents an example of finding the sample size required to detect a difference of 0 in an equivalence test with a margin of 0.2 with 80% power at a significance level of 0.05 when the standard deviation of paired differences is 0.5. They compute the required sample size to be 27 per sequence.

Setup

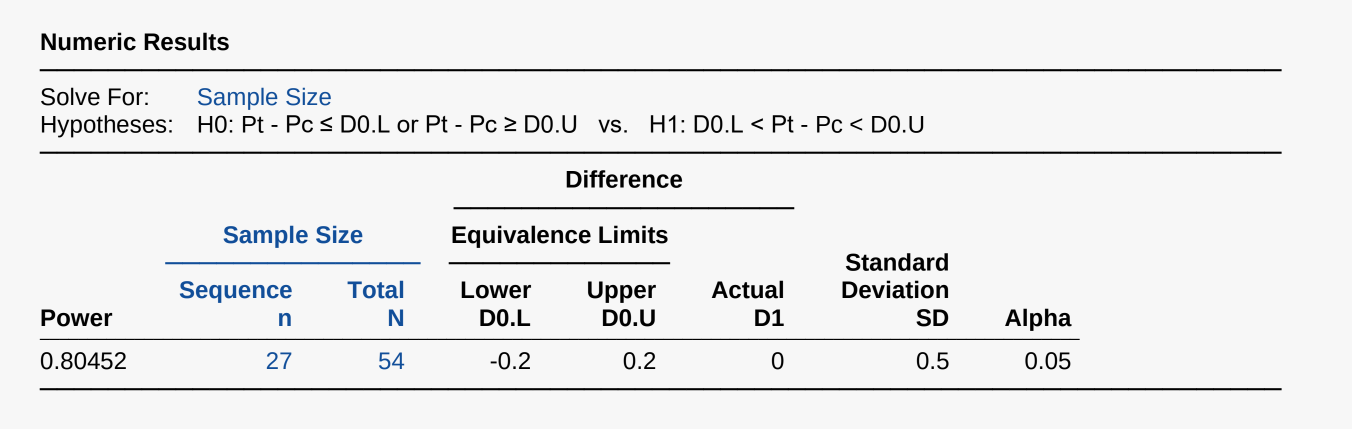
If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 3** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.



```
Design Tab
Solve For ......................................................Sample Size
Power...........................................................0.8
Alpha...........................................................0.05
D0.U (Upper Equivalence Difference)............0.2
D0.L (Lower Equivalence Difference) ............-D0.U
D1 (Actual Difference)..................................0
Standard Deviation (SD)...............................0.5
```

Output

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



Numeric Results

Solve For: Sample Size
Hypotheses: H0: Pt - Pc ≤ D0.L or Pt - Pc ≥ D0.U vs. H1: D0.L < Pt - Pc < D0.U

			Difference				
	Sample Size		Equivalence Limits				
Power	Sequence n	Total N	Lower D0.L	Upper D0.U	Actual D1	Standard Deviation SD	Alpha
0.80452	27	54	-0.2	0.2	0	0.5	0.05

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

© NCSS, LLC. All Rights Reserved.