Chapter 500

Tests for the Difference Between Two Means 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 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 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 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.

Technical Details

The 2x2 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 sequence is replicated *m* times. So, if *m* = 3, the sequences are ABABAB and BABABA.

The usual method of analysis is the analysis of variance. However, the power and sample size formulas that follow are based on the t-test, not the F-test. This is done because, in the balanced case, the t-test and the analysis of variance F-test are equivalent. Also, the F-test is limited to a two-sided hypothesis, while the t-test allows both one-sided and two-sided hypotheses. This is important because one-sided hypotheses are used for non-inferiority and equivalence testing.

Cross-Over Analysis

The following discussion summarizes the presentation of Chow and Liu (1999). The general linear model for the standard 2x2 cross-over design is

$$Y_{ijk} = \mu + S_{ik} + P_j + \mu_{(j,k)} + C_{(j-1,k)} + e_{ijk}$$

where *i* represents a subject (1 to N_k), *j* represents the period (1 or 2), and *k* represents the sequence (1 or 2). The S_{ik} represent the random effects of the subjects. The P_j represent the effects of the two periods. The $\mu_{(j,k)}$ represent the means of the two treatments. In the case of the 2x2 cross-over design

$$\mu_{(j,k)} = \begin{cases} \mu_1 & \text{if } k = j \\ \mu_2 & \text{if } k \neq j \end{cases}$$

where the subscripts 1 and 2 represent treatments A and B, respectively.

The $C_{(j-1,k)}$ represent the carry-over effects. In the case of the 2x2 cross-over design

$$C_{(j-1,k)} = \begin{cases} C_1 & \text{if } j = 2, k = 1 \\ C_2 & \text{if } j = 2, k = 2 \\ 0 & \text{otherwise} \end{cases}$$

where the subscripts 1 and 2 represent treatments A and B, respectively.

Assuming that the average effect of the subjects is zero, the four means from the 2x2 cross-over design can be summarized using the following table.

Sequence	Period 1	Period 2
1 (AB)	$\mu_{11} = \mu + P_1 + \mu_1$	$\mu_{21} = \mu + P_2 + \mu_2 + C_1$
2 (BA)	$\mu_{12} = \mu + P_1 + \mu_2$	$\mu_{22} = \mu + P_2 + \mu_1 + C_2$

where $P_1 + P_2 = 0$ and $C_1 + C_2 = 0$.

Test Statistic

The presence of a treatment effect can be studied by testing whether $\mu_1 - \mu_2 = \delta$ using a *t*-test or an F-test. If the F-test is used, only a two-sided test is possible. The t statistic is calculated as follows

$$t_d = \frac{(\bar{x}_T - \bar{x}_R) - \delta}{\hat{\sigma}_w \sqrt{\frac{2}{N}}}$$

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

The two-sided null hypothesis is rejected at the α significance level if $|t_d| > t_{\alpha/2,N-2}$. Similar results are available for a one-sided hypothesis test.

Computing the Within-Subject Variance (σ_w^2)

The ANOVA F-test is calculated using a standard repeated-measures analysis of variance table in which the between factor is the sequence and the within factor is the treatment. The within mean square error provides an estimate of the within-subject variance, σ_w^2, where

$$\sigma_w^2 = \text{Variance}\left(e_{ijk}\right)$$

If prior studies used a *t*-test rather than an ANOVA to analyze the data, you may not have a direct estimate of σ_w^2. Instead, you may have an estimate of the variance of the period differences from the *t*-test ($\hat{\sigma}_P^2$), an estimate of the variance of the paired differences ($\hat{\sigma}_D^2$), or an estimate of the variances of the paired variables ($\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$) and the correlation between the paired variables ($\hat{\rho}$). The within-subject variance, σ_w^2, is functionally related to these other variances as described below. Any of these different variances may be entered directly into this procedure.

Using the Variance of the Period Differences (σ_P^2)

The variance of the period differences for each subject within each sequence (σ_P^2) is defined as

$$\sigma_P^2 = \text{Variance}\left(\frac{Y_{i2k} - Y_{i1k}}{2}\right).$$

σ_P^2 has a functional relationship with the within-subject population variance (σ_w^2), namely

$$\sigma_P^2 = \frac{\sigma_w^2}{2},$$

such that

$$\sigma_w^2 = 2\sigma_P^2 .$$

The within-subject standard deviation (σ_w) is then

$$\sigma_w = \sqrt{2\sigma_P^2} .$$

Using the Variance of the Paired Differences (σ_D^2)

The variance of the paired differences (σ_D^2) is defined as

$$\sigma_D^2 = \text{Variance}(Y_{i2k} - Y_{i1k}) .$$

σ_D^2 has a functional relationship with the within-subject population variance (σ_w^2), namely

$$\sigma_D^2 = 2\sigma_w^2 ,$$

such that

$$\sigma_w^2 = \frac{\sigma_D^2}{2} .$$

The within-subject standard deviation (σ_w) is then

$$\sigma_w = \sqrt{\frac{\sigma_D^2}{2}} .$$

Using the Variances of the Paired Variables (σ_1^2 and σ_2^2) and the Correlation Between the Paired Variables (ρ)

The variances of the paired variables (σ_1^2 and σ_2^2) and the correlation between the paired variables (ρ) are defined as

$$\sigma_1^2 = \text{Variance}(Y_{i1k})$$

$$\sigma_2^2 = \text{Variance}(Y_{i2k})$$

$$\rho = \text{Correlation}(Y_{i1k}, Y_{i2k})$$

The variance of paired differences (σ_D^2) can be computed from σ_1^2, σ_2^2 and ρ as

$$\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 ,$$

such that the within-subject population variance (σ_w^2) can be computed as

$$\sigma_w^2 = \frac{\sigma_D^2}{2} = \frac{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}{2} .$$

The within-subject standard deviation (σ_w) is then

$$\sigma_w = \sqrt{\frac{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}{2}} .$$

If $\sigma_1^2 = \sigma_2^2 = \sigma_x^2$, then with

$$\sigma_x^2 = \text{Variance}\left(Y_{ijk}\right),$$

the formula for σ_w^2 reduces to

$$\sigma_w^2 = \sigma_x^2(1-\rho).$$

The within-subject standard deviation (σ_w) is then

$$\sigma_w = \sqrt{\sigma_x^2(1-\rho)}.$$

Computing the Power

The power is calculated as follows for a directional alternative (one-sided test).

1. Find t_α such that $1 - T_{df}(t_\alpha) = \alpha$, where $T_{df}(x)$ is the area left of *x* under a central-*t* curve and *df* = *N* - 2.
2. Calculate the noncentrality parameter: $\lambda = \frac{\delta\sqrt{N}}{\sigma_w\sqrt{2}}$.
3. Calculate: $\text{Power} = 1 - T'_{df,\lambda}(t_\alpha)$, where $T'_{df,\lambda}(x)$ is the area to the left of *x* under a noncentral-*t* curve with degrees of freedom *df* and noncentrality parameter λ.

Example 1 – Power Analysis

Suppose you want to consider the power of a balanced cross-over design that will be analyzed using the two-sided t-test approach. The difference between the treatment means under H0 is 0. Similar experiments have had a standard deviation of the period differences (σ_P) of 10. Compute the power when the true differences are 5 and 10 at sample sizes between 5 and 50. 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**
Alternative Hypothesis	**H1: δ ≠ δ0 (Two-Sided)**
Alpha	**0.05**
N (Total Sample Size)	**5 10 15 20 30 40 50**
δ0 (Mean Difference|H0)	**0**
δ1 (Mean Difference|H1)	**5 10**
Standard Deviation Input Type	**Enter the SD of Period Differences**
σP (SD of Period Differences)	**10**

Output

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

Numeric Results

Solve For: Power
Hypotheses: H0: δ = δ0 vs. H1: δ ≠ δ0

Power	Total Sample Size N	Mean Difference Null δ0	Mean Difference Actual δ1	Standard Deviation σP	Standard Deviation σw	Effect Size |δ1 - δ0|/σw	Alpha	Beta
0.06912	5	0	5	10	14.14214	0.35355	0.05	0.93088
0.10769	10	0	5	10	14.14214	0.35355	0.05	0.89231
0.14630	15	0	5	10	14.14214	0.35355	0.05	0.85370
0.18510	20	0	5	10	14.14214	0.35355	0.05	0.81490
0.26244	30	0	5	10	14.14214	0.35355	0.05	0.73756
0.33794	40	0	5	10	14.14214	0.35355	0.05	0.66206
0.41010	50	0	5	10	14.14214	0.35355	0.05	0.58990
0.12657	5	0	10	10	14.14214	0.70711	0.05	0.87343
0.28630	10	0	10	10	14.14214	0.70711	0.05	0.71370
0.43392	15	0	10	10	14.14214	0.70711	0.05	0.56608
0.56201	20	0	10	10	14.14214	0.70711	0.05	0.43799
0.75292	30	0	10	10	14.14214	0.70711	0.05	0.24708
0.86895	40	0	10	10	14.14214	0.70711	0.05	0.13105
0.93371	50	0	10	10	14.14214	0.70711	0.05	0.06629

Power	The probability of rejecting a false null hypothesis when the alternative hypothesis is true.
N	The total sample size drawn from all sequences. The sample is divided equally among sequences.
δ	The difference in means. Difference (δ) = Treatment Mean (μT) - Reference Mean (μR).
δ0	The mean difference under the null hypothesis, H0.
δ1	The actual mean difference under the alternative hypothesis at which the power is computed.
σP	The standard deviation of the period differences for each subject within each sequence. σP = √[var([Y_i2k - Y_i1k]/2)].
σw	The within-subject population standard deviation, which was calculated from σP using the equation, σw = σP × √2.
|δ1 - δ0|/σw	The Effect Size, i.e., the relative magnitude of the effect under the alternative hypothesis.
Alpha	The probability of rejecting a true null hypothesis.
Beta	The probability of failing to reject the null hypothesis when the alternative hypothesis is true.

Summary Statements

A 2×2 cross-over design will be used to test whether the treatment mean (μT) is different from the reference mean (μR), with a null difference of 0 (H0: δ = 0 versus Ha: δ ≠ 0, δ = μT - μR). The comparison will be made using a two-sided t-test with a Type I error rate (α) of 0.05. The within-subject population standard deviation is assumed to be 14.14214 (calculated using σP = 10 and the equation, σw = σP × √2). To detect a difference in means of 5, with a total sample size of 5 (allocated equally to the two sequences), the power is 0.06912.

Dropout-Inflated Sample Size

Dropout Rate	Sample Size N	Dropout-Inflated Enrollment Sample Size N'	Expected Number of Dropouts D
20%	5	7	2
20%	10	13	3
20%	15	19	4
20%	20	25	5
20%	30	38	8
20%	40	50	10
20%	50	63	13

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, 7 subjects should be enrolled to obtain a final sample size of 5 subjects.

References

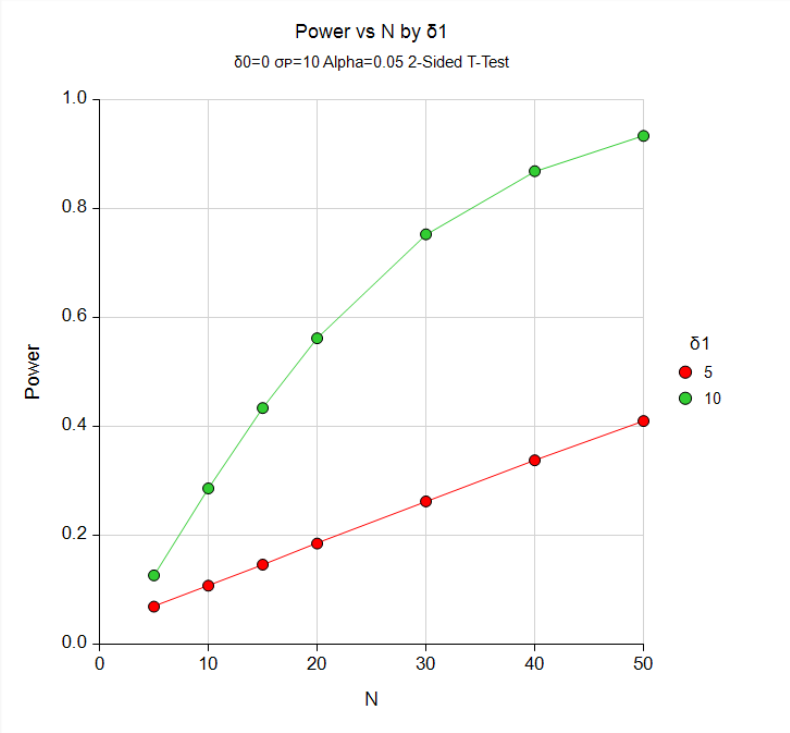
Chow, S.C. and Liu, J.P. 1999. Design and Analysis of Bioavailability and Bioequivalence Studies. Marcel Dekker. New York

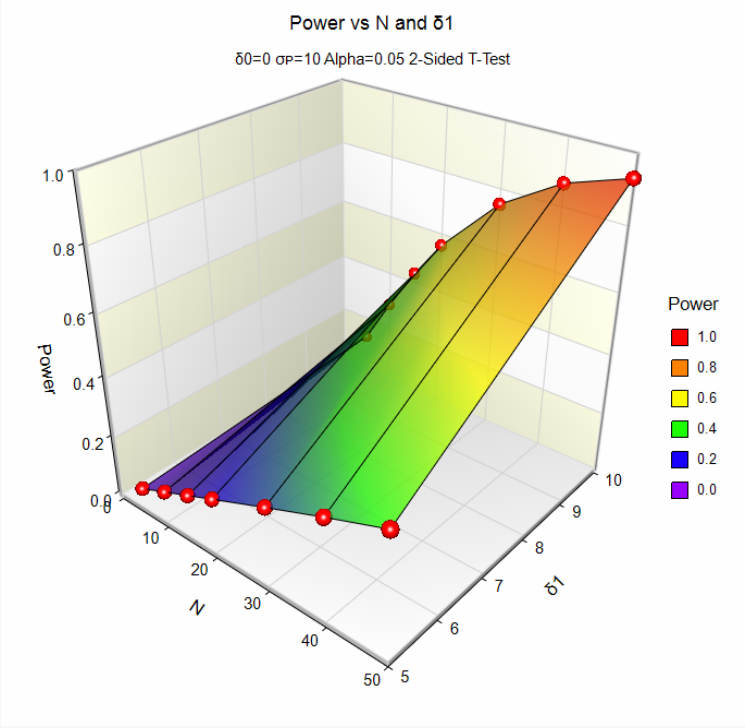
Chow, S.C., Shao, J., and Wang, H. 2003. Sample Size Calculations in Clinical Research. Marcel Dekker. New York.

Julious, Steven A. 2004. 'Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data.' Statistics in Medicine, 23:1921-1986.

Senn, Stephen. 2002. Cross-over Trials in Clinical Research. Second Edition. John Wiley & Sons. New York.

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 about 46 is needed when δ1 = 10 for 90% power, while δ1 = 5 never reaches 90% power in this range of sample sizes.

Example 2 – Finding the Sample Size

Continuing with Example 1, suppose the researchers want to find the exact sample size necessary to achieve 90% power for both values of δ1.

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**
Alternative Hypothesis	**H1: δ ≠ δ0 (Two-Sided)**
Power	**0.90**
Alpha	**0.05**
δ0 (Mean Difference|H0)	**0**
δ1 (Mean Difference|H1)	**5 10**
Standard Deviation Input Type	**Enter the SD of Period Differences**
σP (SD of Period Differences)	**10**

Output

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

Numeric Results

Solve For: Sample Size
Hypotheses: H0: δ = δ0 vs. H1: δ ≠ δ0

Power	Total Sample Size N	Mean Difference: Null δ0	Mean Difference: Actual δ1	Standard Deviation: σP	Standard Deviation: σw	Effect Size |δ1 - δ0|/σw	Alpha	Beta
0.90323	172	0	5	10	14.14214	0.35355	0.05	0.09677
0.91250	46	0	10	10	14.14214	0.70711	0.05	0.08750

This report shows the exact sample size necessary for each scenario.

Note that the search for *N* is conducted across only even values of *N* since the design is assumed to be balanced.

Example 3 – Validation using Julious (2004)

Julious (2004) page 1933 presents an example in which δ0 = 0.0, δ1 = 10, σ_w = 20, alpha = 0.05, and beta = 0.10. Julious obtains a sample size of 86.

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**
Alternative Hypothesis	**H1: δ ≠ δ0 (Two-Sided)**
Power	**0.90**
Alpha	**0.05**
δ0 (Mean Difference|H0)	**0**
δ1 (Mean Difference|H1)	**10**
Standard Deviation Input Type	**Enter the Within-Subject Population SD**
σw (Within-Subject Population SD)	**20**

Output

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

Numeric Results

Solve For: Sample Size
Hypotheses: H0: δ = δ0 vs. H1: δ ≠ δ0

Power	Total Sample Size N	Mean Difference: Null δ0	Mean Difference: Actual δ1	Standard Deviation σw	Effect Size |δ1 - δ0|/σw	Alpha	Beta
0.90648	88	0	10	20	0.5	0.05	0.09352

PASS obtains a sample size of 88, two higher than that obtained by Julious (2004). However, if you look at the power achieved by an N of 86, you will find that it is 0.899997—slightly less than the goal of 0.90.