Chapter 237

Analysis of 2x2 Cross-Over Designs using T-Tests for Equivalence

Introduction

This procedure analyzes data from a two-treatment, two-period (2x2) cross-over design where the goal is to demonstrate equivalence between a treatment and a control or reference. The response is assumed to be a continuous random variable that follows the normal distribution. When the normality assumption is suspect, the nonparametric Mann-Whitney U (or Wilcoxon Rank-Sum) Test may be employed.

In the two-period cross-over design, subjects are randomly assigned to one of two groups. One group receives treatment *R* followed by treatment *T*. The other group receives treatment *T* followed by treatment *R*. Thus, the response is measured at least twice on each subject.

Cross-over designs are used when the treatments alleviate a condition, rather than effect a cure. After the response to one treatment is measured, the treatment is removed, and the subject is allowed to return to a baseline response level. Next, the response to a second treatment is measured. Hence, each subject is measured twice, once with each treatment.

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.

Equivalence

Cross-over designs are popular in the assessment of equivalence. In this case, the effectiveness of a new treatment formulation (drug) is to be compared against the effectiveness of the currently used (reference) formulation. When showing equivalence, it is not necessary to show that the new treatment is better than the current treatment. Rather, the new treatment need only be shown to be as good as the reference so that it can be used in its place.

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.

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 time effect and the carry-over effect of the previous treatment.

The design 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 effect a cure.

Because subjects must be measured at least twice, it may be 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

Suppose you want to evaluate the non-inferiority of a treatment, *T*, as compared to a control or reference, *R*, using data on subjects in a 2x2 cross-over design, where a period effect may be present. This procedure allows you to perform this type of analysis.

Cross-Over Analysis

In the discussion that follows, we summarize the presentation of Chow and Liu (1999). We suggest that you review their book for a more detailed presentation.

The general linear model for the standard 2x2 cross-over design is

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

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 $F_{(j,k)}$ represent the effects of the two formulations (treatments). In the case of the 2x2 cross-over design

$$F_{(j,k)} = \begin{cases} F_R & \text{if } k = j \\ F_T & \text{if } k \neq j \end{cases}$$

where the subscripts *R* and *T* represent the *reference* and *treatment* formulations, respectively.

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

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

where the subscripts *R* and *T* represent the *reference* and *treatment* formulations, 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
$1(RT)$	$\mu_{11} = \mu + P_1 + F_R$	$\mu_{21} = \mu + P_2 + F_T + C_R$
$2(TR)$	$\mu_{12} = \mu + P_1 + F_T$	$\mu_{22} = \mu + P_2 + F_R + C_T$

where $P_1 + P_2 = 0$, $F_T + F_R = 0$, and $C_T + C_R = 0$.

Treatment Effect

Two-Sample T-Test for Treatment Effect

The presence of a treatment (drug) effect can be studied by testing whether $F_T - F_R = M$ using a *t* test. This test is calculated as follows

$$T_d = \frac{\hat{F} - M}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$\hat{F} = \bar{d}_{.1} - \bar{d}_{.2}$$

$$\bar{d}_{.k} = \frac{1}{n_k} \sum_{i=1}^{n_k} d_{ik}$$

$$\hat{\sigma}_d^2 = \frac{1}{(n_1 - n_2 - 2)} \sum_{k=1}^{2} \sum_{i=1}^{n_k} \left(d_{ik} - \bar{d}_{.k}\right)^2$$

$$d_{ik} = \frac{Y_{i2k} - Y_{i1k}}{2}$$

The null hypothesis of no drug effect is rejected at the α significance level if

$$|T_d| > t_{\alpha/2, n_1+n_2-2} \, .$$

A $100(1-\alpha)\%$ confidence interval for $F = F_T - F_R$ is given by

$$\hat{F} \pm \left(t_{\alpha/2, n_1+n_2-2}\right) \hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Mann-Whitney U (or Wilcoxon Rank-Sum) Test for Treatment Effect

Senn (2002) pages 113-114 describes Koch's adaptation of the Wilcoxon-Mann-Whitney rank sum test that tests treatment effects in the presence of period effects. The test is based on the period differences and assumes that there are no carryover effects. Koch's method calculates the ranks of the period differences for all subjects in the trial and then uses the Mann-Whitney U (or Wilcoxon Rank-Sum) Test to analyze these differences between the two sequence groups. The Mann-Whitney U (or Wilcoxon Rank-Sum) Test is described in detail in the Two-Sample T-Test chapter of the documentation.

Carryover Effect

The 2x2 cross-over design should only be used when there is no carryover effect from one period to the next. The presence of a carryover effect can be studied by testing whether $C_T = C_R = 0$ using a *t* test. This test is calculated as follows

$$T_c = \frac{\hat{C}}{\hat{\sigma}_u \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$\hat{C} = \bar{U}_{\cdot 2} - \bar{U}_{\cdot 1}$$

$$\bar{U}_{\cdot k} = \frac{1}{n_k} \sum_{i=1}^{n_k} U_{ik}$$

$$\hat{\sigma}_u^2 = \frac{1}{(n_1 - n_2 - 2)} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (U_{ik} - \bar{U}_{\cdot k})^2$$

$$U_{ik} = Y_{i1k} + Y_{i2k}$$

The null hypothesis of no carryover effect is rejected at the α significance level if

$$|T_c| > t_{\alpha/2, n_1 + n_2 - 2} .$$

A $100(1 - \alpha)\%$ confidence interval for $C = C_T - C_R$ is given by

$$\hat{C} \pm \left(t_{\alpha/2, n_1 + n_2 - 2}\right) \hat{\sigma}_u \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Period Effect

The presence of a period effect can be studied by testing whether $P_1 = P_2 = 0$ using a *t* test. This test is calculated as follows

$$T_P = \frac{\hat{P}}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$\hat{P} = \bar{O}_{.1} - \bar{O}_{.2}$$

$$\bar{O}_{.1} = \bar{d}_1$$

$$\bar{O}_{.2} = -\bar{d}_2$$

$$\hat{\sigma}_d^2 = \frac{1}{(n_1 - n_2 - 2)} \sum_{k=1}^{2} \sum_{i=1}^{n_k} \left(d_{ik} - \bar{d}_{.k}\right)^2$$

$$d_{ik} = \frac{Y_{i2k} - Y_{i1k}}{2}$$

The null hypothesis of no drug effect is rejected at the α significance level if

$$|T_P| > t_{\alpha/2, n_1+n_2-2} \, .$$

A $100(1-\alpha)\%$ confidence interval for $P = P_2 - P_1$ is given by

$$\hat{P} \pm \left(t_{\alpha/2, n_1+n_2-2}\right) \hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Bioequivalence

The *t* test of formulations (treatments) may be thought of as a preliminary assessment of bioequivalence. However, this *t* test investigates whether the two treatments are different. It does not assess whether the two treatments are the same—bioequivalent. That is, failure to reject the hypothesis of equal means does not imply bioequivalence. In order to establish bioequivalence, different statistical tests must be used.

Before discussing these tests, it is important to understand that, unlike most statistical hypothesis tests, when testing bioequivalence, you want to establish that the response to the two treatments is the same. Hence, the null hypothesis is that the mean responses are different, and the alternative hypothesis is that the mean responses are equal. This is just the opposite from the usual *t* test. This is why bioequivalence testing requires the special statistical techniques discussed here.

When using a cross-over design to test for bioequivalence, a washout period between the first and second periods must be used that is long enough to eliminate the residual effects of the first treatment from the

response to the second treatment. Because of this washout period, there is no carryover effect. Without a carryover effect, the general linear model reduces to

$$Y_{ijk} = \mu + S_{ik} + P_j + F_{(j,k)} + e_{ijk}$$

There are many types of bioequivalence. The 2x2 cross-over design is used to assess *average bioequivalence.* Remember that average bioequivalence is a statement about the population average. It does not make reference to the variability in responses to the two treatments. The 1992 FDA guidance uses the ± 20 rule which allows an average response to a test formulation to vary up to 20% from the average response of the reference formulation. This rule requires that ratio of the two averages $\mu_T \,/\, \mu_R$ be between 0.8 and 1.2 (80% to 120%). Another way of stating this is that the μ_T is within 20% of μ_R. The FDA requires that the significance level be 0.10 or less.

Several methods have been proposed to test for bioequivalence. Although the program provides several methods, you should select only the one that is most appropriate for your work.

Confidence Interval Approach

The confidence interval approach, first suggested by Westlake (1981), states that bioequivalence may be concluded if a $(1 - 2\alpha) \times 100\%$ confidence interval for the difference $\mu_T - \mu_R$ or ratio $\mu_T \,/\, \mu_R$ is within acceptance limits (α is usually set to 0.05). If the ± 20 rule is used, this means that the confidence interval for the difference must be between -0.2 and 0.2. Likewise, the confidence interval for the ratio must be between 0.8 and 1.2 (or 80% and 120%). Several methods have been suggested for computing the above confidence interval. The program provides the results for five of these. Perhaps the best of the five is the one based on Fieller's Theorem since it makes the fewest, and most general, assumptions about the distribution of the responses.

Classic (Shortest) Confidence Interval of the Difference

$$L_1 = (\bar{Y}_T - \bar{Y}_R) - \left(t_{\alpha,n_1+n_2-2}\right)\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$U_1 = (\bar{Y}_T - \bar{Y}_R) + \left(t_{\alpha,n_1+n_2-2}\right)\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Classic (Shortest) Confidence Interval of the Ratio

A confidence interval for the ratio may be calculated from the confidence interval on the difference using the formula

$$L_2 = (L_1 \,/\, \bar{Y}_R + 1) \times 100\%$$

$$U_2 = (U_1 \,/\, \bar{Y}_R + 1) \times 100\%$$

Westlake's Symmetric Confidence Interval of the Difference

First, compute values of k_1 and k_2 so that

$$1 - 2\alpha = \int_{k_2}^{k_1} T_{n_1+n_2-2}\, dt$$

Next, compute Δ using

$$\Delta = k_1 \hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} - (\bar{Y}_R - \bar{Y}_T)$$

$$= -k_2 \hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} + 2(\bar{Y}_R - \bar{Y}_T)$$

Finally, conclude bioequivalence if

$$|\Delta| < 0.2\mu_R$$

Westlake's Symmetric Confidence Interval of the Ratio

A confidence interval for the ratio may be calculated from the confidence interval on the difference using the formula

$$L_4 = (-|\Delta| \,/\, \bar{Y}_R + 1) \times 100\%$$

$$U_4 = (|\Delta| \,/\, \bar{Y}_R + 1) \times 100\%$$

Confidence Interval of the Ratio Based on Fieller's Theorem

Both the classic and Westlake's confidence interval for the ratio do not take into account the variability of $\bar{Y}_R$ and the correlation between $\bar{Y}_R$ and $\bar{Y}_T - \bar{Y}_R$. Locke (1984) provides formulas using Fieller's theorem that does take into account the variability of $\bar{Y}_R$. This confidence interval is popular not only because it takes into account the variability of $\bar{Y}_R$, but also the intersubject variability. Also, it only assumes that the data are normal, but not that the group variances are equal as do the other two approaches.

The $(1 - 2\alpha) \times 100\%$ confidence limits for $\delta = \mu_T \,/\, \mu_R$ are the roots of the quadratic equation

$$(\bar{Y}_T - \delta\bar{Y}_R)^2 - \left(t_{\alpha,n_1+n_2-2}\right)^2 \omega(S_{TT} - 2\delta S_{TR} + \delta^2 S_{RR}) = 0$$

where

$$\omega = \frac{1}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

$$S_{RR} = \frac{1}{(n_1 - n_2 - 2)}\left[\sum_{i=1}^{n_1}(Y_{i11} - \bar{Y}_{\cdot 11})^2 + \sum_{i=1}^{n_2}(Y_{i22} - \bar{Y}_{\cdot 22})^2\right]$$

$$S_{TT} = \frac{1}{(n_1 - n_2 - 2)}\left[\sum_{i=1}^{n_1}(Y_{i21} - \bar{Y}_{\cdot 21})^2 + \sum_{i=1}^{n_2}(Y_{i12} - \bar{Y}_{\cdot 12})^2\right]$$

$$S_{TR} = \frac{1}{(n_1 - n_2 - 2)}\left[\sum_{i=1}^{n_1}(Y_{i11} - \bar{Y}_{\cdot 11})(Y_{i21} - \bar{Y}_{\cdot 21}) + \sum_{i=1}^{n_2}(Y_{i12} - \bar{Y}_{\cdot 12})(Y_{i22} - \bar{Y}_{\cdot 22})\right]$$

Additionally, in order for the roots of the quadratic equation to be finite positive real numbers, the above values must obey the conditions

$$\frac{\bar{Y}_R}{\sqrt{\omega S_{RR}}} > t_{\alpha, n_1+n_2-2}$$

and

$$\frac{\bar{Y}_T}{\sqrt{\omega S_{TT}}} > t_{\alpha, n_1+n_2-2}$$

Interval Hypotheses Testing Approach

Schuirmann (1981) introduced the idea of using an interval hypothesis to test for average bioequivalence using the following null and alternative hypotheses

$$H_0: \mu_T - \mu_R \leq \theta_L \quad \text{or} \quad \mu_T - \mu_R \geq \theta_U$$
$$H_a: \theta_L < \mu_T - \mu_R < \theta_U$$

where θ_L and θ_U are limits selected to insure bioequivalence. Often these limits are set at 20% of the reference mean. These hypotheses can be rearranged into two one-sided hypotheses as follows

$$H_{01}: \mu_T - \mu_R \leq \theta_L \quad \text{versus} \quad H_{a1}: \mu_T - \mu_R > \theta_L$$
$$H_{02}: \mu_T - \mu_R \geq \theta_U \quad \text{versus} \quad H_{a2}: \mu_T - \mu_R < \theta_U$$

The first hypothesis test whether the treatment response is too low and the second tests whether the treatment response is too high. If both null hypotheses are rejected, you conclude that the treatment drug is bioequivalent to the reference drug.

Schuirmann's Two One-Sided Tests Procedure

Schuirmann's procedure is to conduct two one-sided tests, each at a significance level of α. If both tests are rejected, the conclusion of bioequivalence is made at the α significance level. That is, you conclude that μ_T and μ_R are average equivalent at the α significance level if

$$T_L = \frac{(\bar{Y}_T - \bar{Y}_R) - \theta_L}{\hat{\sigma}_d\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t_{\alpha,n_1+n_2-2}$$

and

$$T_U = \frac{(\bar{Y}_T - \bar{Y}_R) - \theta_U}{\hat{\sigma}_d\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > -t_{\alpha,n_1+n_2-2}$$

Wilcoxon-Mann-Whitney Two One-Sided Tests Procedure

When the normality assumption is suspect, you can use the nonparametric version of Schuirmann's procedure, known as the Wilcoxon-Mann-Whitney two one-sided tests procedure. This rather complicated procedure is described on pages 110 - 115 of Chow and Liu (1999) and we will not repeat their presentation here.

Anderson and Hauck's Test

Unlike Schuirman's test, Anderson and Hauck (1983) proposed a single procedure that evaluates the null hypothesis of inequivalence versus the alternative hypothesis of equivalence. The significance level of the Anderson and Hauck test is given by

$$\alpha = \Pr\left(|t_{AH}| - \hat{\delta}\right) - \Pr\left(-|t_{AH}| - \hat{\delta}\right)$$

where

$$\Pr(x) = \int_{-\infty}^{x} t_{n_1+n_2-2}\,dt$$

$$\hat{\delta} = \frac{\theta_U - \theta_L}{\hat{\sigma}_d\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

$$t_{AH} = \frac{(\bar{Y}_T - \bar{Y}_R) - (\theta_U + \theta_L)/2}{\hat{\sigma}_d\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Data Structure

The data for a cross-over design is entered into three variables. The first variable contains the sequence number, the second variable contains the response in the first period, and the third variable contains the response in the second period. Note that each row of data represents the complete response for a single subject.

Chow and Liu (1999) give the following data on page 73. These data are contained in the dataset called ChowLiu73.

ChowLiu73 Dataset

Sequence	Period 1	Period 2
1	74.675	73.675
1	96.400	93.250
1	101.950	102.125
1	79.050	69.450
1	79.050	69.025
1	85.950	68.700
1	69.725	59.425
1	86.275	76.125
1	112.675	114.875
1	99.525	116.250
1	89.425	64.175
1	55.175	74.575
2	74.825	37.350
2	86.875	51.925
2	81.675	72.175
2	92.700	77.500
2	50.450	71.875
2	66.125	94.025
2	122.450	124.975
2	99.075	85.225
2	86.350	95.925
2	49.925	67.100
2	42.700	59.425
2	91.725	114.05

Example 1 – 2x2 Cross-Over Analysis for Equivalence – Validation using Chow and Liu (1999))

This section presents an example of how to perform an equivalence test in an analysis of data from a 2x2 cross-over design. Chow and Liu (1999) page 73 provide an example of data from a 2x2 cross-over design. These data were shown in the Data Structure section earlier in this chapter. In this example, we'll assume that the test formulation will be deemed equivalent if it's mean is no more than 20% above or below the standard or reference formulation.

Chow and Liu (1999) indicates on pages 85-87 that the lower and upper 90% confidence limits of the difference for the Shortest C.I. Equivalence test method are -8.698 and 4.123, respectively. The lower and upper 90% confidence limits of the ratio for the Shortest C.I. Equivalence test method are 89.46% and 104.99%, respectively. Westlake's Symmetric 90% C.I. lower and upper limits for the difference are -7.413 and 7.413, respectively. Westlake's Symmetric 90% C.I. lower and upper limits for the ratio are 91.02% and 108.98%, respectively. In all cases, equivalence is concluded.

Chow and Liu (1999) indicates on page 98 that the T-values for the lower- and upper-tailed tests for Schuirmann's Equivalence Test using TOST are 3.810 and -5.036, respectively. The example on page 102 states that p-value for Anderson and Hauck's Equivalence test is 0.000454.

Setup

To run this example, complete the following steps:

1 Open the ChowLiu73 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **ChowLiu73** and click **OK**.

2 Specify the Analysis of 2x2 Cross-Over Designs using T-Tests for Equivalence procedure options

- Find and open the **Analysis of 2x2 Cross-Over Designs using T-Tests for Equivalence** procedure using the menus or the Procedure Navigator.
- The settings 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.

Variables Tab	
Sequence Group Variable	**Sequence**
Period 1 Variable	**Period1**
Period 2 Variable	**Period2**
Upper Equivalence Limit	**20**
%	**Checked**
Lower Equivalence Limit	**-Upper Limit**

Reports Tab	
Show Written Explanations	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Cross-Over Effects and Means Summary

Cross-Over Effects and Means Summary

Parameter	Count	Parameter Estimate	Standard Deviation	Standard Error	T*	Lower 95.0% CL	Upper 95.0% CL
Treatment Effect	24	-2.288	9.145	3.733	2.0739	-10.030	5.455
μ1 (or μR)	24	82.559		4.285			
μ2 (or μT)	24	80.272		4.395			
Period Effect	24	-1.731	9.145	3.733	2.0739	-9.474	6.011
μ|(Period=1)	24	82.281		4.043			
μ|(Period=2)	24	80.550		4.618			
Carryover Effect	24	-9.592	38.390	15.673	2.0739	-42.095	22.911
(R+T)|(Seq=1)	12	167.627	33.234	9.594			
(R+T)|(Seq=2)	12	158.035	42.930	12.393			

This report shows the estimated effects, means, standard deviations, standard errors, and confidence limits of various parameters and subgroups of the data. The least squares mean of treatment R is 82.559 and of treatment T is 80.272. The treatment effect [(μ2 - μ1) or (μT - μR)] is estimated to be -2.288. The period effect [(μ|(Period=2)) - (μ|(Period=1))] is estimated to be -1.731. The carryover effect [((R+T)|(Seq=2)) - ((R+T)|(Seq=1))] is estimated to be -9.592. Note that least squares means are created by taking the simple average of their component means, not by taking the average of the raw data. For example, if the mean of the 20 subjects in period 1 sequence 1 is 50.0 and the mean of the 10 subjects in period 2 sequence 2 is 40.0, the least squares mean is (50.0 + 40.0)/2 = 45.0. That is, no adjustment is made for the unequal sample sizes. Also note that the standard deviation of some of the subgroups is not calculated.

This report summarizes the estimated means and the treatment, period, and carryover effects.

Parameter

These are the items displayed on the corresponding lines. Note that the *Treatment* line is the main focus of the analysis. The *Period* and *Carryover* information is used for preliminary tests of assumptions.

Count

The count gives the number of non-missing values. This value is often referred to as the group sample size or *n*.

Parameter Estimate

These are the estimated values of the corresponding parameters. Formulas for the three effects were given in the Technical Details section earlier in this chapter.

Standard Deviation

The sample standard deviation is the square root of the sample variance. It is a measure of spread.

Standard Error

These are the standard errors of each of the effects. They provide an estimate of the precision of the effect estimate. The formulas were given earlier in the Technical Details section of this chapter.

T*

This is the t-value used to construct the confidence interval. If you were constructing the interval manually, you would obtain this value from a table of the Student's t distribution with *n* - 1 degrees of freedom.

Lower and Upper Confidence Limits

These values provide a confidence interval for the estimated effect.

Interpretation of the Above Report

This section provides a written interpretation of the above report.

Cross-Over Analysis Detail

Cross-Over Analysis Detail

Seq.	Period	Treatment	Count	Least Squares Mean	Standard Deviation	Standard Error
1	1	R	12	85.823	15.691	4.530
2	2	R	12	79.296	25.198	7.274
1	2	T	12	81.804	19.712	5.690
2	1	T	12	78.740	23.207	6.699
1	Difference	(T-R)/2	12	-2.009	6.423	1.854
2	Difference	(T-R)/2	12	0.278	11.225	3.240
1	Total	R+T	12	167.627	33.234	9.594
2	Total	R+T	12	158.035	42.930	12.393
.	.	R	24	82.559		4.285
.	.	T	24	80.272		4.395
1	.	.	24	83.814		
2	.	.	24	79.018		
.	1	.	24	82.281		4.043
.	2	.	24	80.550		4.618

This report shows the estimated effects, means, standard deviations, and standard errors of various subgroups of the data. The least squares mean of treatment R is 82.559 and of treatment T is 80.272. Note that least squares means are created by taking the simple average of their component means, not by taking the average of the raw data. For example, if the mean of the 20 subjects in period 1 sequence 1 is 50.0 and the mean of the 10 subjects in period 2 sequence 2 is 40.0, the least squares mean is (50.0 + 40.0)/2 = 45.0. That is, no adjustment is made for the unequal sample sizes. Also note that the standard deviation and standard error of some of the subgroups are not calculated.

This report provides the least squares means of various subgroups of the data.

Seq.

This is the sequence number of the mean shown on the line. When the dot (period) appears in this line, the results displayed are created by taking the simple average of the appropriate means of the two sequences.

Period

This is the period number of the mean shown on the line. When the dot (period) appears in this line, the results displayed are created by taking the simple average of the appropriate means of the two periods.

Treatment

This is the treatment (or formulation) of the mean shown on the line. When the dot (period) appears in this line, the results displayed are created by taking the simple average of the appropriate means of the two treatments.

When the entry is (*T*-*R*)/2, the mean is computed on the quantities created by dividing the difference in each subject's two scores by 2. When the entry is *R*+*T*, the mean is computed on the sums of the subjects two scores.

Count

The count is the number of subjects in the mean.

Least Squares Mean

Least squares means are created by taking the simple average of their component means, not by taking a weighted average based on the sample size in each component. For example, if the mean of the 20 subjects in period 1 sequence 1 is 50.0 and the mean of the 10 subjects in period 2 sequence 2 is 40.0, the least squares mean is (50.0 + 40.0)/2 = 45.0. That is, no adjustment is made for the unequal sample sizes. Since least squares means are used in all subsequent calculations, these are the means that are reported.

Standard Deviation

This is the estimated standard deviation of the subjects in the mean.

Standard Error

This is the estimated standard error of the least squares mean.

Equivalence Tests using 100(1 - 2α)% Confidence Intervals of the Difference

Equivalence Tests using 100(1 - 2α)% Confidence Intervals of the Difference

Test Type	Lower Equivalence Limit	Lower 90.0% Confidence Limit	Upper 90.0% Confidence Limit	Upper Equivalence Limit	Conclude Equivalence at α = 0.050?
Shortest C.I.	-16.512	-8.698	4.123	16.512	Yes
Westlake C.I.	-16.512	-7.413	7.413	16.512	Yes

Note: Westlake's k2 = -1.3730 and k1 = 2.5984.

Average bioequivalence of the two treatments has been found at the 0.05 significance level using the shortest confidence interval of the difference approach since both confidence limits, -8.698 and 4.123, are between the acceptance limits of -16.512 and 16.512. This experiment used a 2x2 cross-over design with 12 subjects in sequence 1 and 12 subjects in sequence 2.

Average bioequivalence of the two treatments has been found at the 0.05 significance level using Westlake's confidence interval of the difference approach since both confidence limits, -7.413 and 7.413, are between the acceptance limits of -16.512 and 16.512. This experiment used a 2x2 cross-over design with 12 subjects in sequence 1 and 12 subjects in sequence 2.

This report provides the results of two tests for bioequivalence based on confidence limits of the difference between the means of the two formulations. The results match Chow and Liu (1999) exactly.

Test Type

This is the type of test reported on this line. The mathematical details of each test were described earlier in the Technical Details section of this chapter.

Lower and Upper Equivalence Limit

These are the limits on bioequivalence. As long as the difference between the treatment formulation and reference formula is inside these limits, the treatment formulation is bioequivalent. These values were set by you. They are not calculated from the data.

Lower and Upper Confidence Limits

These are the confidence limits on the difference in response to the two formulations computed from the data. Note that the confidence coefficient is $(1 - 2\alpha) \times 100\%$. If both of these limits are inside the two equivalence limits, the treatment formulation is bioequivalent to the reference formulation. Otherwise, it is not.

Conclude Equivalence at α = 0.050?

This column indicates whether bioequivalence can be concluded.

Equivalence Tests using 100(1 - 2α)% Confidence Intervals of the Ratio

Equivalence Tests using 100(1 - 2α)% Confidence Intervals of the Ratio

Test Type	Lower Equivalence Limit	Lower 90.0% Confidence Limit	Upper 90.0% Confidence Limit	Upper Equivalence Limit	Conclude Equivalence at α = 0.050?
Shortest C.I.	80.000	89.464	104.994	120.000	Yes
Westlake C.I.	80.000	91.021	108.979	120.000	Yes
Fieller's C.I.	80.000	90.063	104.917	120.000	Yes

Average bioequivalence of the two treatments has been found at the 0.05 significance level using the shortest confidence interval of the ratio approach since both confidence limits, 89.464 and 104.994, are between the acceptance limits of 80.000 and 120.000. This experiment used a 2x2 cross-over design with 12 subjects in sequence 1 and 12 subjects in sequence 2.

Average bioequivalence of the two treatments has been found at the 0.05 significance level using Westlake's confidence interval of the ratio approach since both confidence limits, 91.021 and 108.979, are between the acceptance limits of 80.000 and 120.000. This experiment used a 2x2 cross-over design with 12 subjects in sequence 1 and 12 subjects in sequence 2.

Average bioequivalence of the two treatments has been found at the 0.05 significance level using Fieller's confidence interval of the ratio approach since both confidence limits, 90.063 and 104.917, are between the acceptance limits of 80.000 and 120.000. This experiment used a 2x2 cross-over design with 12 subjects in sequence 1 and 12 subjects in sequence 2.

This report provides the results of three tests for bioequivalence based on confidence limits of the ratio of the mean responses to the two formulations. The results match Chow and Liu (1999) exactly.

Test Type

This is the type of test report on this line. The mathematical details of each test were described earlier in the Technical Details section of this chapter.

Lower and Upper Equivalence Limit

These are the limits on bioequivalence in percentage form. As long as the percentage of the treatment formulation of the reference formula is between these limits, the treatment formulation is bioequivalent. These values were set by you. They are not calculated from the data.

Lower and Upper Confidence Limits

These are the confidence limits on the ratio of mean responses to the two formulations computed from the data. Note that the confidence coefficient is $(1 - 2\alpha) \times 100\%$. If both of these limits are inside the two equivalence limits, the treatment formulation is bioequivalent to the reference formulation. Otherwise, it is not.

Conclude Equivalence at α = 0.050?

This column indicates whether bioequivalence can be concluded.

Schuirmann's Equivalence Test using TOST (Two One-Sided Tests)

Schuirmann's Equivalence Test using TOST (Two One-Sided Tests)

Alternative Hypothesis	Treatment Mean Diff	SE	Lower Test T-Value	Upper Test T-Value	DF	Prob Level	Conclude Equivalence at α = 0.050?
-16.512 < μT-μR < 16.512	-2.288	3.733	3.8102	-5.0356	22	0.00048	Yes

Average bioequivalence of the two treatments was found at the 0.05 significance level using Schuirmann's two one-sided t-tests procedure. The probability level of the t-test of whether the treatment mean is not too much lower than the reference mean is 0.00048. The probability level of the t-test of whether the treatment mean is not too much higher than the reference mean is 0.00002. Since both of these values are less than 0.05, the null hypothesis of average bioinequivalence was rejected in favor of the alternative hypothesis of average bioequivalence. This experiment used a 2x2 cross-over design with 12 subjects in sequence 1 and 12 subjects in sequence 2.

This report provides the results of Schuirmann's two one-sided hypothesis tests procedure. The results match Chow and Liu (1999) exactly.

Alternative Hypothesis

This is the alternative hypothesis of equivalence that is being tested.

Lower and Upper Test T Value

These are the values of T_L and T_U, the two one-sided test statistics.

DF

This is the value of the degrees of freedom. In this case, the value of the degrees of freedom is $n_1 + n_2 - 2$.

Prob Level

This is the probability level (p-value) of the test. If this value is less than the chosen significance level, then the corresponding effect is said to be significant. For example, if you are testing at a significance level of 0.05, then probabilities that are less than 0.05 are statistically significant. You should choose a value appropriate for your study.

Conclude Equivalence at α = 0.050?

This column indicates whether bioequivalence is concluded.

Anderson and Hauck's Equivalence Test

Anderson and Hauck's Equivalence Test

Alternative Hypothesis	Treatment Mean Diff	SE	Pr(-TL)	Pr(TU)	Prob Level	Conclude Equivalence at α = 0.050?
-16.512 < μT-μR < 16.512	-2.288	3.733	0.00048	0.00002	0.00045	Yes

Average bioequivalence of the two treatments was found at the 0.05 significance level using Anderson and Hauck's test procedure. The actual probability level of the test was 0.00045. This experiment used a 2x2 cross-over design with 12 subjects in sequence 1 and 12 subjects in sequence 2.

This report provides the results of Anderson and Hauck's hypothesis test procedure. The results match Chow and Liu (1999) exactly.

Alternative Hypothesis

This is the alternative hypothesis of equivalence that is being tested.

Treatment Mean Difference

This is the difference between the treatment means, $\hat{\mu}_2 - \hat{\mu}_1$. This is known as the treatment effect.

SE

This is the standard error of the treatment effect. It provides an estimate of the precision of the treatment effect estimate. The formula was given earlier in the Technical Details section of this chapter.

Pr(-TL) and Pr(TU)

These values are subtracted to obtain the significance level of the test.

Prob Level

This is the significance level of the test. Bioequivalence is indicated when this value is less than a given level of α.

Conclude Equivalence at α = 0.050?

This column indicates whether bioequivalence is concluded.

Schuirmann's Wilcoxon-Mann-Whitney Equivalence Test using TOST (Two One-Sided Tests)

Schuirmann's Wilcoxon-Mann-Whitney Equivalence Test using TOST (Two One-Sided Tests)

Test Type	Alternative Hypothesis†	Lower Sum Ranks	Lower Prob Level	Upper Sum Ranks	Upper Prob Level	Conclude Equivalence at α = 0.050?
Exact*	-16.512 < Diff < 16.512	207	0.00025	91	0.00014	Yes
Normal Approximation	-16.512 < Diff < 16.512	207	0.00050	91	0.00033	Yes
Normal Approx. with C.C.	-16.512 < Diff < 16.512	207	0.00055	91	0.00037	Yes

† "Diff" refers to the location difference between the period differences for groups 1 and 2.
* The Exact Test is provided only when there are no ties and the sample size is ≤ 20 in both groups.

Average bioequivalence of the two treatments was found at the 0.05 significance level using the nonparametric version of Schuirmann's two one-sided tests procedure which is based on the Exact Wilcoxon-Mann-Whitney test. The probability level of the test of whether the treatment mean is not too much lower than the reference mean is 0.00025. The probability level of the test of whether the treatment mean is not too much higher than the reference mean is 0.00014. Since both of these values are less than 0.05, the null hypothesis of average bioinequivalence was rejected in favor of the alternative hypothesis of average bioequivalence. This experiment used a 2x2 cross-over design with 12 subjects in sequence 1 and 12 subjects in sequence 2.

This report provides the results of the nonparametric version of Schuirmann's two one-sided hypothesis tests procedure. The results match Chow and Liu (1999) page 113 which states that rank sums are 207 and 91.

Test Type

This is the type of test reported on this line.

Alternative Hypothesis

This is the alternative hypothesis of equivalence that is being tested.

Lower and Upper Sum Ranks

These are sums of the ranks for the lower and upper Mann-Whitney tests.

Lower and Upper Prob Level

These are the upper and lower significance levels of the two one-sided Wilcoxon-Mann-Whitney tests. Bioequivalence is indicated when both of these values are less than a given level of α.

Conclude Equivalence at α = 0.050?

This column indicates whether bioequivalence is concluded.

T-Tests for Period and Carryover Effects (Two-Sided)

T-Tests for Period and Carryover Effects (Two-Sided)

Parameter	Estimate	Standard Error	T-Value	DF	Prob Level	Reject H0 at α = 0.050?
Period Effect	-1.731	3.733	-0.4637	22	0.64739	No
Carryover Effect	-9.592	15.673	-0.6120	22	0.54681	No

A preliminary test failed to reject the assumption of equal period effects at the 0.05 significance level (the actual probability level was 0.64739). A preliminary test failed to reject the assumption of equal carryover effects at the 0.05 significance level (the actual probability level was 0.54681).

This report presents the T-tests for the period and carryover effects. In this case, both are not significantly different from zero.

Parameter

These are the items being tested. The *Period* and *Carryover* lines are preliminary tests of assumptions.

Estimate

These are the estimated values of the corresponding effects. Formulas for these effects were given in the Technical Details section earlier in this chapter.

Standard Error

These are the standard errors of each of the effects. They provide an estimate of the precision of the effect estimate. The formulas were given earlier in the Technical Details section of this chapter.

T-Value

These are the test statistics calculated from the data that are used to test whether the effect is different from zero.

DF

The *DF* is the value of the degrees of freedom. This is two less than the total number of subjects in the study.

Prob Level

This is the probability level (p-value) of the test. If this value is less than the chosen significance level, then the corresponding effect is said to be significant. Some authors recommend that the tests of assumptions (Period and Carryover) should be done at the 0.10 level of significance.

Tests of Assumptions Section

This section presents the results of tests for checking the normality assumption.

Tests of the Normality Assumption for the Period Differences in Sequence 1

Normality Test	Test Statistic	Prob Level	Reject H0 of Normality at α = 0.050?
Shapiro-Wilk	0.9418	0.52170	No
Skewness	-0.7849	0.43251	No
Kurtosis	0.3616	0.71767	No
Omnibus (Skewness or Kurtosis)	0.7468	0.68839	No

One of the underlying assumptions for the T-test is that the differences are normally distributed. This report presents the results of four different normality tests for the period differences within each sequence: Shapiro-Wilk Normality, Skewness Normality, Kurtosis Normality, Omnibus Normality.

Tests of the Normality Assumption for the Period Differences in Sequence 2

Normality Test	Test Statistic	Prob Level	Reject H0 of Normality at α = 0.050?
Shapiro-Wilk	0.9091	0.20784	No
Skewness	0.9127	0.36138	No
Kurtosis	-0.8364	0.40293	No
Omnibus (Skewness or Kurtosis)	1.5327	0.46472	No

One of the underlying assumptions for the T-test is that the differences are normally distributed. This report presents the results of four different normality tests for the period differences within each sequence: Shapiro-Wilk Normality, Skewness Normality, Kurtosis Normality, Omnibus Normality.

Shapiro-Wilk Normality

This test for normality has been found to be the most powerful test in most situations. It is the ratio of two estimates of the variance of a normal distribution based on a random sample of *n* observations. The numerator is proportional to the square of the best linear estimator of the standard deviation. The denominator is the sum of squares of the observations about the sample mean. The test statistic *W* may be written as the square of the Pearson correlation coefficient between the ordered observations and a set of weights which are used to calculate the numerator. Since these weights are asymptotically proportional to the corresponding expected normal order statistics, *W* is roughly a measure of the straightness of the normal quantile-quantile plot. Hence, the closer *W* is to one, the more normal the sample is.

The probability values for *W* are valid for sample sizes greater than 3. The test was developed by Shapiro and Wilk (1965) for sample sizes up to 20. **NCSS** uses the approximations suggested by Royston (1992) and Royston (1995) which allow unlimited sample sizes. Note that Royston only checked the results for sample sizes up to 5000 but indicated that he saw no reason larger sample sizes should not work. *W* may not be as powerful as other tests when ties occur in your data.

Skewness Normality

This is a skewness test reported by D'Agostino (1990). Skewness implies a lack of symmetry. One characteristic of the normal distribution is that it has no skewness. Hence, one type of non-normality is skewness.

The Value is the test statistic for skewness, while Prob Level is the *p*-value for a two-tailed test for a null hypothesis of normality. If this *p*-value is less than a chosen level of significance, there is evidence of non-normality. Under Decision (α = 0.050), the conclusion about skewness normality is given.

Kurtosis Normality

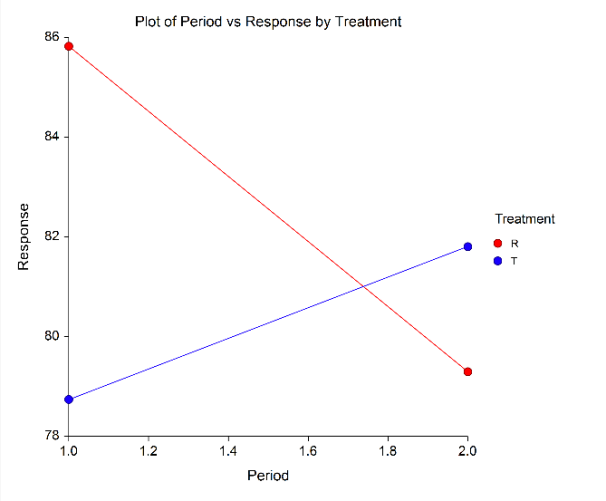
Kurtosis measures the heaviness of the tails of the distribution. D'Agostino (1990) reported a second normality test that examines kurtosis. The Value column gives the test statistic for kurtosis, and Prob Level is the *p*-value for a two-tail test for a null hypothesis of normality. If this *p*-value is less than a chosen level of significance, there is evidence of kurtosis non-normality. Under Decision (α = 0.050), the conclusion about normality is given.

Omnibus Normality

This third normality test, also developed by D'Agostino (1990), combines the skewness and kurtosis tests into a single measure. Here, as well, the null hypothesis is that the underlying distribution is normally distributed. The definitions for Value, Prob Level, and Decision are the same as for the previous two normality tests.

Plot of Sequence-by-Period Means

Plot of Sequence-by-Period Means

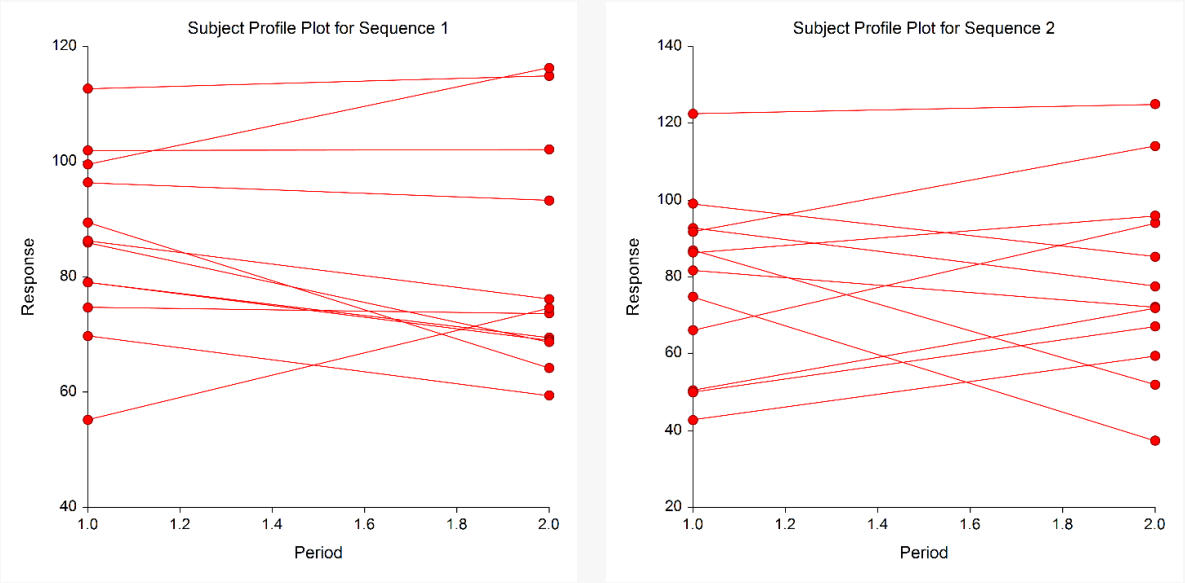


The sequence-by-period means plot shows the mean responses on the vertical axis and the periods on the horizontal axis. The lines connect like treatments. The distance between these lines represents the magnitude of the treatment effect.

If there are no period, carryover, or interaction effects, two horizontal lines will be displayed. The tendency for both lines to slope up or down represents period and carryover effects. The tendency for the lines to cross represents period-by-treatment interaction. This is also a type of carryover effect.

Plot of Subject Profiles

Plot of Subject Profiles



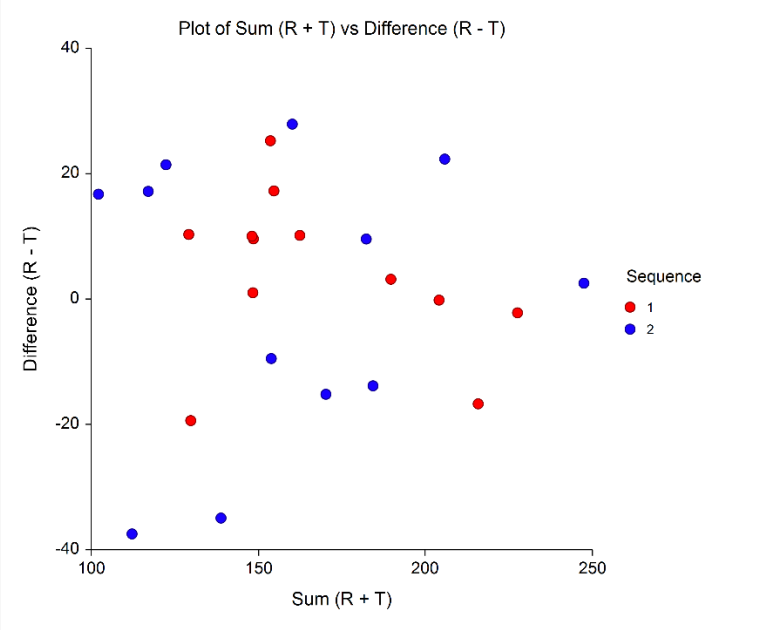
The profile plot displays the raw data for each subject. The response variable is shown along the vertical axis. The two sequences are shown along the horizontal axis. The data for each subject is depicted by two points connected by a line. The subject's response to the reference formulation is shown first followed by their response to the treatment formulation. Hence, for sequence 2, the results for the first period are shown on the right and for the second period on the left.

This plot is used to develop a feel for your data. You should view it first as a tool to check for outliers (points and subjects that are very different from the majority). Note that outliers should be removed from the analysis only if a reason can be found for their deletion. Of course, the first step in dealing with outliers is to double-check the data values to determine if a typing error might have caused them. Also, look for subjects whose lines exhibit a very different pattern from the rest of the subjects in that sequence. These might be a signal of some type of data-recording or data-entry error.

The profile plot allows you to assess the consistency of the responses to the two treatments across subjects. You may also be able to evaluate the degree to which the variation is equal in the two sequences.

Plot of Sums and Differences

Plot of Sums and Differences



The sums and differences plot shows the sum of each subject's two responses on the horizontal axis and the difference between each subject's two responses on the vertical axis. Dot plots of the sums and differences have been added above and to the right, respectively.

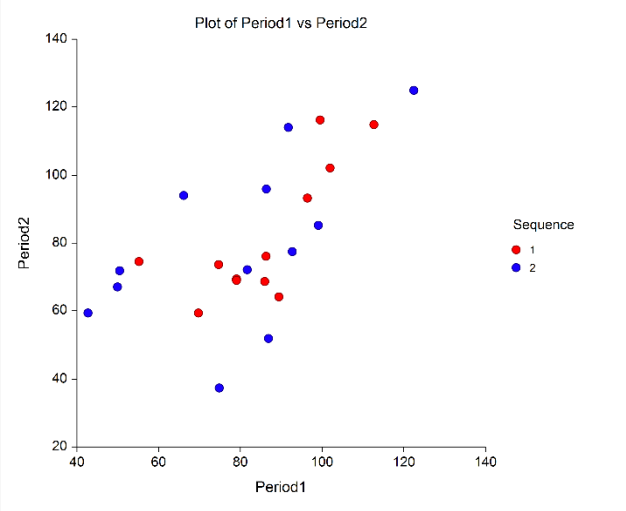
Each point represents the sum and difference of a single subject. Different plotting symbols are used to denote the subject's sequence. A horizontal line has been added at zero to provide an easy reference from which to determine if a difference is positive (favors treatment R) or negative (favors treatment T).

The degree to which the plotting symbols tend to separate along the horizontal axis represents the size of the carryover effect. The degree to which the plotting symbols tend to separate along the vertical axis represents the size of the treatment effect.

Outliers are easily detected on this plot. Outlying subjects should be reviewed for data-entry errors and for special conditions that might have caused their responses to be unusual. Outliers should not be removed from an analysis just because they are different. A compelling reason should be found for their removal and the removal should be well documented.

Period Plot

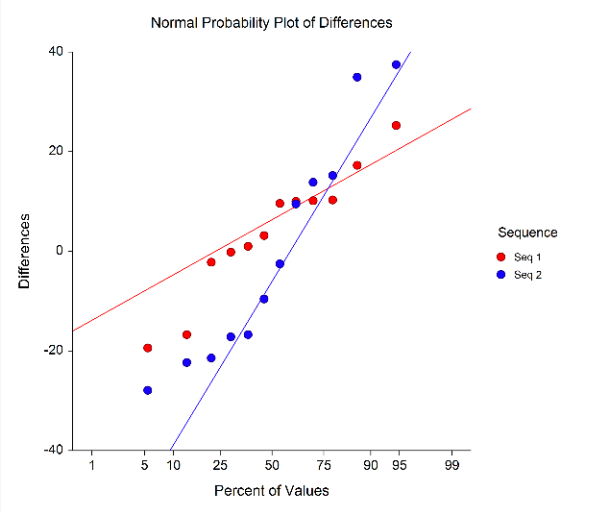
Period Plot



The Period Plot displays a subject's period 1 response on the horizontal axis and the period 2 response on the vertical axis. The plotting symbol is the sequence number. The plot is used to find outliers and other anomalies.

Probability Plots

Probability Plots



These plots show the differences (*P*1-*P*2) on the vertical axis and values on the horizontal axis that would be expected if the differences were normally distributed. The first plot shows the differences for sequence 1 and the second plot shows the differences for sequence 2.

If the assumption of normality holds, the points should fall along a straight line. The degree to which the points are off the line represents the degree to which the normality assumption does not hold. Since the normality of these differences is assumed by the *t*-test used to test for a difference between the treatments, these plots are useful in assessing whether that assumption is valid.

If the plots show a pronounced pattern of non-normality, you might try taking the square roots or the logs of the responses before beginning the analysis.