European Directorate for the Quality of Medicines & HealthCare | Direction européenne de la qualité du médicament & soins de santé

Funded by the European Union and the Council of Europe

EUROPEAN UNION

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Implemented by the Council of Europe

General European OMCL Network (GEON)
GENERAL DOCUMENT

PA/PH/OMCL (21) 26

Statistical analysis of results of biological assays and tests

Full document title and reference	**Statistical analysis of results using CombiStats** *PA/PH/OMCL (21) 26*
Document type	Informative
Legislative basis	Council Directive 2001/83/EC and 2001/82/EC, as amended
Date of first adoption	1st March 2021
Date of original entry into force	1st March 2021
Date of entry into force of revised document	n.a.
Previous titles/other references / last valid version	n.a.
Custodian Organisation	The present document was elaborated by the OMCL Network / EDQM of the Council of Europe
Concerned Network	GEON

1. Assay designs 3
1.1. Randomised block design 3
1.2. Completely randomised design 5
1.3. Latin square design 6
1.4. The value of replication of treatments 7
2. Regression models 11
2.1. Regression analysis 11
2.2. Overview of models 12
2.3. Slope-ratio analysis 12
2.4. Four-parameter logistic regression model 13
2.5. Parallel-line analysis 13
2.6. Other regression models 14
2.7. Data transformations 15
2.8. Weight functions 17
3. ANOVA table 19
3.1. Introduction 19
3.2. Slope-ratio analysis 20
3.3. 4-parameter logistic and parallel-line models 21
3.4. Coefficients of correlation and determination 23
4. Analysis and interpretation of results 25
4.1. Global analysis versus individual analyses 25
4.2. Table of potency estimates 26
4.3. Plot of residuals 28
4.4. Outliers 30
4.5. Non-linearity contrast 31
4.6. Non-parallelism contrast 34
4.7. Use of control charts 38

INTRODUCTION

The purpose of this document is to illustrate some statistical issues presented in Chapter 5.3 of the European Pharmacopoeia - Statistical analysis of results of biological assays and tests. It consists in 4 sections dealing with assay designs, regression models (including assay validity criteria), structure of the ANOVA table and analysis and interpretation of results.

This document could be prepared thanks to the collaboration between statisticians from the OMCL network and from the EDQM. It is intended for information only, and does not replace the statistical requirements found in the chapters and monographs of the European Pharmacopoeia.

1. ASSAY DESIGNS

This section illustrates 3 types of assay designs commonly used in routine testing and inter-laboratory studies, i.e. the completely randomised design, randomised block design, and Latin square design. It addresses also the replication of treatments (defined as doses or dilutions of a test and standard preparations) and the replication of assays.

The randomised block design is presented first, although it is an advanced design compared to the completely at random. The reason is that it allows introducing statistical notions that are needed thereafter to make a clear comparison between the 3 designs.

The explanations and examples given in this section will have met their goal if the reader becomes convinced that i) planning the experiments is crucial to the reliability of results and conclusions, ii) it is too late to address the selection of a type of design once the experiments are done.

1.1. Randomised block design

Getting started

A control laboratory is in charge of testing 2 products (8 vials each). As a maximum of 10 vials can be tested in a run, the analyst is planning one run per product. Each run will require reconstituting a fresh aliquot of a critical reagent, which has a significant effect on the results. Past experiments showed that a difference of 15% could be observed, on average, between results of vials tested on different runs.

Critical assessment

The analyst is interested in calculating the mean difference between two products accurately. However, the proposed design introduces a significant bias. Indeed, in the case where both products would be identical, a mean difference of 15% could still be observed due to confounding with differences between reagent aliquots.

Proposal

Confounding will be avoided by testing vials of each product in parallel, i.e. for a same aliquot of the critical reagent. In this case, two runs can be carried out, in which 4 vials of each product will be tested. Each run represents a block of experiments comparing both products in similar and controlled experimental conditions.

Figure 1 shows that the vials are tested in random order in each block. There is a mean difference of 14% (103.8 / 91.4) between the results of the blocks (due to differences between reagent aliquots), but it does not affect the comparison of products. The mean difference between the two products is rather similar in each block (1.45 and 1.18, respectively). The overall mean difference is 1.31.

Additional blocks of results could be added in order to achieve better precision about the overall means difference. The number of blocks can be determined during the design phase (study protocol) or adjusted after some initial block results are obtained (sequential approach), showing how flexible the randomised block design strategy is.

Design

Block 1	Block 2
P1	P2
P1	P1
P2	P2
P2	P1
P1	P1
P2	P2
P2	P1
P1	P2

P_i: vial of product i

Results

Block 1	Block 2
87.9	103.1
92.8	100.3
90.6	104.7
92.6	103.4
91.9	102.9
93.8	105.8
91.3	103.6
89.9	106.5
Mean	Mean
91.4	103.8

Comparison of means

	Blocks		
Product	**1**	**2**	**1 + 2**
1	90.6	103.2	96.9
2	92.1	104.4	98.2
Mean Diff.	**1.45**	**1.18**	**1.31**

Figure 1. Example of a randomised block design.

CombiStats examples

There are several examples of randomised block designs in the user manual of the CombiStats software. In most cases, all the doses of the standard and test preparations are present in each block (called complete blocks). The blocks allow controlling the heterogeneity of some materials, such as water-bath (Ex. A.1.4), agar plates (Ex. A.2.7) or litters of rats (Ex. A.2.5), which are the block factors in the respective examples.

The user manual introduces the randomised block design in page 19, where the treatments refer to the doses of the preparations (standard, samples):

> *If it is possible to identify an experimental factor that could influence the response of specific groups of units in the same way, the randomised block design may be appropriate. For example, a group of different treatments in a Petri dish might, on average, give a lower response than an identical group of treatments in another Petri dish. Hence, it is important that the treatments be equally distributed over the Petri dishes (the blocks). Applying the same treatment in only one block should absolutely be avoided, as this would confound the effect of the treatment with the block effect, and thus lead to erroneous results.*

Example A.3.4 of the user manual consists in 6 treatments (2 prep. × 3 doses) tested using a randomised block design. There are 6 gels of agar (blocks) on which the 3 doses of the standard and sample preparations are tested in parallel. Figure 2 shows the outcome of the random allocation of the 6 treatments to each block (e.g. 1|1 represents Prep. 1|Dose 1).

The block effect (i.e. heterogeneity between gels) depends on differences between the block means. The mean values range from 204.5 (block 1) to 211 (block 6) and are significantly different according to the analysis of variance (ANOVA table, p-value = 0.001 for the Blocks line). This result shows that the blocking strategy was relevant and useful, given the heterogeneity between gels.

Design	(A)	(B)	(C)	(D)	(E)	(F)
(1)	1|1	1|3	2|2	1|2	2|1	2|3
(2)	2|2	1|2	1|1	2|3	2|1	1|3
(3)	1|2	1|3	2|1	1|1	2|3	2|2
(4)	1|3	2|2	2|1	2|3	1|2	1|1
(5)	2|3	2|1	1|2	1|3	2|2	1|1
(6)	1|3	1|1	2|3	2|2	1|2	2|1

Observ.	(A)	(B)	(C)	(D)	(E)	(F)
(1)	189	222	207	206	185	218
(2)	208	207	191	226	188	220
(3)	208	222	188	190	224	207
(4)	227	210	190	222	210	192
(5)	218	188	202	226	208	191
(6)	227	194	228	208	215	194

Figure 2. Example of randomised blocks in CombiStats. (6 blocks coded from (1) to (6), in which all treatments (Prep.|Dose) are carried out in random order).

1.2. Completely randomised design

Introduction

The completely randomised design can be used when no factor of heterogeneity, such as defined in Section 1.1, is suspected. The doses of the various preparations are thus tested in experimental conditions that are considered to be similar enough. Such conditions can be assumed when, for example, the materials and procedures used for the assay are well standardised.

Specifically, the completely randomised design is recommended when the sources of variation are known to have low and similar contributions to the assay variation. The resulting variation is referred to as the residual error in the penultimate row of the ANOVA table of CombiStats files (section 3.1).

Some bias can still be introduced inadvertently during the course of the assay. It can be prevented by testing the treatments in random order, so that, on average, results remain accurate.

CombiStats example

Example A.2.10 consists in 2 preparations tested at 2 doses (the blank condition, i.e. 0 µg, is excluded). Each of these 4 treatments is to be tested 4 times in a completely randomised design.

The best way to illustrate the fact that the 16 trials are performed in random order is to show the order of trials in a single row (or column) as in Figure 3. For example, the first 2 trials are the first and second replicates of Prep. 2, Dose 2.

Design	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
(A)	2|2|1	2|2|2	1|2|1	2|1|1	2|1|2	1|2|2	1|1|1	1|2|3	2|1|3	1|2|4	2|1|4	1|1|2	1|1|3	2|2|3	2|2|4	1|1|4

Observ.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
(A)	121	124	167	80	88	164	97	159	90	156	82	100	105	122	122	98

Figure 3. Example of a completely randomised design in CombiStats.
Each treatment is tested 4 times (Prep.|Dose|Rep.) in random order.

A more compact layout, i.e. a table with multiple rows and columns, is still possible (Figure 4), although it may be confused with the layout of other designs. In addition, the trial order must be specified in the study protocol: start with the 4 treatments in first row, then those in the second row, and so on.

Design	(1)	(2)	(3)	(4)
(A)	2|2|1	2|2|2	1|2|1	2|1|1
(B)	2|1|2	1|2|2	1|1|1	1|2|3
(C)	2|1|3	1|2|4	2|1|4	1|1|2
(D)	1|1|3	2|2|3	2|2|4	1|1|4

Observ.	(1)	(2)	(3)	(4)
(A)	121	124	167	80
(B)	88	164	97	159
(C)	90	156	82	100
(D)	105	122	122	98

Figure 4. Completely randomised design represented in a 4-by-4 table.
The trial order (by row in this example) must be specified in the study protocol.

1.3. Latin square design

Introduction

This design is of particular interest when the treatments are tested on a support that shows a gradient of heterogeneity. A typical example is the agar gel used in immuno-diffusion assays, which contains an active protein that may not be uniformly distributed/plated.

Figure 5 shows an agar gel with higher (lower) amounts of protein represented by darker (lighter) grey areas. In the first design (Design 1), the 4 replicates of the 4 treatments (2 prep. × 2 doses) are grouped for convenience, but it leads to confounding between the treatment effect and the gradient of protein.

In the Latin square design, the replicates are allocated to $N_{Treat.}^2 = 16$ squares defined on the gel such that each treatment occurs once in each row and column. Doing so, every treatment is tested in areas covering higher and lower amounts of protein. Subsequently, the mean results of the 4 treatments, and thus the potency estimates, will show limited bias.

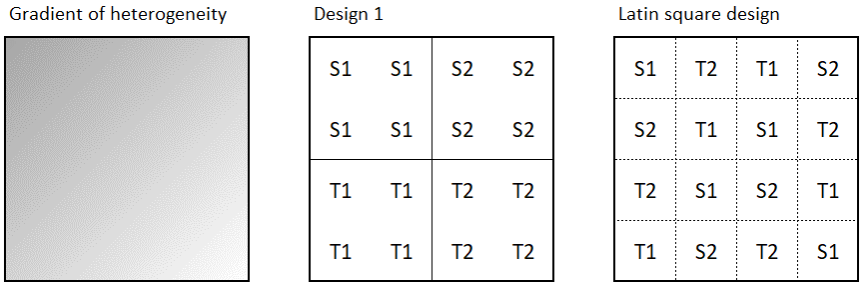


Figure 5. Example of Latin square design used in case of gradient of heterogeneity on an agar gel.

CombiStats example

Example A.1.3 consists in 6 treatments (2 prep. × 3 doses) tested using a 6-by-6 Latin square design, in which each treatment appears once and only once per row and column. CombiStats shows the layout of the design (Figure 6). Note that the treatments should be tested in random order as defined in the study protocol.

Design	(A)	(B)	(C)	(D)	(E)	(F)
(1)	1|1	2|1	2|2	1|3	1|2	2|3
(2)	2|1	2|3	1|1	1|2	2|2	1|3
(3)	2|2	1|3	1|2	1|1	2|3	2|1
(4)	1|3	1|2	2|3	2|1	1|1	2|2
(5)	1|2	2|2	1|3	2|3	2|1	1|1
(6)	2|3	1|1	2|1	2|2	1|3	1|2

Observ.	(A)	(B)	(C)	(D)	(E)	(F)
(1)	161	160	178	187	171	194
(2)	151	192	150	172	170	192
(3)	162	195	174	161	193	151
(4)	194	184	199	160	163	171
(5)	176	181	201	202	154	151
(6)	193	166	161	186	198	182

Figure 6. Example of a 6-by-6 Latin square design in CombiStats.

As a result of the allocation of treatments per row and column, the ANOVA table contains p-values for row and column effects that inform about the direction taken by the gradient of heterogeneity:

- From top to bottom if the row p-value is significant,
- From left to right if the column p-value is significant,
- More complex pattern of heterogeneity if both p-values are significant.

The p-values reported in the ANOVA table (Table 1) indicate a significant 'Row block' effect (p = 0.012) and non-significant 'Column block' effect (p = 0.107). It means that the gradient of heterogeneity on the agar gel is mainly vertical in this case.

Table 1. ANOVA table of the 6-by-6 Latin square design (Ex. A.1.3 of the user manual).

Source of variation	Degrees of freedom	Sum of squares	Mean square	F-ratio	Probability
Preparations	1	11.1111	11.1111	0.535	0.473
Regression	1	8475.04	8475.04	408.108	0.000 (***)
Non-parallelism	1	18.3750	18.3750	0.885	0.358
Non-linearity	2	5.47222	2.73611	0.132	0.877
Standard	1	0.0277778	0.0277778	0.001	0.971
Sample 1	1	5.44444	5.44444	0.262	0.614
Treatments	5	8510.00	1702.00	81.958	0.000 (***)
Rows	5	412.000	82.4000	3.968	0.012 (*)
Columns	5	218.667	43.7333	2.106	0.107
Residual error	20	415.333	20.7667		
Total	35	9556.00	273.029		

1.4. The value of replication of treatments

Replicates are defined as the number of times a treatment is included in the assay. Replicates should be independent, and when each treatment has the same number of replicates, the design is said to be balanced (unbalanced otherwise). These replicates play a significant role in the statistical analysis.

Assay validity

Replicates are involved in the validity assessment of the assay, e.g.:

- The replicated results of a given treatment can be too scattered, denoting a potential experimental error with possible exclusion of results (section 4.4, Ex. A.2.13),
- The variability between replicates can increase with the dose concentration, requiring some refinement of the statistical model (section 2.8, weighted regression in Ex. A.3.17),
- The variability between replicates is used to calculate the Residual error, which is usually the basis for assessing the significance of the various sources of variation listed in the ANOVA table, such as Non-linearity and Non-parallelism of regression lines (Table 1, section 3.1),
- The residual error can be compared with historical data by means of control charts. A residual error being out of the control limits may indicate a problem with the assay (section 0).

Precision of potency estimates

The variability between replicates is also used to calculate 95% confidence limits (95%CL) about potency estimates. These limits represent the precision of the calculated means, and, in many monographs, should fall within some acceptance criteria.

Examples in Ph. Eur. 10th Edition:

- General monograph 2.7.2. Microbiological assay of antibiotics: Unless otherwise stated in the monograph, the confidence limits (P = 0.95) of the assay for potency are not less than 95 per cent and not more than 105 per cent of the estimated potency.
- Monograph 0343. Tetanus antitoxin for veterinary use: The confidence limits (P = 0.95) have been estimated to be: i) 85 per cent and 114 per cent when 2 animals per dose are used; ii) 91.5 per cent and 109 per cent when 3 animals per dose are used; iii) 93 per cent and 108 per cent when 6 animals per dose are used.

The latter example is of particular interest as it shows that the acceptance limits get tighter as more animals (replicates) are included in the assay. This is because the precision improves (95%CL get narrower) with the number of replicates, so that acceptance criteria can be adjusted accordingly.

Residual error (CombiStats example)

Example A.1.1 consists in 6 treatments (3 prep. × 2 doses) tested using a completely randomised design. There are 10 replicates per treatment, which yield a residual error of 765.57. As shown in Table 2, the residual error is the mean of the variances calculated using the 10 replicates of the treatments, and it is expressed in (unit/mg)². The square root of the residual error represents the experimental standard deviation and is equal to SD = 27.7 unit/mg. On average, the analyst can thus expect a difference of about 30 unit/mg between replicated results.

Table 2. Residual error calculation in a completely randomised design.

Treatment	1	2	3	4	5	6
Preparation	S	S	T	T	U	U
Dose (unit)	0.25	1	0.25	1	0.25	1
N rep.	10	10	10	10	10	10
DF	9	9	9	9	9	9
Variance	1026.67	483.822	724.989	718.667	854.622	784.667

DF: degrees of freedom (N rep. - 1)

Residual error = (1026.67 x 9 + 483.822 x 9 + 724.989 x 9 + 718.667 x 9 + 854.622 x 9 + 784.667 x 9) / (6 x 9)

(unit/mg)² = 765.57

The same calculation was performed using the first 5 replicates of each treatment. Table 3 shows the potency estimates and 95%CL calculated by CombiStats for Sample T using both scenarios. Wider 95%CL are observed as the number of replicates has decreased (from 69%-148% of estimate for 10 replicates to 50%-201% for 5 replicates).

Table 3. Potency estimates for 5 and 10 replicated (completely randomised design).

Design	Sample T (first 5 rep.)		
(unit/mg)	Lower limit	Estimate	Upper limit
Potency	0.5208	1.0336	2.0739
Rel. To Est.	50%	100%	201%

Residual error = 741.7 (unit/mg)²

Design	Sample T (10 rep.)		
(unit/mg)	Lower limit	Estimate	Upper limit
Potency	0.7837	1.1421	1.6869
Rel. To Est.	69%	100%	148%

Residual error = 765.6 (unit/mg)²

Note. It is often and erroneously though that the residual error decreases as the number of replicates increases. What actually decreases is the uncertainty about the calculated value, which, on average, tends towards the true experimental error. E.g. a lower estimate is observed for 5 replicates (741.7) than for 10 replicates (765.6). However, the latter should be preferred as it is more precise.

Residual error vs. pure error

The calculation of the residual error was presented for the completely randomised design. In summary, there are 6 treatments of 10 replicates in Ex. A.1.1 for a total of 6 × (10 – 1) = 54 degrees of freedom, as reported for the residual error in the ANOVA table. This error, which derives from replicated results, is the pure error representing the repeatability of the assay. In addition, the number of degrees of freedom informs about how reliable the error term is. With 54 degrees of freedom, the pure error estimated in this example is deemed reliable.

In Ex. A.3.4 of the randomised block design (section 1.1), each of the 6 treatments appears only once in each block. Therefore, there is no pure error and the degrees of freedom are logically equal to 0 in each block (6 × (1 – 1) rep. = 0). However, the ANOVA table contains a residual error reported with 25 degrees of freedom. In this case, the residual error doesn't represent the pure error of the assay (repeatability within a block) but depends on differences between treatments from one block to another. The residual error thus represents the treatment-by-block interaction effect, which number of degrees of freedom is $(N_{Treat.} - 1) \times (N_{Block} - 1) = (6 - 1) \times (6 - 1) = 25$.

The same principle applies to the Latin square design, for which each treatment appears once per row and per column (section 1.3). In the absence of pure error, the residual error consists in the treatment-by-block interaction effect, which number of degrees of freedom is $(N_{Treat.} - 1) \times (N_{Block} - 2) = (6 - 1) \times (6 - 2) = 20$ as in Ex. A.1.3.

Note that the residual error can be made of some pure error and treatment-by-block interaction effect. Assume, for example, a randomised block design consisting in 4 treatments tested in duplicates in 5 complete blocks. The residual error has 32 degrees of freedom, distributed as 4 × (2 – 1) = 4 degrees of freedom per block for the pure error (4 × 5 = 20 in total) and (4 – 1) × (5 – 1) = 12 degrees of freedom for the treatment-by-block interaction effect.

Power and sample size calculation

Sample size calculation consists in defining the number of replicates per treatment required to achieve a given level of quality of information. The example in Table 3 shows that the precision (half-width of the 95%CL) about potency estimates improves as the sample size increases. With acceptance criteria set in many monographs, precision is probably the first element that motivates the calculation of an appropriate sample size.

In addition, CombiStats performs a series of statistical tests to assess the significance of the sources of variation listed in the ANOVA table (Table 1), e.g.:

- Preparations: test for significant differences between the preparation intercepts,
- Regression: test the common slope to 0 (i.e. is there a significant dose effect),
- Non-parallelism: test for significant differences between the preparation slopes.

These tests are characterised by some statistical power, given as a percentage that represents the ability to detect a minimum difference, say 1.5 IU/mL between preparation intercepts. The power is often set to a minimum of 80%, meaning that the test will be significant in more than 80% of routine assays when the difference between intercepts is truly equal to or greater than 1.5 IU/mL.

Once the minimum difference is defined, the number of replicates per treatment is calculated to ensure that the statistical power will be met. The analyst is in charge of defining the minimum difference and statistical power, while the statistician usually performs the sample size calculation.

The number of replicates per treatment will logically increase if the analyst wants:

- To detect a lower difference, e.g. 1 IU/mL instead of 1.5 IU/mL between intercepts,
- To achieve a higher statistical power, e.g. 90% instead of 80%.

The number of replicates depends also on other elements, among which:

- The number of preparations and doses (and distribution along the dose-range),
- The type of design (completely at random, blocked designs),
- The confidence level $(1 - \alpha)$ and,
- The experimental residual error.

In practice, the first 3 elements are usually fixed, and depends mainly on the study objectives and experimental constraints. The residual error can be determined using historical data (validation and/or past routine data). A higher residual error (less reproducible assay) will increase the number of replicates needed to detect the minimum difference or reach the required precision about the potency estimate with adequate statistical power.

2. REGRESSION MODELS

2.1. Regression analysis

CombiStats performs a same regression analysis whatever the type of multiple-dilution assay (slope-ratio, parallel-line, sigmoid curve, quantal response). This analysis includes the estimation of individual slopes and intercepts for the various preparations, as a start to further calculations used to build the ANOVA table and estimate the potency results and/or effective doses.

This regression analysis uses all the dilution points, even those in the lower and upper plateaus of sigmoid curves. In this case, CombiStats performs a linearisation transformation prior to fit the regression model. Non-linearised and linearised values are available from the menu Options > Advanced > Export Matrices > Dataset. They are referred as to *NLinObs* and *LinObs* in the export, respectively.

Table 4 shows the values calculated for Ex. A.3.21 (assay based on quantal responses). Raw data are numbers of respondents out of 8 animals. Non-linearised values are the corresponding proportions (e.g. 8/8 = 1 or 100%) and linearised values are calculated after probit transformation of the proportions (section 2.7). The transformation involves some weighting of the data (*LinWeight* and *NLinWeight* in the export), without which the linearisation cannot be happen.

Table 4. Raw data and linearised values (Ex. A.3.21).

Dose (log)	-3.5	-4	-4.5	-5	-5.5	-6	-6.5	-7	-7.5	-8
Raw data	8/8	8/8	7/8	6/8	2/8	1/8	1/8	0/8	0/8	0/8
Non-Lin. values	1	1	0.875	0.75	0.25	0.125	0.125	0	0	0
Lin. Values	3.05	2.40	1.15	0.67	-0.64	-1.14	-0.78	-2.84	-3.51	-4.21

Linearised and non-linearised values are represented by crosses and dots in Figure 7, respectively. CombiStats performs the regression analysis on linearised values. Predicted values (dotted line on left panel) are referred as to *LinPred* in the export of results. These values are back transformed to obtain non-linearised predicted values (*NLinPred*, continuous line on right panel).

The regression parameters calculated using linearised values can be exported using the menu Options > Advanced > Export Matrices > Linear (slope: Slpe1 = 0.6462; intercept: Icpt1 = 0.1953 for Ex. A.3.21).

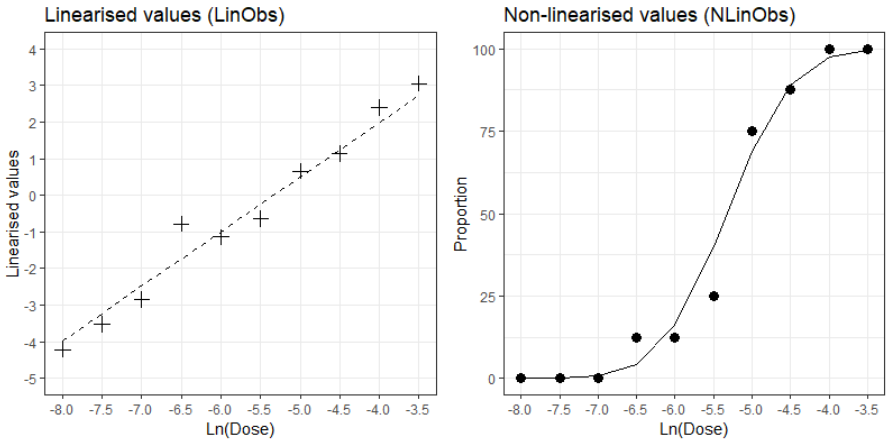


Figure 7. Plot of linearised and non-linearised values (Ex. A.3.21).

2.2. Overview of models

The selection of the regression model depends first on the characteristics of the multi-dilution assay, mainly the type of response variable and dose-response relationship, as shown in Table 5:

- The response variable (Y) can be either quantitative (e.g. absorbance values as in immuno-assays) or quantal reported as a number of events out of a total (e.g. 5 positive cases out of 10 as in *in-vivo* assays).
- The dose-response relationship is usually determined during the development of the assay. If the dilution factor chosen by the analyst is additive and the quantitative response varies in a linear manner with the doses (e.g. 0, 2, 4, 6, 8), then the slope-ratio analysis is appropriate (Ex. A.1.6). If the selected dilution factor is multiplicative (fold-ratio) and the response describes a sigmoid curve after the doses (e.g. 1, 2, 4, 8, 16, 32, etc.) are log-transformed, then the 4-parameter logistic regression is appropriate (Ex. A.1.8 for quantal responses and Ex. A.1.14 for quantitative responses).

Table 5. Regression models in CombiStats.

Response (Y)	Dose scale	x-axis	Shape	Regression model
Quantitative	Additive	Dose	Straight lines	Slope-ratio analysis (SRA)
	Fold-ratio	Ln(Dose)	Straight lines	Parallel-line analysis (PLA)
	Fold-ratio	Ln(Dose)	Sigmoid curve	4-parameter logistic regression (4PL)
Quantal	Fold-ratio	Ln(Dose)	Sigmoid curve	4-parameter logistic regression (4PL)

Doses are equidistant on the x-axis of the regression plot of the slope-ratio model.
Similarly, Ln(Doses) are equidistant on the x-axis of the regression plot of the parallel-line model.

2.3. Slope-ratio analysis

In the slope-ratio analysis, a linear regression model (of the form Y = a + bX + error; a: intercept, b: slope) is fitted for each preparation. As doses are expressed on an additive scale, the regression lines diverge if the preparations have different strengths (Figure 8). For this reason, potency estimates depend on ratios between preparation slopes.

In Ex. A.1.6, the slopes are 0.2467 and 0.2031 for the standard and test preparations, respectively. The relative potency estimate of the test preparation is RP = 0.2031 / 0.2467 = 0.823 (the test preparation is 17.7% less potent than the standard preparation). The potency of the test preparation is equal to the potency of the standard preparation multiplied by the RP, i.e. 1 RP/volume × 0.823 = 0.823 RP/volume.

Linearity of the regression lines is a typical assay validity criterion. In addition, multi-dilution assays assume that the preparations are similar (that they act as dilution of the same substance). In a slope-ratio analysis, similarity is demonstrated if the regression lines intersect each other at zero-dose (common intercept assumption). The equality of regression intercepts is thus evaluated (Intersection source of variation in the ANOVA table) and constitutes another validity criterion.

The additive dose-scale also offers the possibility to consider blank (zero-dose) results during the statistical evaluation. The difference between the mean intercept of the preparations and the mean of the blanks is calculated, comparing the signal of the product matrix to the background signal of the assay.

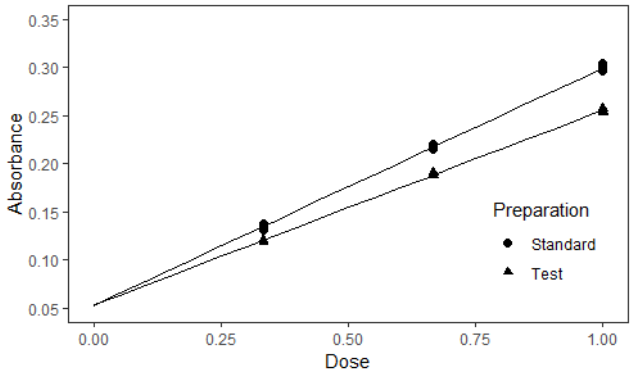


Figure 8. Slope-ratio model.

2.4. Four-parameter logistic regression model

The 4-parameter logistic regression model (4-PL) is used in case of sigmoid curves. This model is defined by lower and upper asymptotes (a and d parameters in the equation below), a slope (b) and a mid-point (c). In case of quantitative response variables, this model requires enough doses (usually about 10 x-fold dilution points including 4 to 5 points in the linear part of the sigmoid) to cover the sigmoid, from product matrix/background signal (lower asymptote) to saturation (upper asymptote).

$$Y = a + \frac{d - a}{1 + \left(\frac{Dose}{c}\right)^{b}} + error$$

As doses are expressed on a log-scale, there is a shift between the sigmoid curves if the preparations have different strengths. In other words, the (relative) potency depends on the difference between the preparation mid-points.

In Ex. A.1.14, the mid-points are $5.39071E^{-3}$ IU and $1.84758E^{-3}$ IU for the standard and test preparations, respectively (the assumed potency of the test preparation was set to 0.2 IU/mL in the data table, instead of "? IU/mL", for the sake of the presentation). The relative potency is RP = 5.39071 / 1.84758 = 2.918. On the regression plot (Figure 9), the two regression lines are distant by ln(RP) = 1.07 Ln(IU) on the x-axis. The potency of the test preparation is equal to the assumed potency of the preparation multiplied by the RP, i.e. 0.2 IU/mL × 2.918 = 0.5836 IU/mL.

As for the slope-ratio analysis, linearity of regression lines and similarity of preparations should be demonstrated. In 4-PL regression models, these features are evaluated after linearization of the sigmoid curves (section 2.1). In particular, similarity is demonstrated if the linearised regression lines are parallel (common slope assumption). Non-linearity and non-parallelism are thus evaluated (ANOVA table) and constitute two assay validity criteria.

The sigmoid curve of the 4-parameter logistic model is symmetrical on each side of the mid-point. A fifth parameter (i.e. asymmetry factor as in Ex. A.3.32) can be added to the model in case of lack of fit between the observed results and the fitted curve.

2.5. Parallel-line analysis

The 4-parameter logistic and parallel-line models share the same selection criteria: a quantitative response and log-transformed doses. The use of one or the other will usually depends on additional considerations. First, let us recall that the 4-parameter logistic model usually requires about 10 dilution points. The analyst may not be interested in the dilution points corresponding to the lower and upper asymptotes of the sigmoid curve, especially if he is testing similar preparations on a routine

basis. In such a case, the analyst may decide to keep the dilution points corresponding to the steepest and linear part of the sigmoid (Figure 9). The parallel-line model is then appropriate (Ex. A.1.1). Because this model usually requires about 5 dilution points, the analyst can test more preparations on the same plate or in the same run.

In the parallel-line analysis, a linear regression model (of the form Y = a + bX + error; a: intercept, b: slope) is fitted for each preparation. As for the 4-parameter logistic model, doses are expressed on a log-scale, with the same two main outcomes:

i) The regression lines should be parallel (common slope assumption). This validity criterion, if not met, may indicate non-similarity of preparations,

ii) The potency estimates depend on distances between the regression lines of the preparations, that is, differences between preparation intercepts (relative potency = $(a_{Test} - a_{Std}) / b_{Common}$).

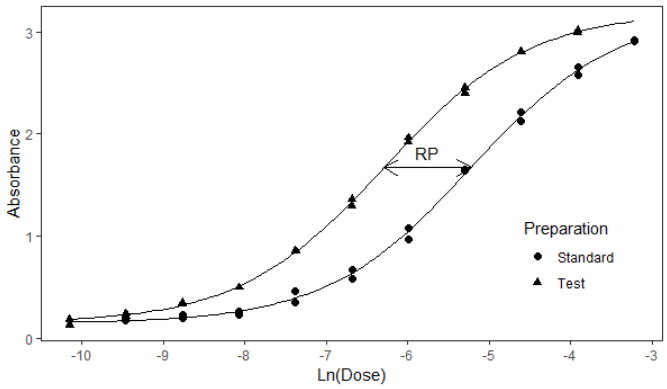


Figure 9. Sigmoid curves (4-parameter logistic model). The relative potency (RP) depends on the difference between the mid-points of the preparations. The steepest part of the curves can be used to fit a parallel-line model.

Note. The parallel-line model can be used on a subset of dilution points selected in the steepest part of the sigmoid curve (4-PL model). It may not work in the other direction, i.e. adding extra dilution points on each side of a series of dilution points showing satisfactory linearity of the dose-response relationship may not result in a sigmoid curve.

2.6. Other regression models

Multiple dilution assays

For multiple dilution assays based on quantal responses, CombiStats can estimate potency results and effective doses using the Spearman-Kaerber method. The user cannot select this method, which is not described in Ph. Eur. Chapter 5.3 as it is an empirical approach and not a regression analysis. As a result, no ANOVA table is produced, with no possibility to assess non-linearity nor non-parallelism of regression lines.

However, CombiStats will automatically invoke this method when it cannot fit the 4-parameter logistic regression. A typical example is the lack non-extreme responses for the selected dose range, which precludes the calculation of the slope parameter. Example A.3.31 illustrates this issue, also known as complete or quasi-complete separation (Table 6). From a design perspective, the analyst may consider revising the dose range.

"Spearman-Kaerber method used" is the message that appears on the CombiStats sheet, above the tables of potency estimates. 95% confidence limits of potency estimates are calculated using the Irwin/Cheeseman formula.

Table 6. Example of complete separation in models based on quantal responses (Results come down to 2 groups with 0% and 100% probability levels).

Dose	1/1	1/2	1/4	1/8	1/16	1/32	1/64	1/320
Resp.	0/4	0/4	0/4	0/4	4/4	4/4	4/4	4/4

Single dilution assays

CombiStats can also compare the results of different preparations in case of single dose assays. Specifically, the software performs the one-sided test of Wilcoxon-Mann-Whitney to assess the null hypothesis of equality between:

- The actual potency of the test preparation, which is unknown, and,
- The assumed potency of the test preparation, which is calculated using information about predilutions and the assigned potency of the standard preparation.

If the null hypothesis is rejected, the analyst can conclude that the test preparation contains significantly more (or less) than the assumed potency. Therefore, this test is referred as to a limit test in CombiStats. Prior to run the test, the analyst should select Completely randomised from the Design tab of the Options Wizard as well as Parallel-lines or Quantal responses from the Model tab for quantitative and qualitative results, respectively.

2.7. Data transformations

The statistical analysis of multi-dilution assay results involves different types of data transformations. As shown in Table 5, doses (x-axis) are log-transformed in all models, with the exception of the slope-ratio model. Thus, the transformation applied to the doses is model-dependent. The response variable (y-axis) can be transformed as well for different purposes:

- The analyst can apply a data transformation to improve the fit of the slope-ratio model or parallel-line model. Example A.1.5 can be run with and without the logarithm transformation to assess its benefit on the linearity and parallelism of the regression lines. Although the logarithm transformation is likely to be the most frequent, several other transformations of the response variable are available in CombiStats. More generally, these transformations are referred as to power transformations, e.g.:

Label	Power	Comments
Inverse (1/y)	$\rightarrow y^{-1}$	The log transformation is a special case where the power transformation is defined as y^0.
Square root ($\sqrt{y}$)	$\rightarrow y^{0.5}$	
Square (y^2)	$\rightarrow y^2$	

- Sigmoid models for quantitative and quantal responses are characterised by a linearization of the dose-response relationship, prior to fit the regression model (section 2.1). This linearization always involves a data transformation of the response variable. Thus, the transformation is model-dependent. The default transformations are Logit and Probit for quantitative and quantal responses, respectively. However, other transformations are available as shown in Table 7.

Table 7. Data transformations. Model-dependent transformations are in italic.

Regression model	X-scale	Y-scale
Slope-ratio analysis (SRA)	*Dose*	Power transformations (y^{-1}, y^0, $y^{0.5}$, y^2, etc.)
Parallel-line analysis (PLA)	*Ln(Dose)*	Power transformations (y^{-1}, y^0, $y^{0.5}$, y^2, etc.)
Sigmoid (4PL), quantitative	*Ln(Dose)*	*Probit, Logit (default), Angular, Rectangular, Gompertz*
Sigmoid (4PL), quantal	*Ln(Dose)*	*Probit (default), Logit, Angular, Rectangular, Gompertz*

Model-dependent transformations are combined with some weighting of the data required to linearise the dose-response relationship.

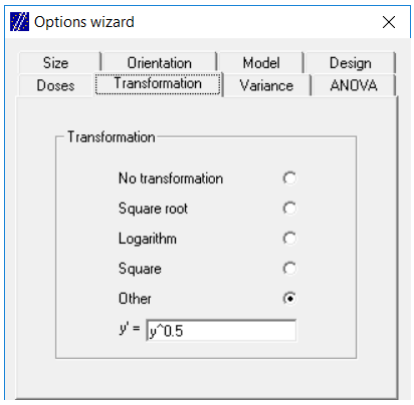
Figure 10 shows the two dialog boxes available from the Options wizard to select the transformations of the response variable. With regards the linearization of sigmoid curves, the Gompertz transformation is applicable to asymmetrical curves with a shorter lower tail and longer upper tail. The Probit, Logit, Angular and Rectangular transformations are applicable to symmetrical curves, as shown in Figure 11:

- Probit: short tails, i.e. sigmoid curves reach the minimum and maximum y-values rapidly,
- Logit: long tails, i.e. sigmoid curves reach the minimum and maximum y-values slowly,
- Angular and rectangular: almost no tails, i.e. sigmoid curves reach the minimum and maximum y-values even more rapidly than in the case of the Probit.

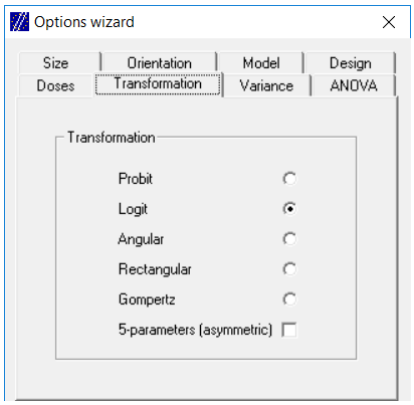
Note. The “5-parameters” check box adds an additional regression parameter to the 4-parameter logistic model to account for potential asymmetry of the sigmoid curve.

The shape of the sigmoid curves obtained during the validation exercise and first routine assays together with the knowledge about the response variable (e.g. based on growth, tolerance) can be used to support the selection of the most appropriate transformation.

Transformations of response variables



Linearization of sigmoid curves



y^0.5 corresponds to the square root of y.

Figure 10. Dialog boxes for the selection of transformations.

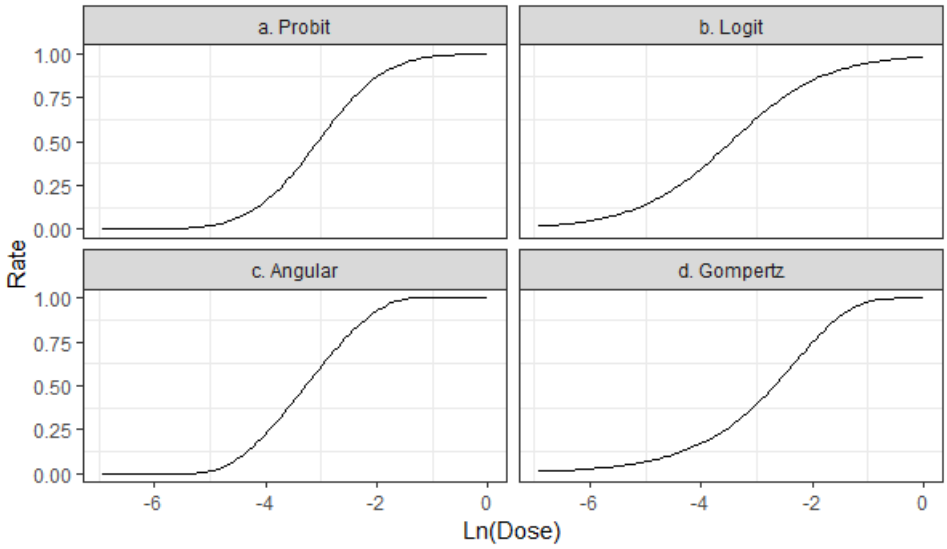


Figure 11. Examples of possible shapes of sigmoid curves.

In summary, data transformations are used to improve the model fit. The most appropriate transformations (of x- and y-values) are usually defined during the development of the method and should not be modified from one routine assay to another. Routine assays should be analysed in a consistent way. Otherwise:

- Sporadic changes of the transformation may be indicative of errors during the experimental work (e.g. dilution error), reporting or analysis of results. This assumption can be reinforced by unexpected values obtained for other statistical parameters.
- Time-to-time changes of the transformation may be indicative of a poorly developed assay or lack of command of the method. Assay optimisation (e.g. redefinition of the optimal dose-range) and better standardisation of the operating procedure should be carried out to improve the reliability and comparability of future assay results.
- Recurrent use of another transformation than the one selected initially may be indicative of a permanent change. The analyst should evaluate how this change affects the assay results (that may be systematically under or overestimated) and take appropriate actions, if needed.

2.8. Weight functions

The regression models applied to quantitative data assume a constant variance among replicated results across the dose range. Graphically, the spread of the data points look rather similar from one dose to another (Ex. A.1.1). This variance is the residual error in the ANOVA table (complete randomised design). Weights are set to 1 (w = 1), and all the observations contribute equally to the regression fit.

In some assays, however, the variance among replicates can depend on the dose. In Ex. A.3.17, the spread of the data points, represented by the standard deviation in Figure 12, increases with the dose (in a ratio of 1 to 6 between D01 and D11 for Sample 1 and in a ratio of 1 to 9 for Sample 2). In such a case, it is usual to calculate the relative standard deviation, which tends towards a common value (rsd = 4% here). CombiStats offers the possibility to address this heterogeneity of variance by performing a weighted regression analysis with weights inversely proportional to the variance observed at each dose. To do so, the analyst can indicate w = 1 / (m*m) as weight function.

Note. Weights are available from the Advanced options dialog box only (F12 shortcut).

Example A.3.30 shows another application of the weight function used to alleviate the negative effect of outliers on the regression fit. To do so, the analyst can indicate w = h as weight function. CombiStats performs a regression analysis where outlying results are replaced by values calculated according to the Huber's robust approach.

	Doses, Sample 1										
Rep.	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11
1	0.217	0.209	0.214	0.233	0.291	0.395	0.497	0.696	0.644	0.877	0.786
2	0.208	0.225	0.222	0.235	0.289	0.415	0.561	0.675	0.762	0.801	0.853
3	0.206	0.222	0.212	0.254	0.310	0.429	0.564	0.681	0.760	0.891	0.793
4	0.212	0.219	0.212	0.235	0.309	0.416	0.570	0.682	0.742	0.788	0.815
Std.dev	0.005	0.007	0.005	0.010	0.011	0.014	0.034	0.009	0.056	0.052	0.030
Rsd	**2%**	**3%**	**2%**	**4%**	**4%**	**3%**	**6%**	**1%**	**8%**	**6%**	**4%**

	Doses, Sample 2										
Rep.	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11
1	0.204	0.209	0.214	0.253	0.320	0.414	0.516	0.673	0.708	0.795	0.847
2	0.210	0.212	0.216	0.241	0.295	0.386	0.506	0.643	0.708	0.843	0.858
3	0.210	0.230	0.215	0.235	0.277	0.396	0.514	0.619	0.701	0.808	0.777
4	0.215	0.211	0.219	0.243	0.262	0.378	0.532	0.606	0.710	0.768	0.784
Std.dev	0.005	0.010	0.002	0.007	0.025	0.016	0.011	0.029	0.004	0.031	0.042
Rsd	**2%**	**5%**	**1%**	**3%**	**9%**	**4%**	**2%**	**5%**	**1%**	**4%**	**5%**

Figure 12. Raw data in Ex. A.3.17.

The two previous examples show that the analyst can define his own weight functions. However, this function can also be model-dependent. It is the case for quantal responses, which differ from quantitative responses by, among others, the nature of the probability distribution:

- Quantitative responses follow normal distributions defined by a mean and a variance that are independent parameters. Observed mean values can be calculated at each dose and predicted by the regression line fitted across the dose range. If the variances are equal from one dose to another, no weight function is required (w = 1). Otherwise, the analyst can adjust it.

- Rates (e.g. 1/10, 3/10, 5/10) in quantal responses follow binomial distributions, which means and variances are dependent parameters. The observed mean values are simply the rates, e.g. $m_1 = 0.1$, $m_2 = 0.3$ and $m_3 = 0.5$. Variances are calculated as $v = m \times (1 – m)$, i.e. $v_1 = 0.09$, $v_2 = 0.21$ and $v_3 = 0.25$. Since variances are unequal from one dose to another, CombiStats performs a weighted regression analysis with weights inversely proportional to the variances. The weight function is model-dependent and equal to $w = n / (m*(1 – m))$ (where n is the size of the group, e.g. 10 animals/group).

In addition, the weight function should be adapted when the response variable represents counts like in bioassays based on CFU or PFU (colony- or plaque-forming units). Counts follow a Poisson distribution, which variance is equal to the mean (v = m). Therefore, the weight function should be set to w = 1/m for a correct evaluation of the assay results.

In summary, weight functions, just like data transformations, can be used to improve the model fit. Some weight functions are model-dependents. Others can be defined during the development of the method and should be used consistently in routine (unless intended modifications of the method affect the weight function). Otherwise, the use of another weight function in routine may be indicative of an issue. The analyst should run an investigation and take appropriate actions, if needed.

3. ANOVA TABLE

3.1. Introduction

The analysis of variance (ANOVA) table provides a summary of the fitted model, with focus on the evaluation of the validity criteria. The content of the ANOVA table should be interpreted in logical order.

The first question to address is to know whether the different preparations tested at different doses led to some significant change of the signal/response variable. As the combinations of preparations and doses are referred to as treatments in Ph. Eur. Chapter 5.3, the requested information appears in the "Treatments" line of the ANOVA table (Table 8). However, depending on the display options in CombiStats, the Treatments effect can be labelled as "Full factorial" as well.

Table 8. ANOVA Table – Treatment effect (Ex. A.2.8).

Source of variation	DF	Mean squares	F-ratio	Probability
Treatments	8	12.0863	621.579	0.000 (***)
Residual error	9	0.0194444		
Total	17	5.69794		

DF: degrees of freedom

A summary of the assay design is required to evaluate the content of the ANOVA table correctly. Example A.2.8 is a completely randomised design in which the standard and sample preparations were tested at 5 and 3 doses, respectively. There were 2 replicates per dose and 2 blank results (zero-dose).

Degrees of freedom

With regards the degrees of freedom, there are:

- 9 treatments in total, i.e. (1 std × 5 doses) + (1 test × 3 doses) + (1 zero-dose). The Treatments effect has 9 – 1 = 8 degrees of freedom,
- 9 treatments × (2 rep. – 1) = 9 degrees of freedom for the residual error. In addition, the residual error represents the pure error (repeatability) of the assay as the design is completely randomised (Section 1.4). The residual error is s^2 = 0.0194444 $Unit^2$. The corresponding experimental SD is s = 0.139 Unit.

Mean squares

An introduction to Mean squares (MS) is needed to continue interpreting the ANOVA table:

- Residual error: with 9 treatments of 2 replicates, 9 variances of repeatability can be calculated. The mean squared error is the average of the 9 variances, also called pooled variance.
- Treatments: the overall mean is first calculated (mean of 18 results) and then the mean of each treatment (2 results/treatment). MS-Treatments depends on the squares of the differences between the 9 treatment means and the overall mean. The greater the differences between the treatment means, the higher the MS value, the more likely the treatments to have a statistically significant effect on the assay signal.

The significance of the Treatments effect is obtained by calculating the ratio between MS-Treatments and MS-Error = 12.0863 / 0.0194444 = 621.579. This ratio consists in performing an F-test (F-ratio)

that tests the hypothesis of "no treatment effect", i.e. "no differences between the 9 treatment means". This hypothesis is unlikely given the high ratio (622) between MS-Treatments and the variance of repeatability. A probability (p-value = 0.000, i.e. less than 0.1%) can be calculated to further illustrate this outcome.

p-values

The treatment means are reported in the next table. The p-value represents the probability of getting these values just by chance, i.e. if we assume that no differences are expected at all. This scenario is of course unlikely, in light of the experimental SD (0.139). More precisely, it has less than 0.1% chances to occur.

Treat.	S-0.05	S-0.10	S-0.15	S-0.20	S-0.25	T-1.0	T-1.5	T-2.0	Blank
Mean	3.4	4.9	6.2	7.9	9.5	4.9	6.4	7.7	1.5

E.g. S-0.05 stands for the standard preparation at 0.05 µg, T-1.0 for the Test preparation at 1.0 mL in Ex A.2.8.

P-value are reported with 3 decimal places in CombiStats, with very low p-values appearing as 0.000. P-values that indicate a statistically significant effect are followed by 1, 2 or 3 stars:

- **No star:** the p-value is greater than or equal to 0.05 (5%), which is the usual significant threshold in statistics. The null hypothesis of "no effect" cannot be rejected.
- **(*):** the p-value is greater than or equal to 0.01 (1%) and lower than 0.05 (5%). The effect is significant.
- **(**):** the p-value is greater than or equal to 0.001 (0.1%) and lower than 0.01 (1%). The effect is highly significant.
- **(***):** the p-value is lower than 0.001 (0.01%). The effect is very highly significant.

In summary

The F-test has concluded to significant differences among the 9 treatment means. If it would not be the case, the analyst could stop the statistical analysis and most likely invalid the assay.

The information contained in the Treatments effect remains rather non-specific, however. The next step consists in assessing the significance of additional effects addressing specific questions, in relation with the assay design.

3.2. Slope-ratio analysis

With 2 preparations tested at several doses plus some blank results as in Ex. A.2.8, the Treatments effect can be split into 4 effects that can be used to validate the assay (Table 9). Note that the sum of the degrees of freedom of these effects is equal to the number of degrees of freedom of the parent Treatment effect. More generally, the ANOVA table of the slope-ratio analysis depends on the assay design. For example,

- The Blanks effect requires a zero-dose experimental condition,
- In case of one preparation only, there will be no Intersection effect (no preparation intercepts to compare),
- In case of 2 doses per preparation, there will be no Non-linearity effect (a minimum of 3 doses are needed to assess deviation from linearity).

Table 9. Effects in the ANOVA table of a slope-ratio analysis (Ex. A.2.8).

Treatments	DF = 8	Questions
├ Regression	2	Are preparation slopes significantly different from 0 (significant dose effect)?
├ Blanks	1	Is there any significant difference between the mean intercept and the mean of blank results?
├ Intersection	1	Is there any significant difference between preparation intercepts?
└ Non-linearity	4	Is there any significant deviation from linearity of the regression lines?
Residual Error	9	

CombiStats performs F-tests to assess the significance of the various effects (Table 10). For example, F-ratio = 0.144689 / 0.0194444 = 7.441 for the Blanks effect. The p-value (0.023) is below the 0.05 significant threshold, meaning that there is a significant difference between the mean intercept and the mean of blank results.

As part of the validity assessment of the assay results, the Regression effect should be significant (p-value ≤ 0.05), while the Intersection and Non-linearity effects should not be significant (p-value > 0.05). With p-values equal to 0.000, 0.287 and 0.309, respectively, these validity criteria are met. Note that CombiStats further splits the Non-linearity effect into individual effects (i.e. due to the standard and due to sample 1).

Table 10. ANOVA table of a slope-ratio analysis (Ex. A.2.8).

Source of variation	Degrees of freedom	Sum of squares	Mean square	F-ratio	Probability
Regression	2	96.4116	48.2058	>1000	0.000 (***)
Blanks	1	0.144689	0.144689	7.441	0.023 (*)
Intersection	1	0.0248773	0.0248773	1.279	0.287
Non-linearity	4	0.108833	0.0272083	1.399	0.309
Standard	3	0.0880000	0.0293333	1.509	0.278
Sample 1	1	0.0208333	0.0208333	1.071	0.328
Treatments	8	96.6900	12.0863	621.579	0.000 (***)
Residual error	9	0.175000	0.0194444		
Total	17	96.8650	5.69794		

3.3. 4-parameter logistic and parallel-line models

The 4-parameter logistic model and parallel-line model both share the same structure of ANOVA table.

Table 11 shows how the Treatments effect can be turned into specific information. Example A.1.14 consists in 20 treatments (2 prep. × 10 doses) tested using a completely randomised design. There are 2 replicates per treatment and the results are fitted using a 4-parameter logistic model.

More generally, the ANOVA table depends on the assay design. For example,

- In case of one preparation only, there will be no Preparations effect, nor Non-parallelism effect. Indeed, with only one intercept and slope, there are no further intercepts (or slopes) to compare.

- In case of 2 doses per preparation, there will be no Non-linearity effect. Indeed, a minimum of 3 doses are needed to assess deviation from linearity in parallel-line models (4-parameter logistic model require even more doses to reach the lower and upper asymptotes).

Table 11. Effects in the ANOVA table of a 4-parameter logistic model (Ex. A.1.14).

Treatments	DF =	19	Questions
├ Preparations		1	Is there any significant difference between preparation intercepts?
├ Regression		1	Is the common slope significantly different from 0 (significant dose effect)?
├ Non-parallelism		1	Is there any significant difference between preparation slopes?
└ Non-linearity		16	Is there any significant deviation from linearity of the regression lines?
Residual Error		20	

CombiStats assesses the significance of the various effects (Table 12), using F-tests for parallel-line analyses and Chi-square tests for 4-parameter logistic models. For example, Chi-square = ($MS_{Prep.}$ × $DF_{Prep.}$) / MS_{Error} = (0.000756861 × 1) / 0.00142898 = 0.529653 for the Preparations effect. With such a low ratio, it is likely that the effect is not significant. Indeed, the p-value (0.467) is almost ten times higher than the 0.05 significant threshold.

As part of the validity assessment of the assay, the Regression effect should be significant (p-value ≤ 0.05), while the Non-parallelism and Non-linearity effects should not (p-value > 0.05). With p-values equal to 0.000, 0.830 and 0.918, respectively, these validity criteria are met.

Table 12. ANOVA table of a 4-parameter logistic model (Ex. A.1.14).

Source of variation	Degrees of freedom	Sum of squares	Mean square	Chi-square	Probability
Preparations	1	0.000756861	0.000756861	0.529653	0.467
Regression	1	9.43054	9.43054	6599.51	0.000 (***)
Non-parallelism	1	6.55525E-05	6.55525E-05	0.0458738	0.830
Non-linearity	16	0.0127084	0.000794275	8.89337	0.918
Standard	8	0.00764179	0.000955224	5.34774	0.720
Sample 1	8	0.00506661	0.000633326	3.54562	0.896
Treatments	19	9.44407	0.497056	6608.98	0.000 (***)
Residual error	20	0.0285795	0.00142898		
Total	39	9.47265	0.242888		

The analyst should check the validity criteria in logical order. In practice, he should start by the bottom of the ANOVA table and go up line-by-line, stopping at the line where a validity criterion is not met:

1. **Residual error:** the analyst can check whether the error term is consistent with those of previous assays (by means of a control chart for example). A high value may be indicative of outlying results due to some mistakes (e.g. experimental, reporting). A low value may be of concern too as the residual error appears at the denominator of the statistical test. F-ratios or chi-square values will increase artificially with Non-linearity and Non-parallelism criteria very likely to fail. This issue is further discussed in sections 4.5 and 4.6.
2. **Treatments:** this overall effect should be significant, due to, at least, the dose range selected to get a significant slope (significant dose-response relationship). The steeper the

slope, the better the precision about the potency estimates (i.e. the narrower the 95% confidence limits).

The Treatments effect is split into further effects in the next lines of the ANOVA table, which help assessing the validity of the assay. The first of these effects is the Non-linearity effect.

3. **Non-linearity:** should be non-significant (p-value > 0.05) to conclude to linearity of the dose-response relationship. This effect is applicable to all regression models, including sigmoid models for which CombiStats performs a linearization transformation first. In other words, all the dilution points, including those in the asymptotes, take part in the evaluation of the Non-linearity effect. The number of degrees of freedom is equal to the number of dilution points – 2 for each preparation. Therefore, there are 8 degrees of freedom for each preparation (16 in total) for the sigmoid model presented in Table 12.

   Note. The Non-linearity overall effect (i.e. pooled across the preparations) is used in Ph. Eur. Chapter 5.3 as part of the assay validity assessment. However, the analyst may pay attention to the individual effects (i.e. of each preparation) reported in the ANOVA table of CombiStats, when the p-value of the overall effect is close to the significance threshold. Indeed, some of the individual effects may be statistically significant. The analyst may decide to take appropriate actions in this case.

4. **Non-parallelism:** a key assumption is that the test and standard preparations are similar. In this case, they act as dilution of the same substance, which implies that the regression lines are parallel. Therefore, the Non-parallelism effect should be non-significant (p-value > 0.05) to conclude to equality of slopes, i.e. similarity of preparations.

5. **Regression:** the p-value should be ≤ 0.05 to conclude to a significant dose-response relationship. The steeper the common slope, the better the precision of the potency estimates (narrower 95%-CL).

6. **Preparations:** compares differences between regression intercepts linked to the strength (potency) of the preparations. The p-value should be ≤ 0.05 if some difference is expected, > 0.05, otherwise. The analyst may also decide not to set any validity criterion for this effect.

3.4. Coefficients of correlation and determination

In classical slope-ratio and parallel-line models, CombiStats calculates the coefficient of correlation |r|, and reports its absolute value comprised between 0 and 1. This coefficient is linked mainly to the regression effect of the ANOVA table: a very significant slope usually results in a high coefficient of correlation.

The square of r (r^2) is the coefficient of determination, which represents the percentage of the variation of the results that is explained by the fitted model. In Ex. A.1.1, r = 0.7654 (76.54%) and r^2 = 0.5858 (58.58%). The regression model explains less than 60% of the variation of the experimental results. This rather low percentage is explained mainly by the high experimental/residual error, which accounts for 34.55% of the variation (SS_{Error} / SS_{Total} = 41340.9 / 119647 = 0.3455).

The r^2 is calculated using the formula: $r^2 = SS_{Model} / SS_{Total} = (SS_{Preparations} + SS_{Regression}) / SS_{Total}$

The sums of squares (SS) of Preparations and Regression are given in the Normal ANOVA table. Their sum (SS_{Model}) can be displayed by selecting the Complete ANOVA from the Options wizard. Table 13 shows the Complete ANOVA, with focus on the Model, Deviation from model and Residual Error sums of squares. It appears that r^2 represents the percentage of variation of the results that is explained by

the fitted model after subtracting lack-of-fit effects (deviation from linearity and parallelism for parallel-line models, deviation from linearity and intersection for slope-ratio models).

Note. In Ex. A.1.1, deviations from model are limited to non-parallelism, as non-linearity cannot be evaluated with 2 doses for each preparation only.

Table 13. Sums of squares of the Complete ANOVA table of Ex. A.1.1.

	DF	Sum of squares (SS)	Contributions
Model	**3**	70087.5	**58.58% (r^2)**
├ Preparations	2	63830.8	53.35%
└ Regression	1	6256.63	5.23%
Deviation from model	**2**	8218.23	**6.87%**
└ Non-parallelism	2	8218.23	6.87%
Residual Error	**54**	41340.9	**34.55%**
Total	59	119647	100%

For sigmoid models that involve a linearization of dose-response relationship, CombiStats calculates a weighted coefficient of correlation. This coefficient can be obtained by taking the square root of the coefficient of determination calculated as above. In Ex. A.1.14, the coefficient of determination is equal to SS_{Model} / SS_{Total} = 9.43129 / 9.47265 = 0.995634 and the weighted coefficient of correlation is equal to 0.997815. The simple or unweighted coefficient of correlation, introduced previously, is also reported, but for information only. In addition, it cannot be calculated from the sums of squares of the ANOVA table anymore.

If the model involves some weighting of the response variable, CombiStats performs a weighted regression and thus reports the weighted coefficient of correlation only. Weights can be defined using the Weight function of the Advanced options (F12). CombiStats performs a weighted regression when the weight function is different from w = 1 (no weight factor). It is the case for quantal responses, for which w = n / (m × (1 – m)), by default (Ex. A.1.8). Weighted regression applied to quantitative data are found in Ex. A.3.17, A.3.24 and A.3.29, in which the weight function ($w = 1/m^2$) is used to stabilize the variance when it increases with the dose. In Ex. A.3.30, Huber's weights (w = h) are used to reduce the influence of outliers without having to exclude them.

Table 14 provides a summary of the coefficients of correlation reported on CombiStats sheets depending on the regression model and weight function used.

Table 14. Coefficients of correlation reported on CombiStats sheets.

Model	No weights (w = 1)	Weights (w ≠ 1)
Slope-ratio	Unweighted r	Weighted r
Parallel-line	Unweighted r	Weighted r
Sigmoid, quantal data	Not applicable	Weighted r
Sigmoid, quantitative data	Weighted r and unweighted r	Weighted r

4. ANALYSIS AND INTERPRETATION OF RESULTS

4.1. Global analysis versus individual analyses

It is common to assay the standard and several test preparations in the same run, to make full use of resources, for example. The question is whether the analyst should perform one global statistical analysis or compare each test preparation to the standard in separated analyses.

The question is raised because the potency estimates and 95% confidence limits calculated for the test preparations will differ depending on the type of analysis. From the information provided in Table 15, the analyst can conclude that:

- Differences in common slopes and intercepts will affect the potency estimates, primarily.
- Differences in residual errors and associated degrees of freedom will affect the 95% CL.

Table 15. Key statistical parameters used to calculate potency estimates and 95% CL. (Parallel-line model, completely randomised design: 4 preparations × 5 doses × 3 replicates).

Parameters	Global analysis	Separated analyses
Output	One set of parameters calculated on 60 data	Three sets of parameters calculated on 30 data (standard: 15, test preparation: 15)
Common slope	Average of 4 individual slopes (standard and 3 test preparations)	Average of 2 individual slopes (standard and selected test preparation)
Intercepts	One vector of 4 intercepts $\{a_{Std}, a_{T1}, a_{T2}, a_{T3}\}$	Three vectors of 2 intercepts $\{a'_{Std}, a'_{T1}\}$; $\{a''_{Std}, a''_{T2}\}$; $\{a'''_{Std}, a'''_{T3}\}$
Residual error	Average of 4 experimental variances (standard and 3 test preparations)	Average of 2 experimental variances (standard and selected test preparation)
Degrees of freedom	4 prep. × 5 doses × (3 rep. – 1) = 40	2 prep. × 5 doses × (3 rep. – 1) = 20

The global analysis would be relevant from a statistical viewpoint, as it reflects the experimental design (all the preparations tested in the same run) with some benefits:

- **Control of the type I error rate.** A statistical test can result in erroneous conclusions. For example, the effect of a factor (e.g. non-linearity or non-parallelism) can be declared as significant whereas it is not. This outcome is defined as the Type I error, which nominal rate is 5% in the global analysis, whatever the number of preparations tested. By performing multiple analyses on the same set of data, the error rate will increase, and so the risk of taking wrong conclusions about the validity of the regression analyses.

- **Robust estimates of slopes and intercepts.** The calculation of the slope and intercept of a given preparation takes into account the experimental results obtained for the other preparations. For similar preparations tested in parallel, the global analysis will provide more robust estimates of the slopes and intercepts than the separated analyses.

- **Robust estimate of the assay repeatability.** Likewise, with the residual error calculated as the average of the experimental variances of the preparations, the global analysis will result in a more robust estimate of the assay repeatability (reflected by the increased number of degrees of freedom).

Overall, better precision (tighter 95%CL) can be expected from the global analysis of the results of a test run as well as more consistent potency estimates from one test run to another. However, the global analysis is meaningful only if all the preparations are similar (i.e. act as dilutions of a same product). This assumption – which is the basis of the regression models presented in Ph. Eur. Chapter 5.3 – is made, for example, when batches of a product are to be tested in routine analyses.

Despite the elements above in favour of the global analysis, the analyst may decide to perform separated analyses (one analysis per test preparation), depending on further considerations, e.g.

- There may be a formal request from a regulatory body to perform separated analyses.
- Separated analyses can result in a more flexible quality control process (e.g. invalid results of one preparation will not affect the processing of other preparations).
- The analyst may be testing unknown products with no guarantee of similarity among them. Separated analyses will allow evaluating the preparations, case by case.

4.2. Table of potency estimates

The potency estimates and 95% lower and upper limits (95%CL) of the test preparations are reported below the ANOVA table. In the case of separated analyses, there is one table of result in each CombiStats file. In case of a global analysis, there are as many tables of results as tested preparations.

As shown in Figure 13, the results are reported using the experimental units of the assay, e.g. 0.214 (95%CL: 0.185, 0.250) ng/unit and 936.6 (95%CL: 874.6, 1003) IU/mg. The 95%CL are also reported as percent of the potency estimate (Rel. to Est.):

- Rel. to Est. = (0.185/0.214, 0.250/0.214) = (86.3%, 116.7%) in Ex. A.2.6,
- Rel. to Est. = (874.6/936.6, 1003/936.6) = (93.4%, 107.1%) in Ex. A.3.2.

The 95%CL represent the precision of the potency result for which validity criteria can be found in monographs when applicable (e.g. 95%-105%, 80%-125%).

Sample 1			
(ng/unit)	Lower limit	Estimate	Upper limit
Potency	0.184890	0.214212	0.249937
Rel. to Ass.	?	?	?
Rel. to Est.	86.3%	100.0%	116.7%

Sample 1			
(IU/mg)	Lower limit	Estimate	Upper limit
Potency	874.648	936.639	1003.23
Rel. to Ass.	87.5%	93.7%	100.3%
Rel. to Est.	93.4%	100.0%	107.1%

Figure 13. Table of potency estimates (Ex. A.2.6 and A.3.2).

In addition, potency results can be reported as percent of the assumed potency of the test preparation specified on top of the table of raw data, e.g. 1000 IU/mg in Ex. A.3.2:

- Rel. to Ass. = (874.6/1000, 936.6/1000, 1003/1000) = (87.5%, 93.7%, 100.3%).

In absence of assumed potency of the test preparation, question marks are reported.

The assumed potency indicated by the analyst on top of the data table for the test preparation should be use with caution as it can influence the calculated potency results, depending on how doses are specified. Four different cases are shown in Table 16, based on Ex. A.1.2:

- Case 1: the assumed potency is 1 unit/mg and doses are reported as final contents (units) (explicit notation). The potency is 1.11181 unit/mg (Rel. to Ass. = 1.11181 / 1 = 111.2%).
- Case 2: the assumed potency is 2 unit/mg and doses are reported as final contents (units) (explicit notation). The potency is now 2.22361 unit/mg (Rel. to Ass. = 2.22361 / 2 = 111.2%).

- Case 3: the assumed potency is 2 unit/mg and doses are reported as dilutions (implicit notation). The potency is 1.11181 unit/mg (Rel. to Ass. = 1.11181 / 2 = 55.6%).
- Case 4: the assumed potency is 1 unit/mg and doses are reported as dilutions (implicit notation). The potency is 1.11181 unit/mg (Rel. to Ass. = 1.11181 / 1 = 111.2%).

In conclusion, when doses are reported as final contents (explicit notation, cases 1 and 2), the assumed potency of the test preparation takes the role of an assigned (true) potency used in the calculation of the potency result. When doses are reported as dilutions (implicit notation, cases 3 and 4), the assumed potency has no effect on the calculated potency result, only on the 'relative to assumed' (Rel. to Ass.). Therefore, the analyst should enter doses as final contents only when he is sure about the expected potency value of the test preparation. He should enter doses as dilutions otherwise.

Table 16. Effect of assumed potency and dose entry on calculated potency results.

Case	Data table	Potency results
1	Sample 1 Ass. pot. | 1 unit/mg Doses | 0.25 unit | 1.0 unit	Sample 1 (unit/mg) | Lower limit | Estimate | Upper limit Potency | 0.824973 | 1.11181 | 1.51357 Rel. to Ass. | 82.5% | 111.2% | 151.4% Rel. to Est. | 74.2% | 100.0% | 136.1%
2	Sample 1 Ass. pot. | 2 unit/mg Doses | 0.25 unit | 1.0 unit	Sample 1 (unit/mg) | Lower limit | Estimate | Upper limit Potency | 1.64995 | 2.22361 | 3.02714 Rel. to Ass. | 82.5% | 111.2% | 151.4% Rel. to Est. | 74.2% | 100.0% | 136.1%
3	Sample 1 Ass. pot. | 2 unit/mg Doses | 1 mg/4 mL | 1 mg/1 mL	Sample 1 (unit/mg) | Lower limit | Estimate | Upper limit Potency | 0.824973 | 1.11181 | 1.51357 Rel. to Ass. | 41.2% | 55.6% | 75.7% Rel. to Est. | 74.2% | 100.0% | 136.1%
4	Sample 1 Ass. pot. | 1 unit/mg Doses | 1 mg/4 mL | 1 mg/ 1 mL	Sample 1 (unit/mg) | Lower limit | Estimate | Upper limit Potency | 0.824973 | 1.11181 | 1.51357 Rel. to Ass. | 82.5% | 111.2% | 151.4% Rel. to Est. | 74.2% | 100.0% | 136.1%

There is the possibility to calculate effective doses in addition to potency estimates. This option is available from the Advanced options dialog box (F12 shortcut). The analyst can enter m = 50 and check 'Perc.' to calculate the classical ED50% value (Figure 14). The interpretation given to this value depends on the type of response variable used in the 4-parameter logistic model:

- Quantitative response: ED50 is the mid-point (parameter C) of the sigmoid curve (the other parameters, i.e. common slope, lower and upper asymptotes, are displayed on the right of the CombiStats file, just below the data tables).
- Quantal responses: ED50 is the dose producing an effect in 50% of subjects (e.g. 6 seropositive animals out of 12). Other effective doses can be calculated (e.g. EDm=10, EDm=90).

Figure 14 shows how the ED50 value and 95%CL are displayed according to selected options in the Advanced options dialog box. In this example, the ED50 value is ¼ of the assumed potency, i.e. 0.25 IU (Option 1: ED50 = 0.25 IU). Therefore, a dose of 1 IU contains 4 times the effective dose (Option 2: 1 dose = 4 ED50) (Rel. to Ass. = 400%).

Data table	Option 1: how many IU for the ED50?	Option 2: how many ED50 in a dose?

Sample 1	
Ass. pot.	1 IU/dose
Doses	(1)
1/1	1/10
1/2	3/10
1/4	5/10
1/8	7/10
1/16	9/10

Effective Dose: m = 50 Inv. ☑ Perc. ☑

Sample 1			
(IU/dose)	Lower limit	Estimate	Upper limit
IU/ED50	0.143570	0.250000	0.435328
Rel. to Ass.	229.7%	400.0%	696.5%
Rel. to Est.	57.4%	100.0%	174.1%

Effective Dose: m = 50 Inv. ☐ Perc. ☑

Sample 1			
(IU/dose)	Lower limit	Estimate	Upper limit
ED50/dose	2.29712	4.00000	6.96526
Rel. to Ass.	229.7%	400.0%	696.5%
Rel. to Est.	57.4%	100.0%	174.1%

Figure 14. Table of ED50 value and 95%CL.

The analyst may decide to untick the 'Perc.' checkbox. In that case, the effective dose is the dose required to reach an expected value (m) of the response variable. This option is of limited interest for quantal responses but is useful for quantitative ones as it offers the possibility to perform inverse predictions (interpolation of concentrations).

Example A.1.6

The analyst wants to calculate the dose corresponding to an absorbance value of 0.2. This dose is in-between S1 (0.33 RP) and S2 (0.66 RP) for the standard preparation and in-between T2 (0.66 RP) and T3 (1 RP) for the test preparation, according to the tables of raw data. After indicating m = 0.2 for the effective dose and unticking the 'Perc.' checkbox, the analyst obtains 0.596 and 0.724 RP for the standard and test preparations, respectively.

Note that the data transformation influences the values of ED50. As a result, the analyst should be very cautious when comparing or plotting (e.g. on a control chart) the ED50 values of routine assays analysed using different data transformations, as they may be expressed on different scales.

4.3. Plot of residuals

The term "residual" represents the vertical distance from one result to the regression line. There are as many residual values as experimental results, as shown in Figure 15, and the squares of these values are used to calculate the experimental/residual error (pure error in the ANOVA table) at the basis of the F-tests calculation. The lower the residuals, the lower the experimental error, the higher the ability of F-tests to detect signals (e.g. regression, deviation from linearity or parallelism). Low residuals also result in higher values of the coefficient of determination.

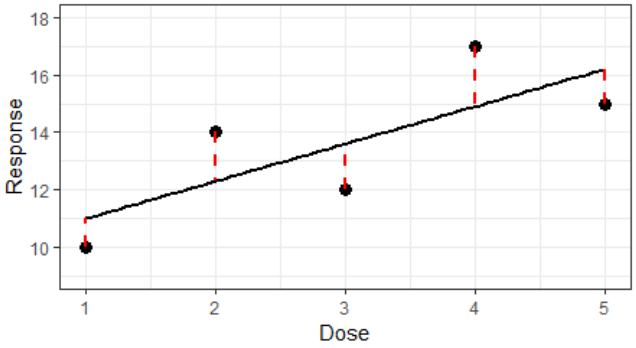


Figure 15. Regression plot with residuals represented by vertical dotted lines.

The main objective of the regression analysis is to summarise the experimental results, replacing them by mean values constituting the regression line. This substitution is relevant if it does not cause too much loss of information. This loss depends on the magnitude of the residual values and is calculated

as one minus the value of coefficient of determination (see section 3.4). For example, to a coefficient of determination of 97% corresponds a loss of information of 3%. From a graphical viewpoint, a coefficient of 100% would imply all the results (data points) to be on the regression line (experimentally unlikely).

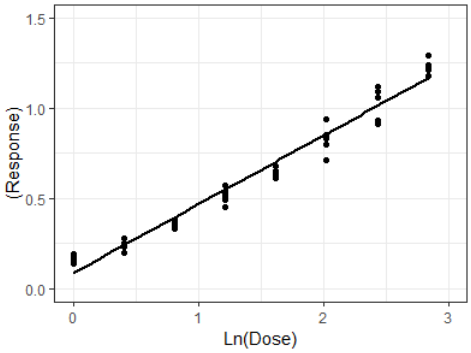
However, a good coefficient of determination is not enough for the model to be fully valid. In particular, the analyst should check that the regression line follows the trend of the data points or that their spread is managed correctly at each dose-level (see weight functions in case of heterogeneous variances in section 2.8).

Example A.2.1

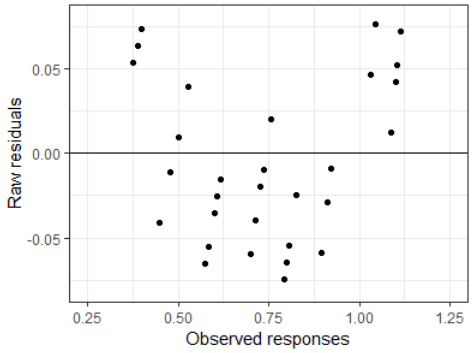
The dose-response relationship was fitted using a simple linear regression, which explains $r^2 = 96.8\%$ of the variation of the data. The residuals represent the remaining and unexplained 3.2%.

$$\text{Resp.} = a + b \times \ln(\text{Dose}) + \text{error}.$$

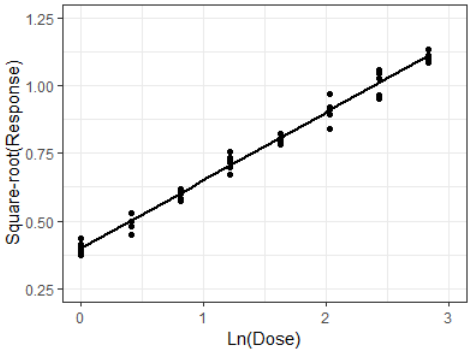
The data points show some curvature (regression plot in Figure 16-a) confirmed by the highly significant non-linearity contrast (ANOVA table, F-ratio = 7.288, $p \leq 0.001$). As the model equation doesn't include a quadratic term (i.e. $c \times \ln(\text{Dose})^2$), the curvature is likely to account for most of the 3.2% unexplained variation contained into the residuals. The curvature observed on the residual plot in Figure 16-b confirms this assumption.



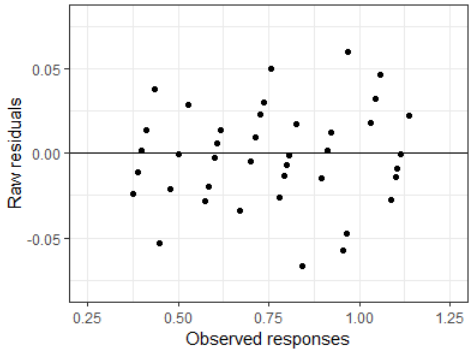
(a) Ex. A.2.1. Regression plot



(b) Ex. A.2.1. Residual plot



(c) Ex. A.2.2. Regression plot



(d) Ex. A.2.2. Residual plot

Figure 16. Regression plots and residual plots of Ex. A.2.1. and Ex. A.2.2.

Example A.2.2

The response variable was transformed ($\sqrt{y}$) to improve the linear fit (section 2.7). The simple linear regression explains $r^2 = 98.6\%$ of the variation of the data. The residuals represent the remaining and unexplained 1.4%. The regression and residual plots in Figure 16-c-d show that the model fit has

improved significantly following the data transformation. In particular, the residual points are randomly distributed (expected distribution for a valid model). Therefore, the 1.4% unexplained variation can be attributed to random experimental variations (repeatability). The good fit is confirmed by the non-significant non-linearity contrast in the ANOVA table (F-ratio = 0.621, p = 0.712).

In conclusion

The residual plot is a key element of the validity assessment of the fitted model. Any trend in the data not properly addressed by the analyst will be visible on the residual plot, helping to identify possible root-causes (e.g. curvature, heterogeneous variances across the dose range) and to take relevant actions (e.g. data transformation, weight function).

There is the possibility to represent the residuals (y-axis) according to different variables on the x-axis, like the observed responses or predicted responses. The residuals can be plotted also as raw or standardised values (recommended). The analyst can look at the different plots to check the goodness-of-fit of the model.

Notes.

The plots of residuals can be displayed using Tools > Graphs > Graph > Residual plot from the Menu bar. In addition, the residuals can be exported using Options > Advanced > Export Matrices > Dataset. They appear on the right part of the table as “RawRes” and “StandRes” for raw and standardised residuals, respectively. In CombiStats, standardised residuals are raw residuals adjusted for leverages (relative distances of data points on the x-axis (dose range)) and weighting function.

4.4. Outliers

Outliers are results which distances from the regression line are unexpectedly high. Since such distances are called residuals, an effective way to detect outliers is to look at the residual plot (section 4.3).

How to detect outliers

Residual plots usually come along with lower and upper limits (centred on 0), similar to the control limits of a control chart. Residual points located beyond these limits are considered as potential outliers.

In the absence of such calculated limits on the residual plots created by CombiStats, the analyst should consider the spread of the residual points on each side of the central line in order to define an empirical symmetrical interval representing the range of random variation of the assay. This interval can be further supported by approximated limits equal to k = 2 or 3 times the experimental standard deviation (“k-sigma limits”).

$$\{\text{Lower, Upper}\} \text{ limits} = \{- k \times \text{SQRT}(\text{residual error}), + k \times \text{SQRT}(\text{residual error})\}$$

Examples

The distribution of residuals in Figure 16-d looks rather homogeneous, ranging from about -0.06 to 0.06. Obviously, there are no outliers and an interval ranging, for example, from -0.05 to 0.05 would be too strict. A more practical range could be -0.075 to 0.075, supported by 2-sigma and 3-sigma limits equal to ± 0.058 and ± 0.087, respectively (residual error = 0.00085 in Ex. A.2.2).

Two more examples of residual plots are shown in Figure 17. The distribution of residuals shows no particular trend on the first plot. A practical range of random variation could be (-0.7, 0.7) (residual error = 0.064, 2-sigma limits ± 0.51 and 3-sigma limits ± 0.76).

Four clusters are visible on the second plot. They are explained by the selected dose range and do not represent an issue. A practical range of random variation could be (-5, 5), denoting the presence of two outliers: dose T4, rep. 5 (y = 188, resid. = 30.2) and dose S3, rep. 2 (y = 187, resid. = 10.2).

Note that the k-sigma limits would be too large in this last example, as the residual error is affected by the presence of the two outliers (residual error = 70, 2-sigma limits ± 17). This example reemphases the need to control chart residual errors (section 0) with the opportunity to replace the observed residual error by the average value of the control chart.

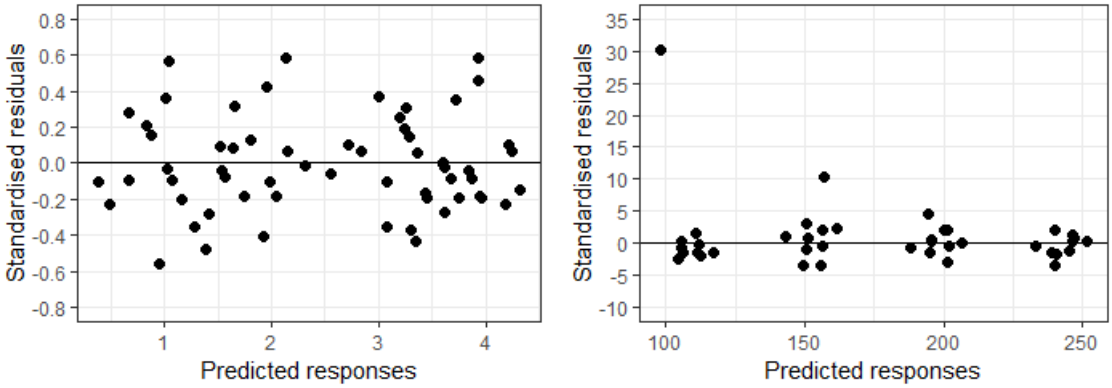


Figure 17. Residual plots of Ex. A.2.14. and Ex. A.3.30.

Declaration of outliers

The analyst should not exclude an outlier from statistical considerations only. He should investigate and identify likely causes of exclusion. Examples of classical root-causes are:

- Experimental work: preparation/dilution errors, stability issues, contamination, edge effect on microplates.
- Data analysis: calculation/reporting errors (possible correction), inappropriate statistical model/options (possible improvement).

The analyst may decide to perform a robust regression analysis in case he fails to identify the root-cause. By using Huber's weights (section 2.8, w = h), the influence of the outlier will be reduced without having to exclude its value (Ex. A.3.30).

4.5. Non-linearity contrast

Linearity of dose-response relationships is a key assumption of the regression analysis performed by CombiStats (section 2.1). When departure from linearity is observed (non-linearity contrast p-value ≤ 0.05), it is recommended to look at the regression plot (of linearised values) to figure out the possible root-cause (e.g. presence of outlier(s), curvature). As shown in Figure 18, the significance of the non-linearity contrast depends on the location of treatment means (red crosses) and associated distances (vertical dotted lines) on both sides of the regression line. In this example, the clear curvature observed on the left panel is solved by applying a log transformation to the response variable.

Significance of the non-linearity contrast is assessed against the experimental error in the ANOVA table, i.e. F-ratio = $MS_{Non\text{-}Lin.}$ / MS_{Error}. Table 17 shows the calculated results for the two above graphical representations. For example, in absence of transformation of the response variable, the high F-ratio (13.79) indicates a strong non-linearity signal, with subsequent low and significant p-value (0.001).

The analyst may find opposite conclusions when looking at the regression plot, which doesn't indicate any particular linearity issue, and the statistical test, which p-value is below the 0.05 significant

threshold. This paradoxical situation can be frequent when the assay is “very repeatable”. Indeed, very close replicated measurements result in a very low residual error, i.e. very low denominator of the F-ratio. Thus, the statistical test can detect very small departures from linearity, which are not practically meaningful.

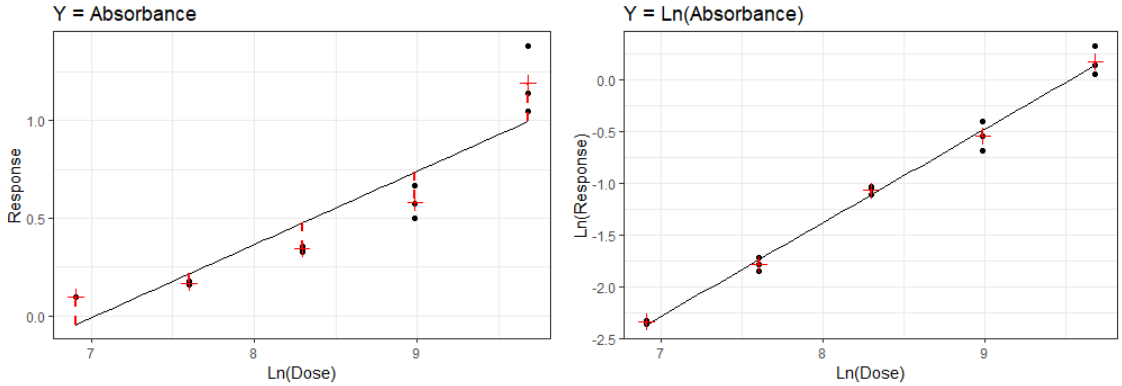


Figure 18. Non-Linearity contrast (Ex. A.1.5, Sample T). Significance depends on location of treatment means (red crosses) and associated distances (vertical dotted lines) on both sides of the regression line.

Table 17. Non-Linearity contrast significance (Ex. A.1.5, Sample T, ANOVA table).

Source	DF	Y = Absorbance			Y = Ln(Absorbance)		
		MS	F-ratio	p-value	MS	F-ratio	p-value
Non-Linearity	3	0.102506	13.79	0.001	0.001776	1.018	0.425
Residual error	10	0.007433		(***)	0.001746		

Dealing with non-linearity

In case of unusual departure from linearity, the analyst can look for errors during the experimental phase, the reporting of results or statistical analysis (e.g. inappropriate data transformation or model options). Excluding results during the statistical analysis may be acceptable in the plateaus of the sigmoid curve provided some rational exists (e.g. some reading systems can give erratic signals at saturation). Unexpected signal in the linear part of the sigmoid curve may require further investigation and the analyst may take appropriate actions, in accordance with their quality assurance procedures.

If departures from linearity occur frequently and are confirmed by the regression plot, the analyst can try a data transformation. If the data transformation doesn’t solve the lack of linearity, the range of selected doses may be reconsidered. The statistical analysis of a series of routine assays may be needed to better assess the expected dose-response relationship and find a suitable range of doses. In addition, a minimum of 3 doses should be kept in order to assess non-linearity (for parallel-line and slope-ratio analyses).

If the regression plot doesn’t confirm the significant non-linearity contrast (ANOVA table), the analyst can check whether replicates are independent and obtained with all relevant uncertainty contributors taken into account. Otherwise, significant non-linearity contrasts may be attributable to an underestimated experimental error. For example, replicates of signal measurements may not be enough and additional uncertainty contributors be worth considering (e.g. independent preparations, pre-dilutions).

If non-linearity contrasts remain frequently significant, the analyst may decide to introduce one decision rule in addition to the classical F-test (F-ratio). Different options may be suitable, including:

- Option 1. Replacing the exceptionally low residual error by a value that best represents the experimental error (standard uncertainty). This value can be estimated during the validation exercise and re-evaluated periodically using routine data (e.g. control chart of residual errors). It can be entered in the Variance tab of the Options wizard under Theoretical variance (left panel in Figure 19).

- Option 2. Checking that the mean square of the quadratic curvature is negligible compared to the mean square of the linear regression. Specifically, the ratio of mean squares should be less than 1/100 and the difference between preparations should be small [Bliss 1956, Hewitt 1981]. The quadratic curvature can be displayed by selecting the Extended or Complete ANOVA table from the Option wizard (right panel in Figure 19).

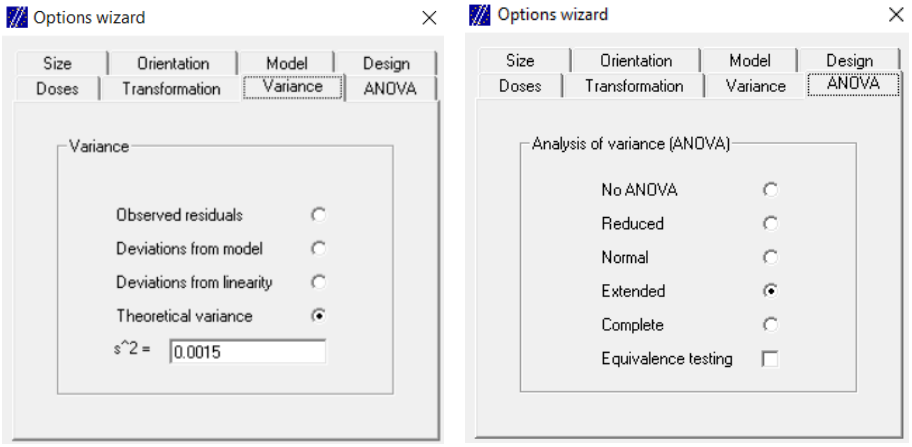


Figure 19. Additional rules to the classical F-test. Option 1. Using a theoretical variance that best reflects the experimental error (Left panel). Option 2. Checking whether the quadratic term is negligible compared to the linear term (Right panel).

The complete ANOVA table provides the information required by Option 2 (Figure 20). CombiStats actually displays the regression parameters of a polynomial model, e.g. for one preparation:

$$\text{Response} = a_0 + a_1 \times \text{Ln(Dose)} + a_2 \times \text{Ln(Dose)}^2 + \text{error, where:}$$

- a_0 is the regression intercept,
- a_1 is the slope, which significance is reported under “Regression” in the ANOVA table,
- a_2 is the quadratic term, which significance is reported under “Quadratic curvature”.

The requested mean squares are 0.280443 and 2.03737 for the quadratic term and slope, respectively, of the regression plot on the left panel of Figure 18. The ratio of mean squares (0.138) is far above the 1/100 threshold recommended in Option 2, confirming the significance of the curvature.

Source of variation	Degrees of freedom	Sum of squares	Mean square	F-ratio	Probability
Regression	1	2.03737	2.03737	274.096	0.000 (***)
Non-linearity	3	0.307519	0.102506	13.791	0.001 (***)
Quadratic curvature	1	0.280443	0.280443	37.729	0.000 (***)
Lack of quadratic fit	2	0.0270760	0.0135380	1.821	0.212

Figure 20. Extract from the complete ANOVA table used to compare quadratic term to linear term.

The degrees of freedom (DF) of the Non-linearity contrast is equal to the number of dilution points minus 2 (i.e. DF = 5 – 2 = 3). This contrast is further split into the Quadratic curvature (DF = 1), which

let 2 extra DF to assess the lack of fit of the polynomial model (Lack of quadratic fit). With a p-value of 0.212, there is no evidence of such a lack of fit, i.e. there is no need to add extra terms to the polynomial model.

However, a polynomial model that would require more than two terms (e.g. $a_0 + a_1 \times \text{Ln(Dose)} + a_2 \times \text{Ln(Dose)}^2 + a_3 \times \text{Ln(Dose)}^3 + \text{error}$) is practically unlikely and would rather denote some errors during the experimental phase, evaluation or reporting of results.

Note. The tables and figures of this section were created using the data of Sample T of Ex. A.1.5.

4.6. Non-parallelism contrast

Introduction

Parallelism of regression lines is applicable to parallel line and sigmoid models (for quantitative and quantal responses), i.e. models characterised by a log-dose transformation (section 3.3). Parallelism results from similarity between the test and standard preparations, which is a key condition to potency calculations.

When departure from parallelism is observed (non-parallelism contrast p-value ≤ 0.05), it is recommended to look at the regression plot (of linearised values) to figure out the possible root-cause. For example, there is very good parallelism between the regression lines of Preparations S and T in Figure 21. Preparation U shows an atypical trend affected by the lower results observed at the first dose. The analyst could look for possible experimental errors (e.g. pre-dilution error), unless the trend is expected.

Significance of the non-parallelism contrast is assessed against the experimental error in the ANOVA table, i.e. F-ratio = $MS_{\text{Non-Par.}}$ / MS_{Error}. Table 18 shows the calculated results for Ex. A.1.1 (Preparations S, T and U) and Ex. A.1.2 (Preparations S, T). Taking the 3 preparations in account, the high F-ratio (5.367) indicates a strong non-parallelism signal, with subsequent low and significant p-value (0.007).

The analyst may find opposite conclusions when looking at the regression plot, which doesn't indicate any particular non-parallelism issue, and the statistical test, which p-value is below the 0.05 significant threshold. This can be frequent when the assay is "very repeatable". Indeed, very close replicated measurements result in a very low residual error, i.e. very low denominator of the F-ratio. Thus, the statistical test can detect very small departures from parallelism, which are not practically meaningful.

Example A.3.17 provides an illustration of this paradoxical situation. While the regression plot doesn't indicate any particular non-parallelism issue, the contrast in the ANOVA table is slightly significant (F-ratio = 3.90, p-value = 0.048), as a consequence of a low residual error.

Notes.

The residual error is the experimental variance, i.e. the variance between replicated results (Var. = 0.001865 absorbance² in Ex. A.3.17). The standard deviation may be easier to interpret as expressed in absorbance like the replicated results (SD = sqrt(Var.) = 0.0432 absorbance). Taking the absorbance value (about 0.5) corresponding to the ED50, the relative standard deviation is RSD = 0.0432 / 0.5 = 8.6%.

The RSD value is not as low as thought initially, which may explain why the non-parallelism contrast is 'just' significant (i.e. p-value slightly lower than the 0.05 significant threshold). Therefore, there is no clear evidence that this paradoxical situation would happen frequently. On contrary, it is likely to be the case when the RSD is about 2.5% or lower.

Table 18. Non-Parallelism contrast significance (ANOVA table).

	Preparations S, T, U (Ex. A.1.1)				Preparations S, T (Ex. A.1.2)			
Source	DF	MS	F-ratio	p-value	DF	MS	F-ratio	p-value
Non-Parallelism	2	4109.12	5.367	0.007	1	34.2250	0.046	0.831
Residual error	54	765.572		(**)	39	738.536		

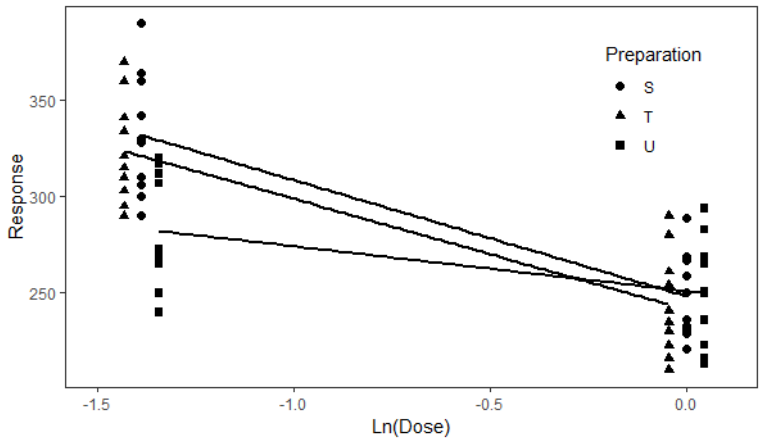


Figure 21. Example of a regression plot (parallel line assay in Ex. A.1.1).

Dealing with non-parallelism

In case of unusual departure from parallelism, the analyst can look for errors during the experimental phase, the reporting of results or statistical analysis (e.g. inappropriate data transformation or model options). Excluding results during the statistical analysis may be acceptable in the plateaus of the sigmoid curve provided some rational exists (e.g. some reading systems can give erratic signals at saturation). Unexpected signal in the linear part of the sigmoid curve may require further investigation and the analyst may take appropriate actions, in accordance with their quality assurance procedures.

If departures from parallelism occur frequently and are confirmed by the regression plot, the analyst can try a data transformation. If the data transformation doesn't solve the lack of parallelism, the range of selected doses may be reconsidered. Actually, the analyst should address these points during the development of the assay such that an appropriate dose range and data transformation are used consistently in routine testing. Therefore, a lack of parallelism in routine assays may be indicative of an uncontrolled change in the procedure or manufacturing process or storage condition of materials.

If the regression plot doesn't confirm the significant non-parallelism contrast (ANOVA table), the analyst can check whether replicates are independent and obtained with all relevant uncertainty contributors taken into account. Otherwise, significant non-parallelism contrasts may be attributable to an underestimated experimental error. For example, replicates of signal measurements may not be enough and additional uncertainty contributors be worth considering (e.g. independent preparations, pre-dilutions).

If non-parallelism contrasts remain frequently significant, the analyst may decide to introduce one decision rule in addition to the classical F-test (F-ratio). Different options may be suitable, including the replacement of the exceptionally low residual error by a value that best represents the experimental error. This value can be estimated during method development and first routine assays

and re-evaluated periodically, e.g. using e.g. control chart of residual errors. It can be entered in the Variance tab of the Options wizard under Theoretical variance (left panel in Figure 19).

Another approach consists in using an equivalence test in place of the classical F-test. To do so, the analyst can check the "Equivalence testing" box in the ANOVA tab of the Option wizard (right panel in Figure 19). CombiStats will calculate differences of slopes (test preparations vs. standard) together with 90% confidence intervals. It will also calculate ratios of slopes (and 90%CI), as both calculations can be found in the scientific literature. However, the analyst should not use them in parallel and should choose between differences or ratios for routine analyses.

The equivalence between the slopes of the test and standard preparations is demonstrated if the slope difference (or ratio) and 90%CI are within some pre-defined equivalence limits (-θ, + θ) (Figure 22). These limits should mimic the shape of the 90%CI and thus be symmetrical or asymmetrical (defined as a fold-ratio) for differences and ratios of slopes, respectively. In addition, the range of the equivalence limits is critical to the ability of the test to detect non-parallelism of practical significance. These limits can be determined using relevant information obtained during method development and first routine assays and then checked periodically.

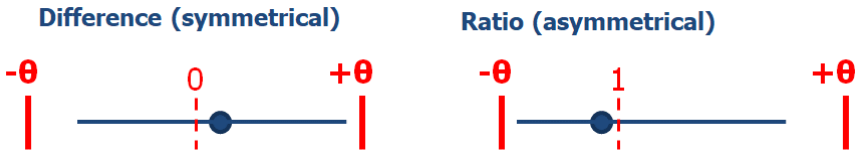


Figure 22. Equivalence of slopes (non-parallelism evaluation). The dot and horizontal line represent the slope difference (or ratio) and 90% confidence interval. The vertical dotted line represents the expected value is case of perfect parallelism (equality of slopes). Outer vertical lines represent the equivalence limits (-θ, + θ).

Figure 23 shows the results of the equivalence test performed to compare the slopes of preparations T and U (Samples 1 and 2) to the slope of preparation S (Standard). For example, the slope ratio between T and S is 0.956 (90%CI: 0.659-1.377). That is, the slope of preparation T is equal to 95.6% (90%CI: 65.9%-137.7%) of the slope of preparation S. The slope ratio between U and S is much more pronounced (38.5%, 90%CI: 13.5%-68.5%) and would likely fail the equivalence limits.

Recall that the equivalence limits are assay-dependent and that they should be defined in a way to detect departures from parallelism of practical relevance. Last, the classical F-test and equivalence test cannot be used in parallel for the evaluation of non-parallelism in routine tests of a given assay.

	Slope per Sample	Difference with Standard	Ratio with Standard
Standard	-60.3047 (-75.2427 to -45.3666)	0	1
Sample 1	-57.6357 (-72.5738 to -42.6976)	2.66899 (-18.4567 to 23.7946)	0.955742 (0.659070 to 1.37737)
Sample 2	-23.2274 (-38.1655 to -8.28930)	37.0773 (15.9516 to 58.2029)	0.385167 (0.135203 to 0.685490)

Figure 23. Individual slopes, differences and ratios (Ex. A.1.1).
(confidence intervals are reported in brackets).

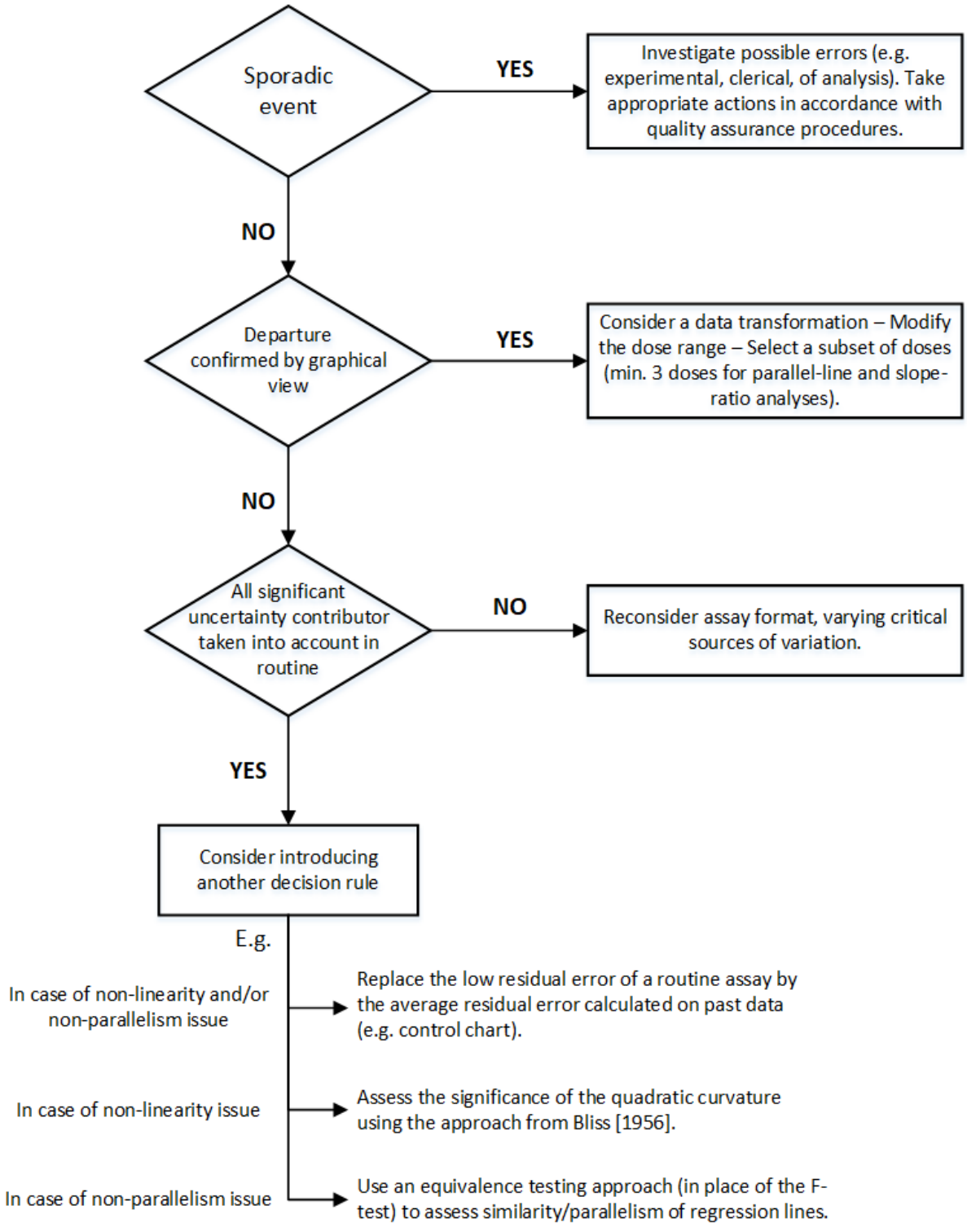


Figure 24. Non-linearity and non-parallelism assessment flowchart.

4.7. Use of control charts

Control charts can be used to monitor the validity and performance of the assay. The analyst should start by identifying relevant parameters, which would be indicative of unexpected events (e.g. experimental error) or potential trends (e.g. material degradation). In that view, the results of an internal control, the slope of the standard preparation and the experimental (residual) error are frequently monitored parameters.

Different control charts may be suitable, including:

- The Shewhart charts (e.g. IMR-chart), which 3-sigma limits can be used to detect unexpected events,
- Multirules (e.g. 6 increasing or decreasing results in a row) or advanced control charts (e.g. EWMA chart), which can be used to detect the onset of a trend or mean shift.

Example 1. Table 19 shows the experimental variance (mean square error) and associated degrees of freedom (DF) of 30 independent assays. Each assay consists in two preparations tested in duplicates at 6 doses. There is one variance between duplicates for each of the 12 treatments (2 prep. × 6 doses), calculated with 1 DF (number of replicates – 1). The experimental variance is the average of the individual variances, calculated with 12 DF.

In practice, the analyst can exclude some doses from the parallel-line analysis. With 2 replicates per dose, the number of DF reported in the table informs about the number of doses kept for analysis (e.g. 9 for the first assay). There are 8 to 9 doses kept in 19 out of 30 analyses (63%) and 7 to 10 doses kept in 28 out of 30 analyses (93%).

For convenience, the experimental variances was multiplied by 10,000 in Table 19. In addition, as the analyst routinely performs a log transformation of the response (absorbance) – to improve linearity and parallelism of regression lines – the experimental variances are reported in $[\ln(\text{Absorbance})]^2$. A practical way to represent the assay variability is to calculate the geometric coefficients of variation or GCV (below formula), which values range from 0.7% (Assay A24) to 5.7% (assay A14).

$$GCV = 100 \times SQRT(EXP(\text{mean squared error}) - 1),$$

Where residual error is the value found in the ANOVA table in the column Mean Square.

Table 19. Mean squared errors (MSE), degrees of freedom (DF) and geometric coefficients of variation (GCV) of 30 parallel-line assays.

Assay	MSE	DF	GCV	Assay	MSE	DF	GCV	Assay	MSE	DF	GCV
A01	3.34	9	1.8%	A11	7.65	9	2.8%	A21	2.57	10	1.6%
A02	18.38	10	4.3%	A12	3.10	9	1.8%	A22	5.00	8	2.2%
A03	3.97	9	2.0%	A13	9.91	10	3.1%	A23	2.80	9	1.7%
A04	4.69	9	2.2%	A14	32.48	12	5.7%	A24	0.55	7	0.7%
A05	2.46	8	1.6%	A15	4.11	8	2.0%	A25	2.42	8	1.6%
A06	1.12	7	1.1%	A16	3.09	9	1.8%	A26	1.89	7	1.4%
A07	3.24	8	1.8%	A17	2.00	8	1.4%	A27	7.36	10	2.7%
A08	7.67	9	2.8%	A18	0.59	7	0.8%	A28	2.28	6	1.5%
A09	3.92	9	2.0%	A19	6.78	10	2.6%	A29	2.73	8	1.7%
A10	2.03	8	1.4%	A20	0.90	8	0.9%	A30	3.75	8	1.9%

GCVs are plotted against numbers of doses kept for analysis in Figure 25, where it appears obvious that both variables are correlated. The higher the number of doses, the higher the experimental

variation. This is somewhat expected as results in the tails of the dose range are often more dispersed than in the middle-range. With this observation, it is not possible to create a classical control chart, which is applicable to simple random samples only. One possibility could be to model the relationship between the experimental variation and the number of doses kept for analysis in order to estimate relevant control limits.

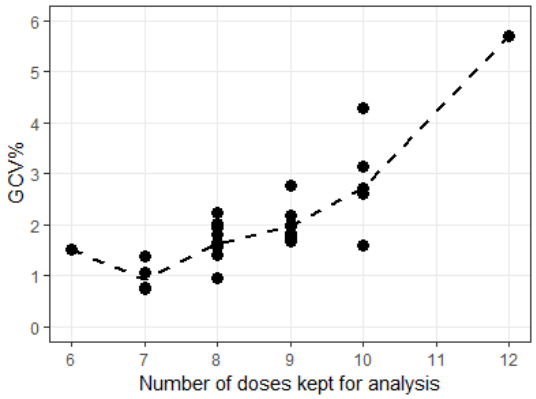


Figure 25. GCVs vs. numbers of doses kept for analysis.

Example 2. The previous example showed that data of a control chart should be obtained using the same statistical design, model and analysis options. In the current example, residual errors are pure errors from completely randomised designs (section 1.2). Each assay consists in 2 preparations tested in triplicates at 3 doses analysed using a parallel-line model with no data transformation nor weighting of the response variable (units: Y).

Experimental variances (MSE in Y^2) are reported in Table 20 together with coefficients of variation (CV%), which values range from 3.2% (Assay A06) to 10.0% (assay A14).

$$CV = 100 \times SQRT(MSE) / Mean,$$

Where Mean is the mean absorbance value calculated at the middle dose.

Note. In Example 1, a geometric coefficient of variation was calculated because of the log-transformation of the response variable.

Table 20. Mean squared errors (MSE) and coefficient of variation (CV).

Assay	MSE	CV	Assay	MSE	CV	Assay	MSE	CV
A01	1.223	5.9%	A11	0.494	3.7%	A21	0.830	4.9%
A02	1.033	5.3%	A12	1.528	6.7%	A22	1.131	5.4%
A03	0.951	4.9%	A13	0.452	3.5%	A23	2.254	8.3%
A04	1.438	6.3%	A14	3.735	10.0%	A24	0.939	4.8%
A05	1.937	7.5%	A15	0.752	4.6%	A25	1.145	5.6%
A06	0.366	3.2%	A16	0.830	4.9%	A26	1.001	5.1%
A07	1.699	6.6%	A17	1.853	7.5%	A27	1.937	7.3%
A08	1.550	6.4%	A18	2.685	8.3%	A28	1.083	5.2%
A09	0.939	4.9%	A19	1.255	5.9%	A29	1.043	5.5%
A10	0.893	5.0%	A20	1.475	6.5%	A30	0.427	3.5%

The skewed distribution shown on the left panel of Figure 26 is typical of MSE values. The analyst should log-transform these values to improve the symmetry of the distribution (right panel of Figure 26) and calculate relevant control limits.

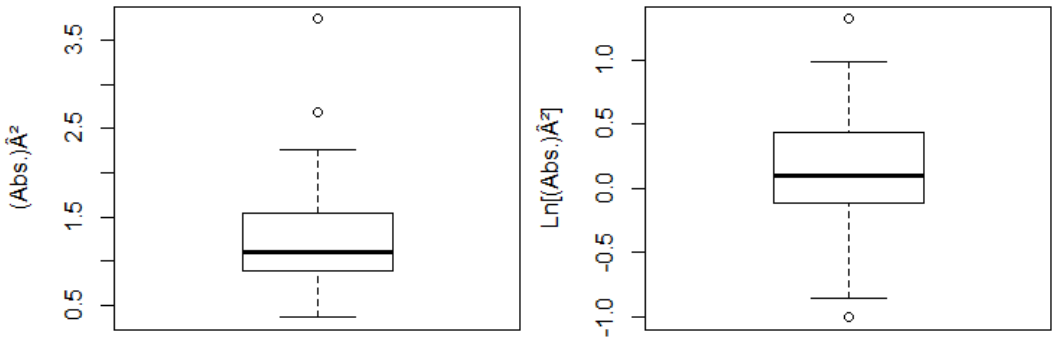


Figure 26. Boxplots of MSE values
No transformation (left panel), log-transformed (right panel).

The control limits can be calculated using the I-MR Shewhart chart, which results are summarised in Table 21. The control limits reported in the table are "2-sigma limits" (95% confidence level) calculated after log-transformation of MSE values, although another confidence level may be used.

Table 21. I-MR Shewhart control chart.

Scale	Units	Mean	Standard deviation (SD)	Lower control limit	Upper control limit
Log-transf.	$Ln(Y^2)$	0.125	0.428	-0.731	0.981
Back-transf.	Y^2	1.13	n.a.	0.481	2.67

Control limits = Mean ± 2 SD. Back-transformation: $e^{-0.731}$ = 0.481 Y^2 and $e^{0.981}$ = 2.67 Y^2. For an overall mean response of 19.1 at the middle dose, the above mean and control limits correspond to the following CVs: 5.6%, 3.6% and 8.6%.

In routine, the analyst may decide to use the control chart of log-transformed values ($ln(Y^2)$) or back-transformed values (Y^2) as shown in Figure 27. Both representations show the classical alert and action limits ("2-sigma" and "3-sigma" limits, respectively) and have exactly the same performance.

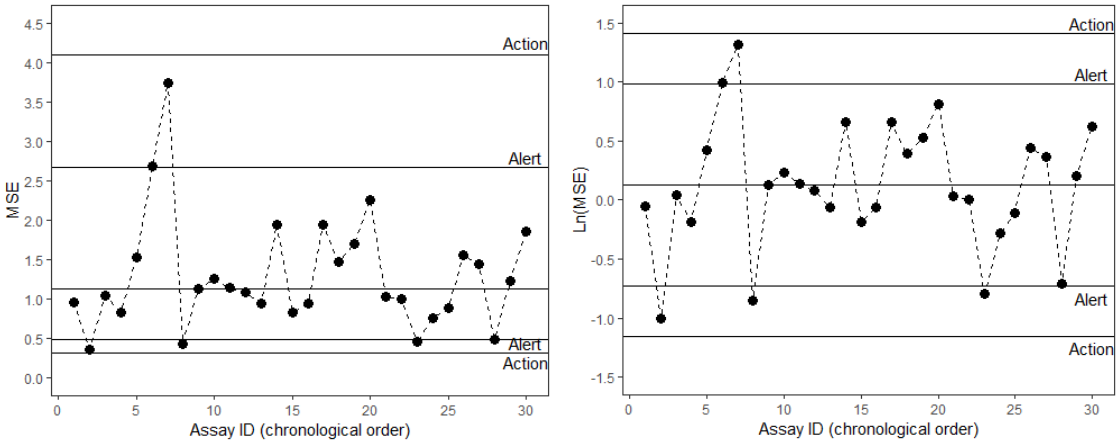


Figure 27. I-chart of MSE values
No transformation (left panel), log-transformed (right panel).

Note. The analyst can find details about how to create control charts in many statistical publications, including, for example, the ISO 7870 guidelines and in Douglas Montgomery's Introduction to Statistical Quality Control.