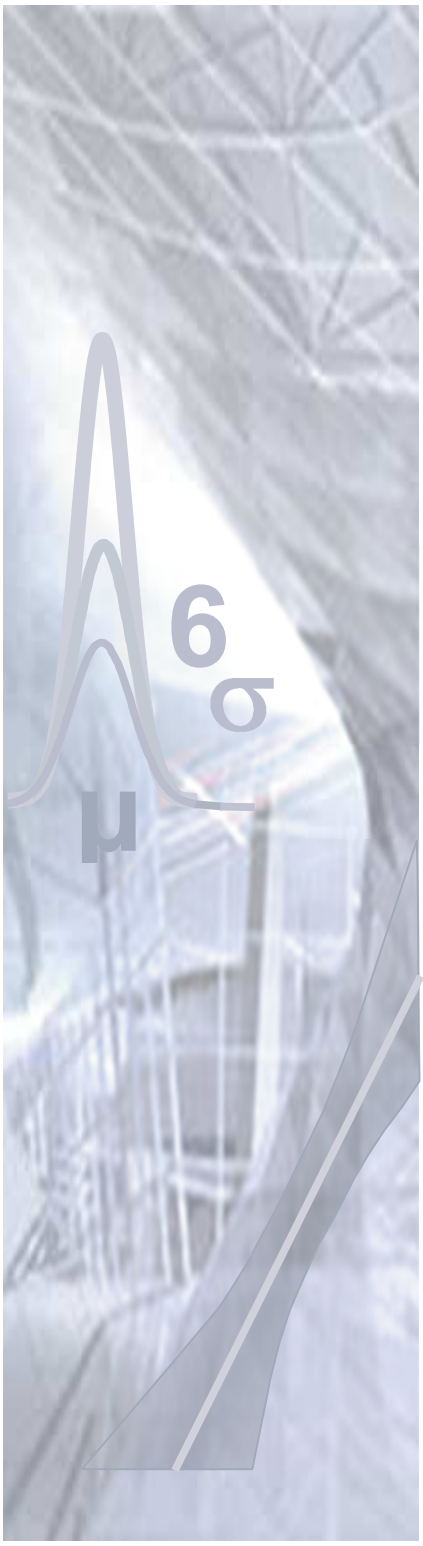
Handbook
of statistical methods

The most important
Methods and procedures for
practice

Curt Ronniger

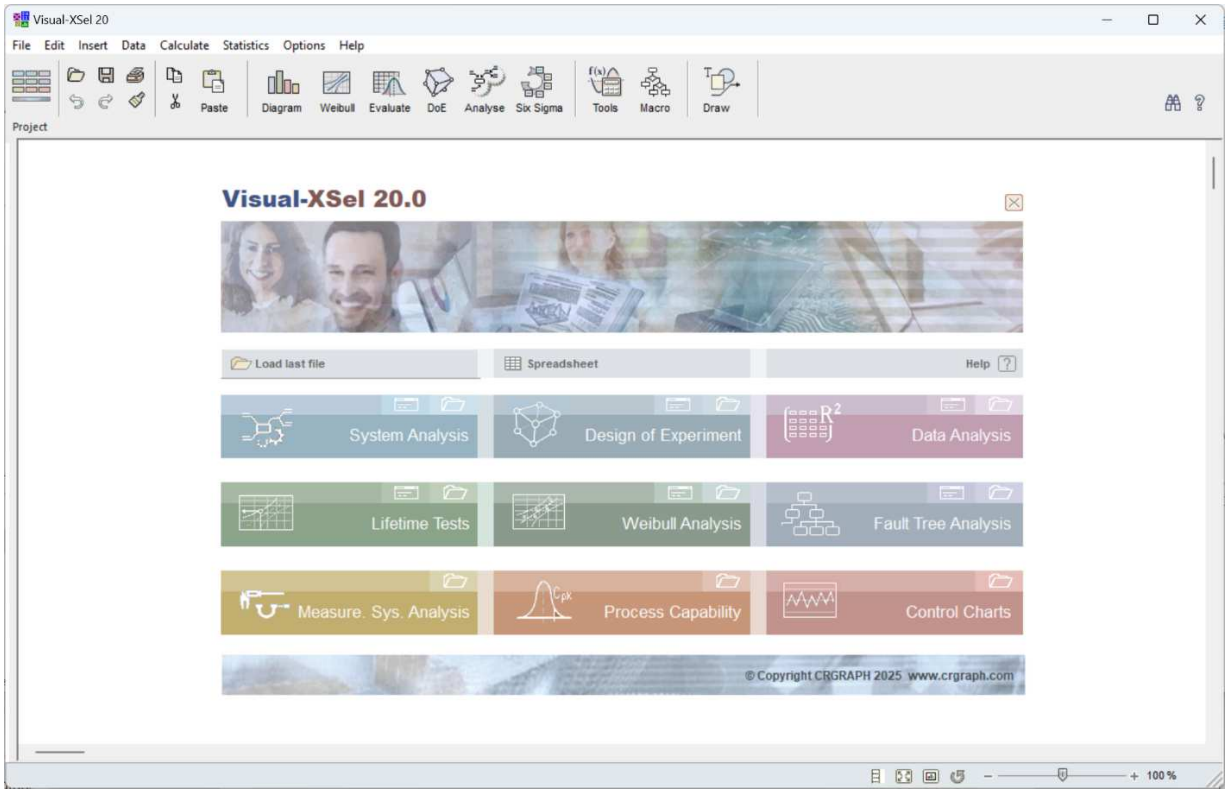
www.crgraph.com



Content

Software

For the methods and procedures, which are shown here, the software Visual-XSel is used.



For the first steps use the icons on the start picture and follow the menus and hints. There are also templates with examples.

Visual-XSel goes much further than other standard programs with many important topics, such as reliability methods & Weibull, as well as Design of Experiments (DoE) and data evaluation.
Some method like hypothesis tests are profided through templates. Those files marked in italics in the overviews and descriptions in blue represent these presentations. The procedure is always the same: Put your data into the table (marked often with yellow background) and start the program with F9. The results are shown then in the main window

For more information, please goto www.crgraph.com

Please ask for a test version via info@crgraph.de

1. Test methods

Under test methods there are statistical methods to understand which were developed for example through Taguchi /3/. These are also known under the system optimization.

The goal is to find the most important influences in technical or other processes, with a minimum of parts and tests.
The products and their productional processes can be improved decisively with these mostly very simple methods.

In the following descriptions there are no derivations of the formulas. The priority is much more the application for the practice. On further-reaching information the literature is therefore referred.

The following issues are treated:

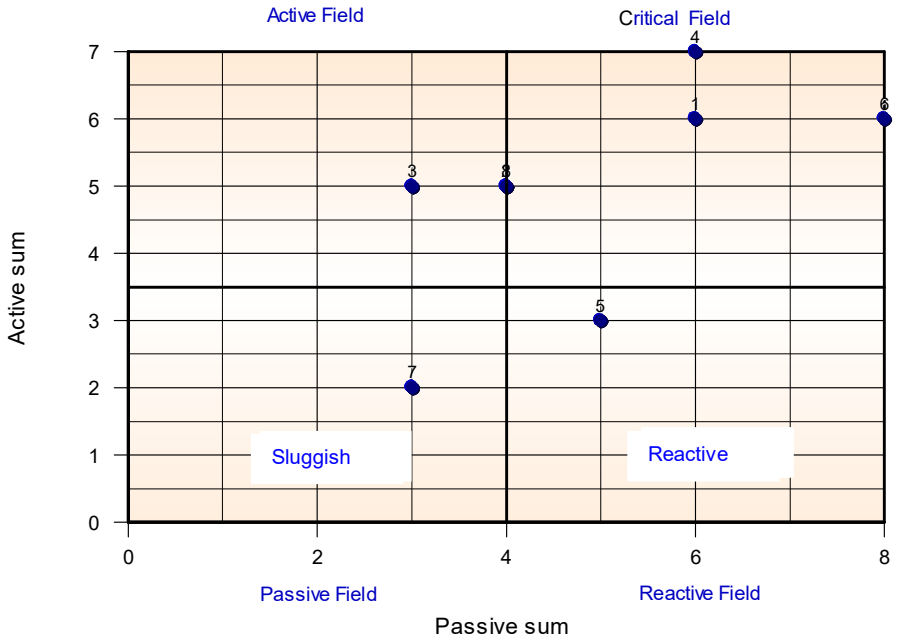
- Taguchi strategy and experiments
- Standard experiments and D-optimal
- ANOVA (Variance – analysis)
- Statistical diagrams and special charts
- Correlation and regression
- Multiple Regression (stepwise regression)
- Multivariate analyses
- Statistical tests and evaluations
- Statistical distributions
- Optimization

Intensity-Relation-Matrix

In a so-called Intensity-Relation Matrix it is the point that the decisive factors for a later investigation should be found, or to reduce the quantity of parameter to the essential ones for an experimental design.
At first the entry of factors with their designations takes place vertically in a table. The same factors have to be entered horizontally in the first row. The particular effects of the factors have to be enlisted in the first column on the factors with the same sequence in the first row.

		Outside-diameters	Roundness	Plan-sprint	Angle pin	Altitude oil-supply	Scope-situation oil-supply	Height-tolerance	Pin-diameter
Outside-diameters	1		3	0	0	1	0	2	0
Roundness	2	3		0	1	1	0	0	0
Plan-sprint	3	0	0		0	2	2	1	0
Angle pin	4	0	0	2		0	3	0	2
Altitude oil-supply	5	0	1	0	0		0	0	2
Scope-situation oil-supply	6	3	0	0	3	0		0	0
Height-tolerance	7	0	0	1	0	1	0		0
Pin-diameter	8	0	0	0	2	0	3	0	

Active
Critical
Passiv-sum
Active-sum



Normally the values for this are estimated by experts or specialists. Possibly the numerical values can be weighted. In a diagram the active summations are spread over the passive summations after a valuation and parts the diagram shares in four big areas. Those depict the active and passive, as well as the critical and reactive field.

For further experimental designs the factors in the active field as well as in the critical field have to be taken into account. Generally, here it is a matter of possible reciprocations. It is possible to renounce the factors in the passive field. The factors in the reactive field can also be performed in the treatment as sub-target factors, which will not be varied in further experimental designs.

This method can be executed directly via the menu *statistics/Intensity-Relation-Matrix* inside the spreadsheet.

Priority Matrix

Different criterions or characteristics are compared in the Priority matrix together and a ranking was formed. The result can be used also for importance's of the criterions for continuing evaluations

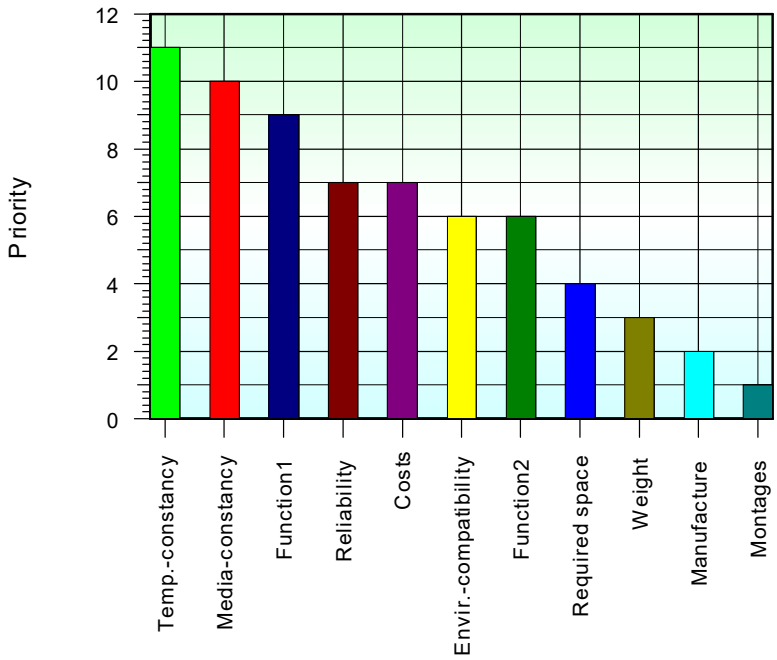
No quantitative measurements are necessary for the comparisons of the characteristics. The test is just a pair-wise comparison and an estimation of experts.

For example: The importance's should be determined for a later comparison of different technical solutions. The characteristics are function, reliability, weight etc. Each criterion has to be compared with each other. That one which is more important gets the number of the criterion (order number in the row). In the first comparison, the function 1 is more important than function 2. Therefore, in row 2 of column 1 is the number of functions 1. The next step is the comparison of function 1 with the reliability.

	Characteristic	1	2	3	4	5	6	7	8	9	10
1	Function1										
2	Function2	1									
3	Reliability	1	3								
4	Weight	1	2	3							
5	Required space	1	2	3	5						
6	Temperature-constancy	6	6	6	6	6					
7	Media-constancy	7	7	7	7	7	6				
8	Environment-compatibility	1	2	3	8	8	6	7			
9	Montages	1	2	3	4	5	6	7	8		
10	Manufacture	1	2	3	4	5	6	7	8	10	
11	Costs	1	11	11	11	11	6	7	8	11	11

The column 2 refers on the evaluation of the function 2 opposite each other criterion. The reliability is more important than function 2. Therefore, in column 2 is the number of the reliability with the value 3.

Now you add up the occurring numbers for each criterion and get the ranking. In this case you can see the following Pareto-Chart.



Each result should be increased of 1, because it is not meaningful to get values with zero (if the results are importance's and you multiply this with other evaluations, you will get also zero), the other point is that in the Pareto-Chart a zero value is not visible.

This method can be executed with help of the template file *Priority_Matrix.vxg*

Matrix diagram

The so called “matrix diagram” is just a representation of a matrix, not really a diagram. However, through an assessment of the rows and columns between each other, there is built a structure.
In the following example the task is to show which method (the first column) is suitable for which use case (the first line).

	to find ideas	present informations	data collection	data evaluation	structuring	illustrate processes	process capability	process control	illustrate connections	system analysis	concept - design	failure cause
Brainstorming	3		1							2	3	1
Monitor, count, measure			3									2
Correlation-diagram		3		2					2			3
Regression/Model-charts		3		3					3			3
Cause-effect-diagram		3			3	2			3	3	3	3
Block-diagram		3			3	1			2	2	3	1
Flow-chart, process-chart		3				3			3	3	3	2
Intensity-relation-matrix		3			2				2	3	2	2
Matrix diagram		3	1		2				3	2	2	1
Priority-evaluation		3			2					2	3	2
Pareto-analysis		2		2								2
Paired comparison				2					1	2	3	2
Histogram		3	3	1				2				1
Quality chart			2	2			2	3				
Probability chart			2	2			3	2				
Weibull			2	2				1				2
		3	2					2				

It is possible to use other items, titles or meaning. The mutual relations are described here as numerical values between 1 and 3. No connection means empty fields or 0. Leaving out 0 has the big advantage that the representation becomes clearer (pattern). Note: The difference is to intensity relation matrix is the titles of the first row are identical with the first column (mutual comparison).
Another evaluation is possible by the row-by-row summation (who brings most points).

2. Analysis of Variance (ANOVA)

The Analysis Of Variance (ANOVA for short) is about determining the variance of groups (factors) against the unexplained variance (residual variance) and „confirming“ or rejecting a significant influence.
Historically, the ANOVA was the evaluation tool for Design of Experiment (DoE). Alternatively, regression methods can usually do more.

The ANOVA is used to determine whether the factors in relation to the scatter (dispersion) have a significant effect on the response.

The known methods of ANOVA are diverse. Only the most important procedures are described in this documentation.
In general, an analysis of variance (dispersion decomposition) is carried out in an ANOVA in order to differentiate systematic influences of factors (treatment) from a random dispersion. The general model is:

Total dispersion = Factors disp. + Error dispersion

$$SS_{Total} = SS_{Factors} + SS_{Error}$$

$$\sum_{j=1}^{z}\sum_{i=1}^{n}\left(y_{ji}-\bar{y}\right)^2 = n\sum_{j=1}^{z}\left(\bar{y}_j-\bar{y}\right)^2 + \sum_{j=1}^{z}\sum_{i=1}^{n}\left(y_{ji}-\bar{y}_j\right)^2$$

SS	Sum of Squares
y_{ji}	Data point from col j and row i
$\bar{y}_j$	Mean of col j
$\bar{y}$	Mean over all

Measurem.	Factors 1	2	3	...	z
1	y_{11}	y_{21}	y_{31}	...	y_{z1}
2	y_{12}	y_{22}	y_{32}	...	y_{z2}
3	y_{13}	y_{23}	y_{33}	...	y_{z3}
...	...	...	...	...	...
n	y_{1n}	y_{2n}	y_{3n}	...	y_{zn}

The variance = Mean Squares MS is the sum of squares based on the degrees of freedom. Here is z = number of factors and n = number of observations:

$$MS_{Total} = \left(\frac{SS_{Total}}{z\,n-1}\right) \quad MS_{Factors} = \left(\frac{SS_{Factors}}{z-1}\right) \quad MS_{Error} = \left(\frac{SS_{Error}}{z\,(n-1)}\right)$$

For a significant test now, the $MS_{Factors}$ are divided through MS_{Error} and it is:

$$F = \frac{MS_{Factors}}{MS_{Error}}$$

The bigger the F-value, the higher the probability of the factor effect. The null hypothesis Ho is: The means of the factors do not differ from one another. Ho is rejected if the probability

from the F-distribution with degrees of freedom $f1 = z\text{-}1$ and $f2 = z\,(n\text{-}1)$ is less than the significance level α.

The so-called coefficient of determination R^2 describes how much the effect of the factors is in the model. The maximum is R^2=1. The bigger the scatter the smaller is the R^2.

$$R^2 = 1 - \frac{SS_{Error}}{SS_{Total}}$$

Balanced One-Way ANOVA (μ1 = μ2 = μ3…)

The null hypothesis is for several data columns of the same size

μ1 = μ2 = μ2 = …

The prerequisite for this test is that the data series are normally distributed. The variances must be the same, which can be checked using the F-test. Alternatively, the t-test is possible, with which different variances are possible. The data series must be independent of one another. For the following example we want to test the null hypothesis that all mean values are equal.
In the following the *Sum of Squares SS* and Degrees of Freedom = DF are calculated with:

z →

	A	B	C	
	1,0	4,0	5,5	
	1,5	5,5	6,5	
	2,5	6,0	8,0	
	4,0	7,0	9,0	
n	5,0	9,0	9,5	
$\bar{y}$	2,8	6,3	7,7	$\bar{\bar{y}}$ = 5,6

$$SS_{Total} = \sum_{j=1}^{z}\sum_{i=1}^{n}\left(y_{j,i} - \bar{\bar{y}}\right)^2 = 100{,}1$$

$$SS_{Factors} = n \cdot (\bar{y} - \bar{\bar{y}})^2 = 5 \cdot 12{,}74 = 63{,}7$$

$$SS_{Error} = SS_{Total} - SS_{Factors} = 36{,}4$$

$$DF_{Total} = n \cdot z - 1 = 14$$

$$DF_{Factors} = z - 1 = 2$$

$$DF_{Error} = DF_{Total} - DF_{Factors} = 12$$

Table of results:

	DF	SS	MS	F	p-val
Factors	2	63,7	31,85	10,50	0,0023
Error	12	36,4	3,03		
Total	14	100,1			

The *p-value* is determined via the Fisher-distribution with:

$$p\text{-}value = 1 - Fisher(F; f1; f2) = 1 - Fisher(10{,}5; 2; 12) = 0{,}0023 \;;\; f1 = DF_{Faktors} \;;\; f2 = DF_{Error}$$

Since the p-value falls below the specified significance level of α = 0.05, the null hypothesis that the mean values are equal has to be rejected.

Balanced Two-Way ANOVA

In contrast to the one-way ANOVA, there is a response variable y in the two-way that affects the factors. The aim here is to determine a relationship between the factors and the response. The factors must have the same number of observations (balanced), must be independent of each other, have comparable scatter, and have to be normally distributed.

For this the variance analysis is:

$$SS_{abs} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \qquad SS_{Total} = \sum_{i=1}^{n} y_i^2 - SS_{abs}$$

$$SS_A = \frac{1}{bk}\sum_{i=1}^{a} \bar{y}_i^2 - SS_{abs}$$

$$SS_B = \frac{1}{ak}\sum_{j=1}^{b} \bar{y}_j^2 - SS_{abs}$$

$$SS_{AB} = \frac{1}{k}\sum_{i=1}^{a}\sum_{j=1}^{b} \bar{y}_{ji}^2 - SS_A - SS_B - SS_{abs}$$

$$SS_{Error} = SS_{tot} - SS_A - SS_B - SS_{AB}$$

n : number of observations

a : number of variations of factor A

b : number of variations of factor B

k : number of repetitions

$\bar{y}_i$: mean of the i-th factor level of factor A

$\bar{y}_j$: mean of the j-th factor level of factor B

The results for a two-factor example with the influences of an additive and the temperature on a process (variable of response) are generally output in the tabular form shown. The F-value is the ratio between the variance (mean square) of the factors and the interaction with the variance of the dispersion (error). From this the probability of error (p-value) is determined via the F-distribution:

	DF	SS	MS	F	p-val
Additive	3	2,608E+02	8,692E+01	4,99	0,008
Temperature	2	8,029E+02	4,014E+02	23,05	0,002
Additive * Temperature	6	3,340E+02	5,567E+01	3,20	0,019
Error	24	4,180E+02	1,742E+01		
Total	35	1,816E+03			

Balanced Two-Way ANOVA Random

Random factors have randomly selected levels, while the levels of fixed factors e.g., have been established by a DoE. The following example resulted in temperatures that were not systematically specified.

Instead of relating the variance MS of the additive to the variance of the error MS_{Error}, the variance of the interaction is used here:

	DF	SS	MS	F	p-val	Typ
Additive	3	2,61E+02	8,69E+01	1,56	0,294	fix
Temperature	2	8,03E+02	4,01E+02	10,5	0,017	random
Additive * Temperature	6	3,34E+02	5,57E+01	3,2	0,019	
Error	24	4,18E+02	1,74E+01			
Total	35	1,82E+03				

This procedure is used in the Measurement System Analysis with ANOVA according to VDA Volume 5. The parts used for the repeatability and the appraisers are random and not the same, e.g., like for a later determination of a process capability.

Nested Two-Way ANOVA

In a so-called nested ANOVA, there is a factor that cannot be freely combined. All factors in the model must be random factors. In this example, the temperature is generated by different heating processes in an oven. Each temperature level is therefore nested in the additives. Instead of referring the variance MS of the additive to MS_{Error}, the second nested factor temperature is used here.

	DF	SS	MS	F	p-val
Additive	2	8,03E+02	4,01E+02	6,075	0,022
Temperature	9	5,95E+02	6,61E+01	3,794	0,004
Error	24	4,18E+02	1,74E+01		
Total	35	1,82E+03			

The last factor is now based on MS_{Error}.

A nested ANOVA is used in particular in the measurement system analysis when the parts that are actually to be measured repeatedly must always be different due to destructive tests.

3. Design of Experiment

Design

After definition of factors the design or the type of the experimental design is fixed. As model ***Linear***, ***Interaction***, ***quadratic*** and ***Cubic*** are standard plans. The orthogonal experimental design according to Taguchi is just available for the linear model, because interactions are mixed with each other.

Type	Attitude	Remark
Linear Y = bo+b 1 x1 + b2 x2.	Factors on respectively only 2 steps, min number of tests p +1*	No nonlinearities and interactions determinable
Change effects Y = . b 4 x1 x2	Factors on respectively only 2 steps, min number of tests p+p (p 1)/2+1 *	No nonlinearities determinable, but interactions
Square Y = .b4 x1 ^ 2.	Factors on respectively only 3 steps min number of tests 2 p+p (p 1)/2+1 *	Nonlinearities recognizable. Incl . interactions
Cubic Y = . . b4 x1 ^ 2 +b5 x1 ^ 3 ..	Factors on respectively only 4 steps, min number of tests 3 p+p (p 1)/2+1 *	Curses of curve with turning point recognizable, incl . interactions

p = number of factors, min = number of tests related to D optimal

According to the choice the required terms are added in a list on the left. Terms can be deleted again, too, e.g., if it is known that certain interactions do not happen. The following design types can be chosen:

Full factorial	All combinations, full orthogonal	High number of tests, effortful best evaluable
Fractional	Half or a smaller number of tests like full factorial, full orthogonal	Mixing of interactions Unsafe of evaluation
Plackett Burmann	Derivation from fractional design. Very low number or tests.	Interactions are not fully confounded
DSD	Definitive Screening Designs Very low number or tests on 3 levels	Quadratic model possible, interactions are confounded only partially
Taguchi	Very low number of tests, multiple fractional full orthogonal	Many interactions mixed with each other and with factors; suitable only for regulation of individual factors

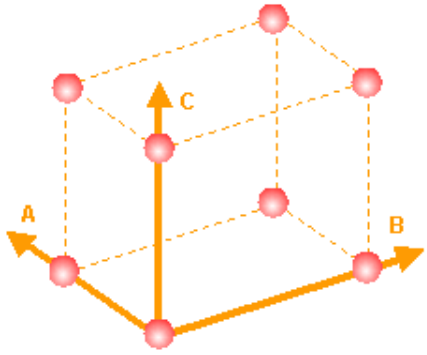
Central Composite Design	The same construction as full-factorial plus cross in the middle. Test space like a ball	High number of tests, effortful good evaluable
Box-Behnken	Evaluation for quadratic models. Middle levels in outlet area.	High number of tests, effortful good evaluable
D-Optimal	Low number of tests, Clear regulation of interactions,	not orthogonal good evaluable
Mixture	Use of factors whose sum must always amount to 100%	not orthogonal, factors dependent on each other good evaluable

Coexistent with the model and type selection the number of so-called candidates and the number of needed trials is shown beneath. The candidates always correspond to those of the full factorial experimental design. So, for a squared model with 3 factors 3^3=27 trials are needed. In addition, also a central point with the middle values and repeats can be chosen. For this see options.

Fullfactorial design

A full factorial test plan is made if all possible attitudes of the factors are combined with each other.
The number of tests required can be calculated through:

$$n = 2^p$$



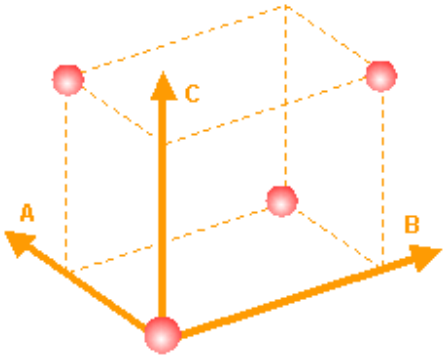
	A	B	C	D	E	F
1	-1	-1	-1	-1	-1	-1
2	1	-1	-1	-1	-1	-1
3	-1	1	-1	-1	-1	-1
4	1	1	-1	-1	-1	-1
5	-1	-1	1	-1	-1	-1
6	1	-1	1	-1	-1	-1
7	-1	1	1	-1	-1	-1
8	1	1	1	-1	-1	-1
9	-1	-1	-1	1	-1	-1
10	1	-1	-1	1	-1	-1
11	-1	1	-1	1	-1	-1
12	1	1	-1	1	-1	-1
13	-1	-1	1	1	-1	-1
14	1	-1	1	1	-1	-1
15	-1	1	1	1	-1	-1
16	1	1	1	1	-1	-1
17	-1	-1	-1	-1	1	-1
18	1	-1	-1	-1	1	-1

At 3 factors, 8 tests arise. Simply one generally prepares a full factorial plan (-1 and 1 standardize) in the following way:

It is the advantage of the complete test plan that all interactions can be explained. So, the influence of A*B*C is just as contained. The number of tests increases with the number of factors, however, very strongly fast, so that the test plan gets too effortful beginning at 5 factors. The question how one can simplify it arises.

This plan is full orthogonal

Fractional design



The triple interaction has only a small influence in most cases. Concerning this statement, one can put another factor instead of the combination which contains A*B*C and then one receives a fractional test plan. In this case the plan is the half size of the full-factorial plan with 2^{4-1}. It is the disadvantage of this test plan that no more triple interactions are determinable and two-factor interactions are confounded with each other: AB with CD, AC with BD and AD with BC, because the respective column products are identical. For a product with at least 4 columns, e.g., F=ABCD two-factor interactions aren't confounded any more. These plans have a so-called resolution of at least V. In general, the number of tests is calculated through

$$n = 2^{p-1}$$

One builds this factorial design at first like the full-factorial plan, but with q factors less. The attitudes of the missing factors q are generated by the product of all previous columns. One also calls these columns "generators". The following table shows an overview for 12 factors:

n \ p	2	3	4	5	6	7	8	9	10	11	12
4	2^2 fullfact.	2^{3-1} III									
8		2^3 fullfact.	2^{4-1} IV	2^{5-2} III	2^{6-3} III	2^{7-4} III					
16			2^4 fullfact.	2^{5-1} V	2^{6-2} IV	2^{7-3} IV	2^{8-4} IV	2^{9-5} III	2^{10-6} III	2^{11-7} III	2^{12-8} III
32				2^5 fullfact.	2^{6-1} VI	2^{7-2} IV	2^{8-3} IV	2^{9-4} IV	2^{10-5} IV	2^{11-6} IV	2^{12-7} IV
64					2^6 fullfact.	2^{7-1} VII	2^{8-2} V	2^{9-3} IV	2^{10-4} IV	2^{11-5} IV	2^{12-6} IV
128						2^7 fullfact.	2^{8-1} VIII	2^{9-2} VI	2^{10-3} V	2^{11-4} V	2^{12-5} IV

Full factorial -> all interactions are evaluable
Fractional plans -> all two-factor interactions evaluable ≥ V
Fractional plans -> two-factor interactions mixed, resolution < V

All fractional plans with resolution V or more are uncritically in the evaluation. Also, here the effort rises up excessive over a number of 6 factors. Therefore D-optimal test plans at which all interactions can always be found out then can be recommended. Plans with resolution less than V gets smaller size but can be used only for searching the most important factors, because interactions are confounded. One also calls this Screening.

Resolution III Design
Main effects are confounded (aliased) with two-factor interactions.

Resolution IV Design
No main effects are aliased with two-factor interactions, but two-factor interactions are aliased with each other.

Resolution V Design
No main effect or two-factor interaction is aliased with any other main effect or two-factor interaction, but two-factor interactions are aliased with three-factor interactions.

With using D-Optimal plans there is still the chance to determine all interactions by the same size of trials like for resolutions < V (see the following chapters).

Plackett-Burman-experiments

Especially Plackett-Burman-Experiments are suitable for preliminary investigations or so-called Screening-plans (only 2 levels). These test plans are derived from fractional plans and can be constructed in steps by 4 tests. With 12 tests there can be determined 11 effects (factors). Nevertheless, it is recommended not to use at least two columns with factors.
Plackett Burman-test plans have compared with the classical fractional plans (resolution III) the great advantage that interactions among each other and with other factors are not completely confounded. For plans with 12 tests and 11 factors a max. correlation of 0,333 arises for two-factor interactions. An evaluation via multiple regression is here normally not a problem. For plans with 20 tests and 19 factors a max. correlation of 0,6 exists. This can be critical to determine interactions. Under circumstances this correlation is too high for evaluations of interactions, in particular if high scatter is given. Indeed, an additional security is given by the evaluation with the method PLS which is non sensitive against correlations.

But there are in each case no confounding's between the factors.

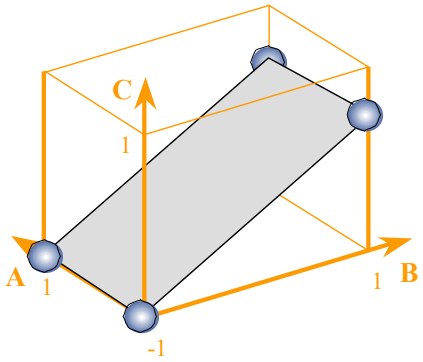
After evaluation with the stepwise regression ordinarily fall out a greater number of 2-factor interactions. Plackett Burman-test plans thereby advantageous when an evaluation should be done before of unknown interactions, but the test expenditure must be very small. Confirmation tests are to be recommended, in any case.
The creation of the plans occurs through the following pattern: A combination order which is repeated column for column around a line down moved is used in each case. The pattern is depending on n:

n=12	+	+	-	+	+	+	-	-	-	+	-													
n=20	+	+	-	-	+	+	+	+	-	+	-	+	-	-	-	-	+	+	-					
n=24	+	+	+	+	+	-	+	-	+	+	-	-	+	+	-	-	+	-	+	-	-	-	-	

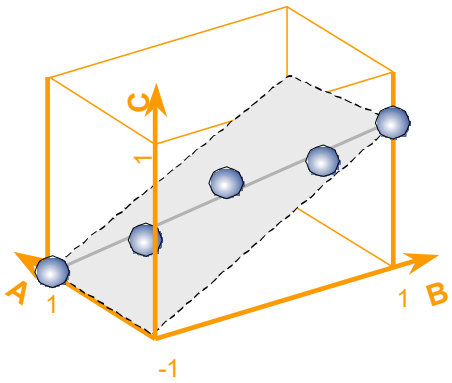
The last field is absent. After cyclic joining together of the columns the surpluses about the line n-1 are added on top again. The last missing line is taken with continuously -1.

Orthogonality

All full-factorial and fractional plans are orthogonal. If there are the factors independent from each other and the correlation coefficients are 0, the plan is full orthogonal. Every factor can have values without changing the attitudes of the other factors. This isn't the case in the right representation. B cannot be changed independently by A. If the plan is not quite orthogonal, e.g., due to a central point, then the evaluation is still possible with the calculation via matrices. At the same deviation of the Y values, the confidence intervals are, however, wider than at orthogonal plans.



orthogonal



not orthogonal

Taguchi

Taguchi plans are, fractional test plans which still more interactions are covered with factors. e.g.:

2^{7-4}

Through this one needs a very low number of tests. A mixture of factors with interaction also arises from it. Therefore, these plans only are recommended if interactions cannot be expected. This plan is full orthogonal.
The plans are marked by Lx in which x is the number of tests. These plans are appropriately orthogonal. 2 examples of orthogonal combinations to Taguchi represent the following plans:

Instead of the standardization -1 ... 1
the attitudes are numbered

$L_4\ (2^3)$

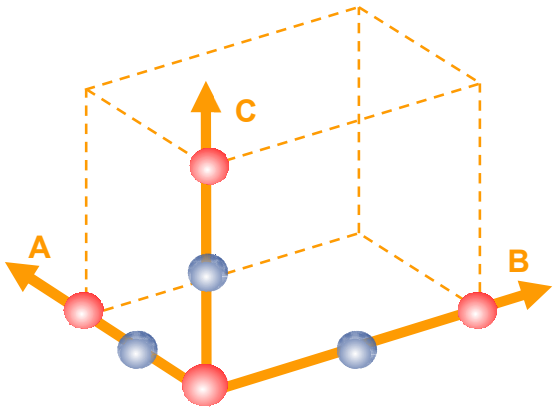
	A	B	C
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

$L_9\ (3^4)$

	A	B	C	D
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

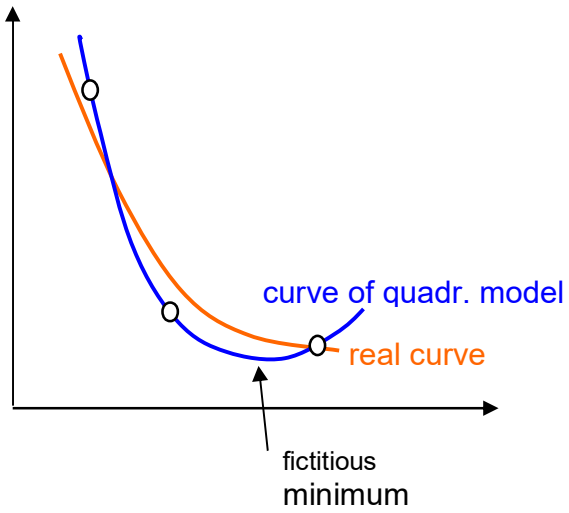
Full-factorial quadratic

In the previous test plans only, linear relations can be explained. In many cases, however, there are nonlinear relations. To take this into account, one additional information each is needed in the test plan. For standardized factor attitudes the levels will be therefore -1, 0, 1. The shown picture illustrate the attitudes without the combinations for detecting the interaction. The necessary number of tests are:



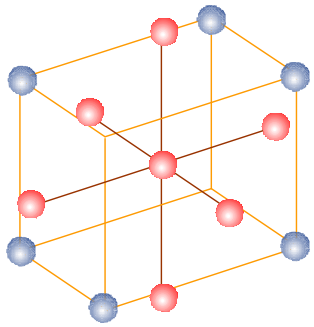
$$n = 3^p$$

The model formation by square terms is in some cases not satisfying. Square terms have the quality that they can produce a maximum or minimum in the used range which not exists in reality. The search for the optimal point then would lie in the bill minimum instead of in the edge area of the course falling in reality steadily. The corresponding data for this factor should be logarithmic. Through this a bent curve which doesn't show any maximum or minimum is produced. Since perhaps the curse of curve, however, isn't shown well enough, the square terms should remain nevertheless contained in the model and perhaps be removed only at the evaluation due to the significance (p-value). At the evaluation the logarithmic transformation must be taken into account in the coefficient ($Y = b_o + b_1 \cdot ln\,(x_1) + b_2 \cdot ln\,(x_1)^2 +$). Another problem can be that the won model equation allows negative values (Y) which cannot be reached in the reality. The logarithmic transformation helps also here.



Central Composite Design

A central composite design consists of a full-factorial terms and a centric star. The shown representation applies to the order of a plan with 3 factors.

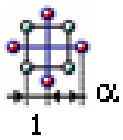


The purpose is the attainment of a roughly spherical test room in which the central point is repeated. As a rule, at a standardized orientation -1 .. +1 the star has an extension of

$\alpha = \pm\sqrt{2}$.

Those plans are also called Central Composite Circumscribed (CCC). Plans with α = 1 is also as Central Composite Face (CCF) plans described

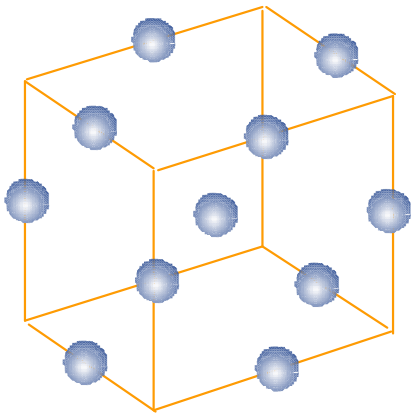
A	B	C	
-1	-1	-1	full factorial
-1	-1	1	
-1	1	-1	
-1	1	1	
1	-1	-1	
1	-1	1	
1	1	-1	
1	1	1	
-1,414	0	0	star
1,414	0	0	
0	-1,414	0	
0	1,414	0	
0	0	-1,414	
0	0	1,414	
0	0	0	centre
0	0	0	
0	0	0	



The evaluability of this type of experiment is very good, however, is even bigger than full factorial.

Box-Behnken design

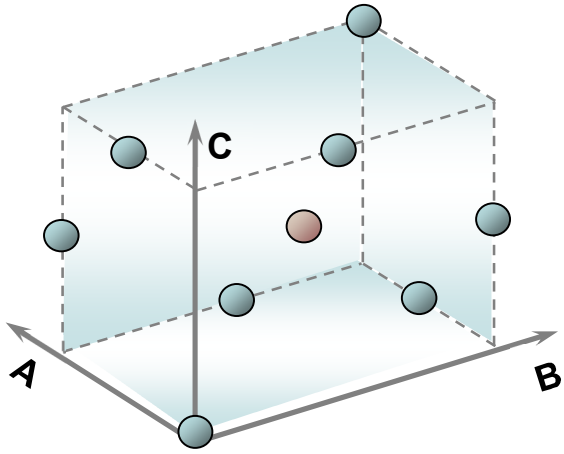
The essential characteristic of the Box-Behnken design is that the middle levels lie in the respective middle of the edge area. Additional there are a center point. With this a square model (non-linear) can be determined (3 levels). Box-Behnke test plans are not derived from fractional designs. The missing corners are ca be advantageous for tests where these extreme combinations are not adjustable.



Box-Behnke test plans can be turned approximately. Under 45° one identifies in the picture on top a CCD plan. In the left table a Box Behnke design (not rotated) is compared with the CCD- design. In the Box-Behnke design a little bit fewer tests are required. If one used in the CCD plan correct-wise 3 central points, the difference precipitates even greater.

CCD			Box-Behnken		
x1	x2	x3	x1	x2	x3
-1	-1	-1	-1	-1	0
1	-1	-1	1	-1	0
-1	1	-1	-1	1	0
1	1	-1	1	1	0
-1	-1	1	-1	0	-1
1	-1	1	1	0	-1
-1	1	1	-1	0	1
1	1	1	1	0	1
-1,4	0	0	0	-1	-1
1,4	0	0	0	1	-1
0	-1,4	0	0	-1	1
0	1,4	0	0	1	1
0	0	-1,4	0	0	0
0	0	1,4			
0	0	0			

Definitive Screening Designs DSD



So-called Definitive Screening Designs are very new test plans developed by Jones and Nachtsheim with a very small number of experiments. They enable the evaluation of quadratic models and are therefore based on 3 levels. There is no confounding (orthogonal) between the main factors among themselves and the quadratic terms. The interactions are not 100% confounded (no correlation). But there are not enough experiments to solve all possible interactions and the correlations r > 0.5.

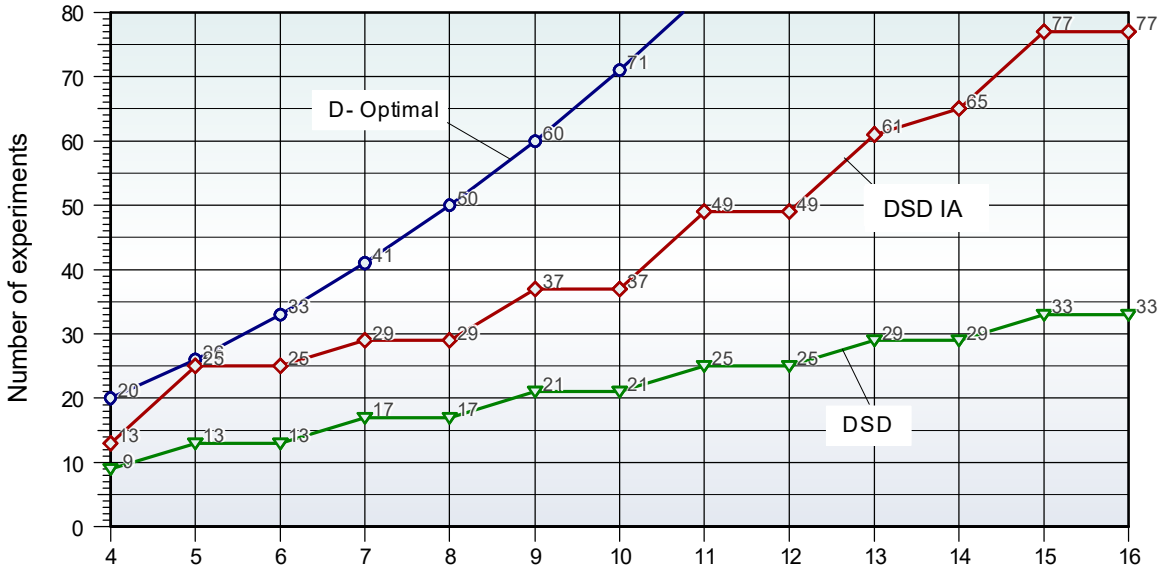
No	A	B	C	D
1	0	1	-1	-1
2	0	-1	1	1
3	-1	0	-1	1
4	1	0	1	-1
5	-1	-1	0	-1
6	1	1	0	1
7	-1	1	1	0
8	1	-1	-1	0
9	0	0	0	0

In the generic generation of these designs (iteratively using the determinant), the number of tests is

n = 2*p+1+(2)

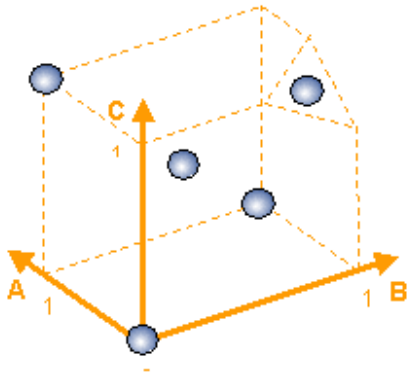
(if the number of factors is odd, two lines without 0 are additionally required).

If you want to be able to evaluate interactions in general, a further variant is to build up the DSD basic plan with more columns than required factors, but not to use the last column(s). These experimental designs mentioned here as DSD IA have significantly fewer experiments than a
D-optimal from 6 factors. However, these are usually sufficient for the significant interactions remaining in the evaluation. The following overview shows the number of attempts compared to D-Optimal (see next chapter)



D-Optimal experiments

Fundamentals



The aim of D-Optimal plans is with minimum effort to prepare test plans which show the desired effects and interactions definitely. This is, a decisive advantage over the fractional design where interactions are confounded with each other partly.

With p = number the number of simple interactions charges itself to factors:

p' = p*(p-1)/2

As a rule, the higher interactions (e.g., ABC, ABD, ACD etc.) are not taken into account since its influence is usually less opposite the simple ones. You also would blow up the size of the tests.

Altogether, the following number of tests is needed for a test plan with two attitudes:

Constant	: 1
Main effects (factors)	: p
Interactions	: p' = p*(p-1)/2
Sum	: p+ p*(p-1)/2+1

In the case of a square model there are still one time p tests (with a middle attitude). Furthermore, gets approximately 5 tests needs to receive sufficient information about the spreads (significances of the factors).

A D-Optimal plan is not generated with a firm scheme but built up iteratively. It has among others the following important qualities:

- Maximization of the determinant (indicator for evaluability)
- Minimization of the correlations and confidence intervals
- Balanced levels (as good as possible)

Due to the target that all interactions shall be recognized at a low-test number prevents particularly that these plans are orthogonal completely., i.e., certain correlations cannot be removed completely. This is, however, a subordinate disadvantage in the evaluation about Multiple Regression.

Advantages of the D-Optimal test plans

- Free choice for the number of the steps per influence factor. The number of levels can be elected factor by factor differently.
- Free choice of the step distances which can equidistantly or not be chosen equidistantly.

- Free choice for the distribution of the test points in the n dimensional test room
- Free choice of the mathematical model
- Expansion capability by new influence factors
- Certain attitudes and combinations can be excluded,
  these are not attainable

Disadvantages of the D-Optimal test plans

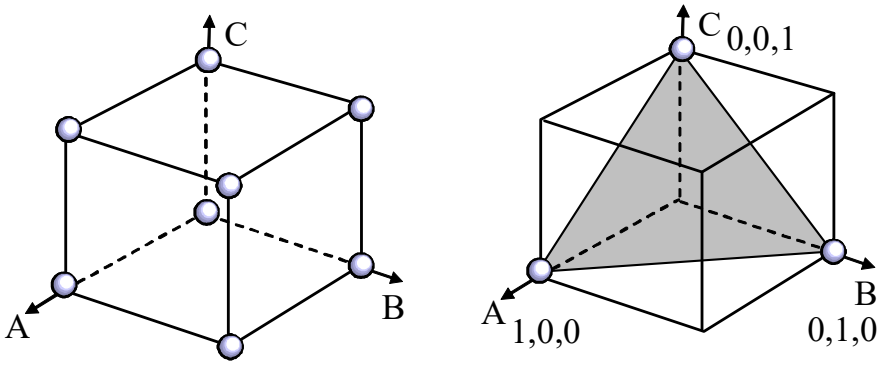
- The test plan is not orthogonal, however, the deviations are usually only small

Mixture experiments

Being indicated in the shares in % at experiments at which it e.g., is about mixtures from chemical liquids. The factors are what in normal test plans, are the different components in mixture plans. All shares must show in sum 100% what leads to the following term

$x_1 + x_2 + ... x_k = 1$ k = count of components

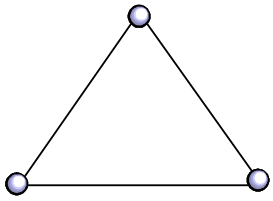
and mean that the components are dependent on each other. This e.g., must be taken into account for the respective tests and can't be treated by standard test plans (only with effort). The possible quota combinations lie in an equilateral triangle.
In most cases there are 3 components. The corresponding test plan looks like represented on the right in comparison with the "conventional" one:



Full factorial Mixed

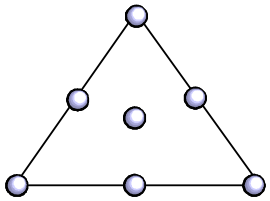
Combinations must be within the range represented grayed. At k=4 components = the possible combinations lie in a tetrahedron. Simplexes are called triangle, tetrahedra and the corresponding arrangements at more than 4 components, the mixture plans are therefore also described as a simplex-plans. For the regulation of only the "main effects" a plan is a so-called type "grade 1" uses. This corresponds to a linear test plan.

No.	comp. A	comp. B	comp. C
1	1	0	0
2	0	1	0
3	0	0	1



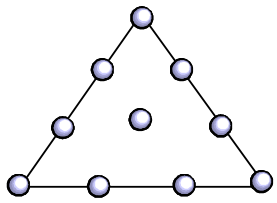
A test plan of the type grade 2 shows the following combinations (in addition with use of all components in the last line):

No.	comp. A	comp. B	comp. C
1	1	0	0
2	0	1	0
3	0	0	1
4	1/2	1/2	0
5	0	1/2	1/2
6	1/2	0	1/2
6	1/3	1/3	1/3



Interactions and nonlinearities can hereby be detected. The next level is grade 3, what is shown in the following table:

Nr.	comp. A	comp. B	comp. C
1	1	0	0
2	0	1	0
3	0	0	1
4	1/3	2/3	0
5	2/3	1/3	0
6	0	1/3	2/3
7	0	2/3	1/3
8	1/3	0	2/3
9	2/3	0	1/3
10	1/3	1/3	1/3



With increasing factors and grade the number of tests increases fast as the following table points:

compon.	Grade 1	Grade 2	Grade 3	Grade 4
2	2	3	4	5
3	3	6	10	15
4	4	10	20	35
5	5	15	35	70
6	6	21	56	126
7	7	28	84	210

Number of tests into dependence of the number of components and of the type

General is the formula

$$m = \frac{k(k+1)(k+2)....(k+g+1)}{1 \cdot 2 \cdot 3.....g}$$

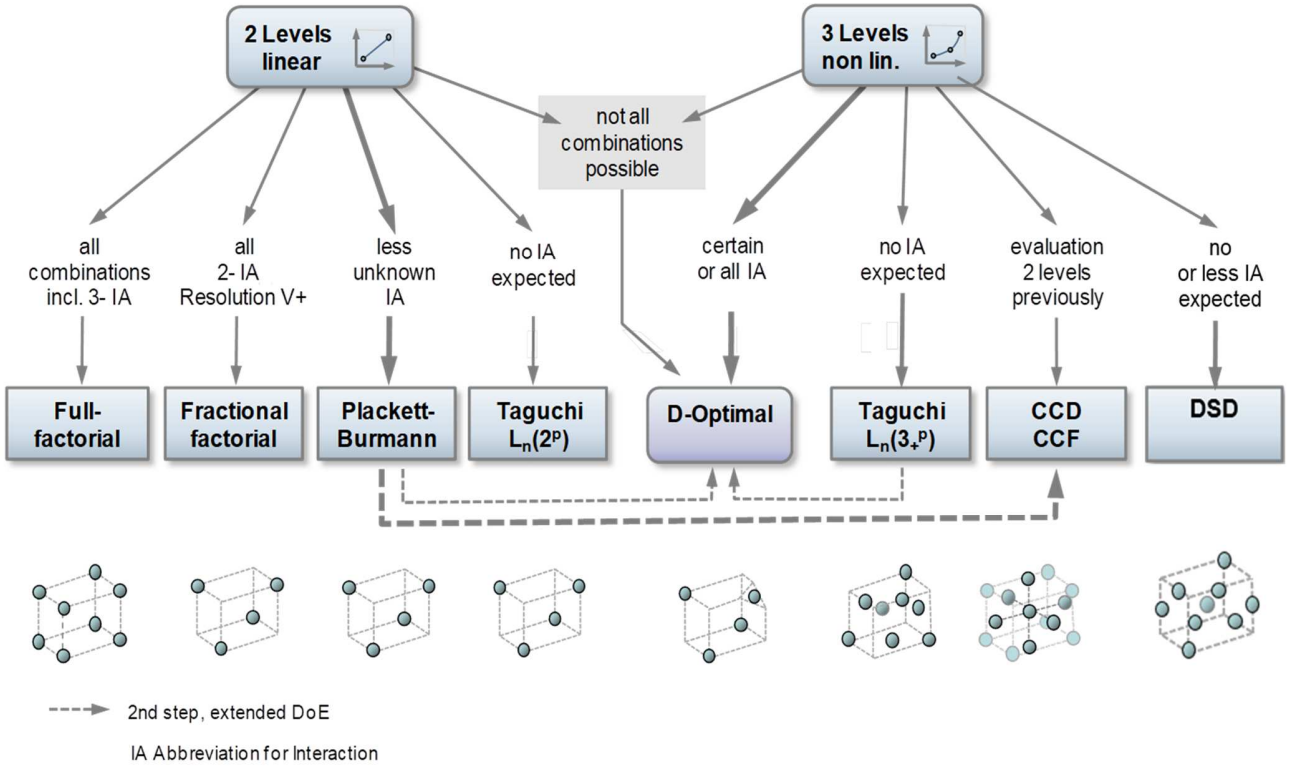
k = number factors, g = grade

To limit the effort, one uses also here D-Optimal. The procedure is comparable with the conventional plans why be further come in here on this shall not.

The evaluation of mixture plans is carried out with the help of the multiple regression. grade 1 corresponds to the model linear, grade 2 squarely etc. The condition $x_1 + x_2 + .. x_k = 1$ is the reason, however, that some of the coefficients generally approach disappear. But the evaluation can be done via Neural Network anyway.

Comparison of Designs

	Full factorial	Fractional	Plackett-Burman	Taguchi	CCD	DSD	D-Opt.
orthogonal	✓	✓	✓	✓	✓	✓	–
quadratic	✓	–	–	partly	✓	✓	✓
cubic	✓	–	–	partly	✓	-	✓
Inter-actions	✓	partly type IV or type V+	partly by enough DF	–	depends on basis	(✓)	✓
Number experiments	very large	middle	small	very little	large	very little	small
partly eval. previously	–	–	–	–	✓	–	–
contrains possible	–	–	–	–	–	–	✓



Correlation

If a connection exists between different factors (dataset), the degree or the strength of this connection can be ascertained with the correlation.

Correlation coefficient after Bravais - Pearson

The measurement of the degree of this connection is the correlation coefficient r. For two dataset x and y, r is calculated after Bravais - Pearson with:

$$r_{xy} = \frac{s_{xy}}{s_x\, s_y} = \frac{\sum\left((x_i - \bar{x})(y_i - \bar{y})\right)}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

With the help of the t-test the hypothesis can be checked: x and y can be considered as two independent datasets. The test statistic is:

$$t_{pr} = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}}\sqrt{n-2}$$

The hypothesis on independence is rejected, if

$$\left|t_{pr}\right| > t_{n-2,1-\alpha/2}$$

The correlation coefficient after Bravais-Pearson strongly reacts to outliers in the observations. Hence, the dataset should be normally distributed.

Rank correlation - Spearman

If the dataset is strongly non normally distributed or if there are categorial attributes, the rank correlation has to be used. Instead of the values the ranking of the sorted data is used. For example, for x = [5;2;7;4] the rank of the value 5 is R=3. The Spearman correlation coefficient is calculated with:

$$r_s = 1 - \frac{6\sum_{i=1}^{n}\left(R(x_i) - R(y_i)\right)^2}{n(n^2-1)}$$

Also, here the t-test is used to check if the datasets x and y can be considered as two independent datasets.
For normally distributed data the difference between Bravais-Pearson and Spearman is low.

Correlation matrix

If there are more than two datasets (factors), each pair can be shown in a matrix. The diagonal contains the value 1.0 (correlation to itself is 100%).
The correlation coefficients of lower left half are same to mirror with upper right half, because $r_{x1x2} = r_{x2x1}$ etc.

	x_1	x_2	x_3	..	x_n
x_1	1.0	r_{x2x1}	r_{x3x1}	..	r_{xnx1}
x_2	r_{x1x2}	1.0	r_{x3x2}	..	r_{xnx2}
x_3	r_{x1x3}	r_{x2x3}	1.0	..	r_{xnx3}
..	..	..	..	1.0	..
x_n	r_{x1xn}	r_{x2xn}	r_{x3xn}	..	1.0

Partial Correlation Coefficient

The partial correlation coefficient describes the dependence of two factors without influence of a third factor. One can also say, how is the influence from x to y if z is eliminated or is held steady. The formula is:

$$r_{xy.z} = \frac{r_{xy} - r_{xz}\, r_{zy}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}$$

Hereby there can be uncovered so-called spurious correlation. Also here is used the t-test. The hypothesis is: x and y are independent without the influence of z. Nevertheless, the degree of freedom is reduced around one and it is:

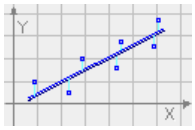
$$t_{pr} = \frac{r_{xy.z}}{\sqrt{1-r_{xy.z}^2}}\sqrt{n-3}$$

The hypothesis on independence is rejected, if

$$\left|t_{pr}\right| > t_{n-3,1-\alpha/2}$$

In Visual-XSel use the menu ***Statistics*** in the spreadsheet.

4. Regression



General

If there is a connection between different features, then the degree or the strength of this connection can be determined with the help of correlation. The correlation coefficient r describes the strength of the connection.
One tries at the regression calculation to put a line or curve adapted to the measurement pairs optimally. This is a compensation straight line in the simplest case at linear slope. One understands the determination of the coefficients of the compensation straight line by an optimal customization in that way that this difference of the straight line becomes a minimum (least square method). The correlation coefficient expresses how good the found equation adapts to the measurements. The nearer r is due to 1, the better the precision is. In any case there must be always more data than model coefficients exist.
There is not always a linear connection. The main problem of the regression calculation is to find the right function. At the choice of the suitable function for the regression one should therefore watch the course of the measurements exactly at first and regard maybe known physical dependencies.

Linear Regression

The linear regression is defined through:

$$Y = a + b\,x$$

The gradient b and the section of the straight lines by the y-axis a is calculated through:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad a = \bar{y} - b\,\bar{x}$$

The confidence interval for the expected value $\hat{y}_i$ at the position x_i is calculated through the min und max-value:

$$Y_u = a + b\,x_i - C \qquad Y_o = a + b\,x_i + C$$

with

$$C = s\,t_{n-2,1-\gamma/2}\sqrt{\frac{1}{n} + \frac{(\bar{x} - x_i)^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}}$$

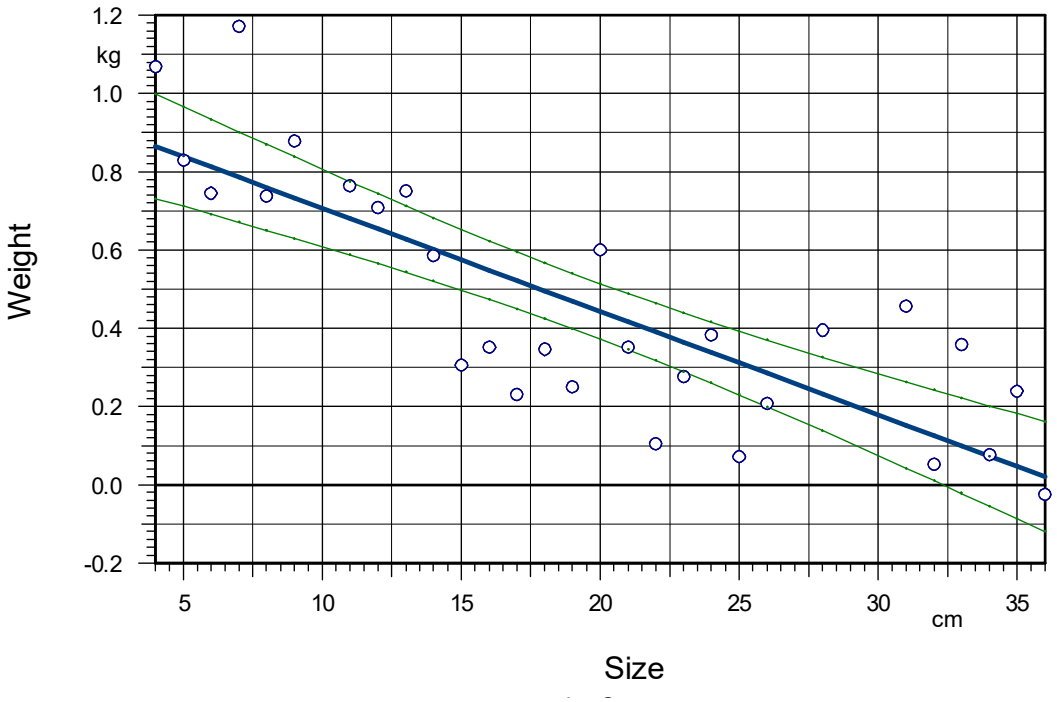
The estimated standard deviation s is calculated from the variance by the deviations of the observations to the compensation straight line:

$$s^2 = \sum_{i=1}^{n} \left(Y_i - (a + b\,x_i)\right)^2$$

Each position of x_i results a different wide confidence bounds along the straight line defined through:

$Y_{unten} = a + b\,x - C$ and $Y_{oben} = a + b\,x + C$

which is at least at $x_i = \bar{x}$:



Linear regression through 0-point

In certain cases, the facts force, that the compensation straight goes by the 0 point. The standard equation $Y = a + b\,x$ becomes:

$$Y = b\,x \quad \text{with} \quad b = \frac{\sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2}$$

Nonlinear regression

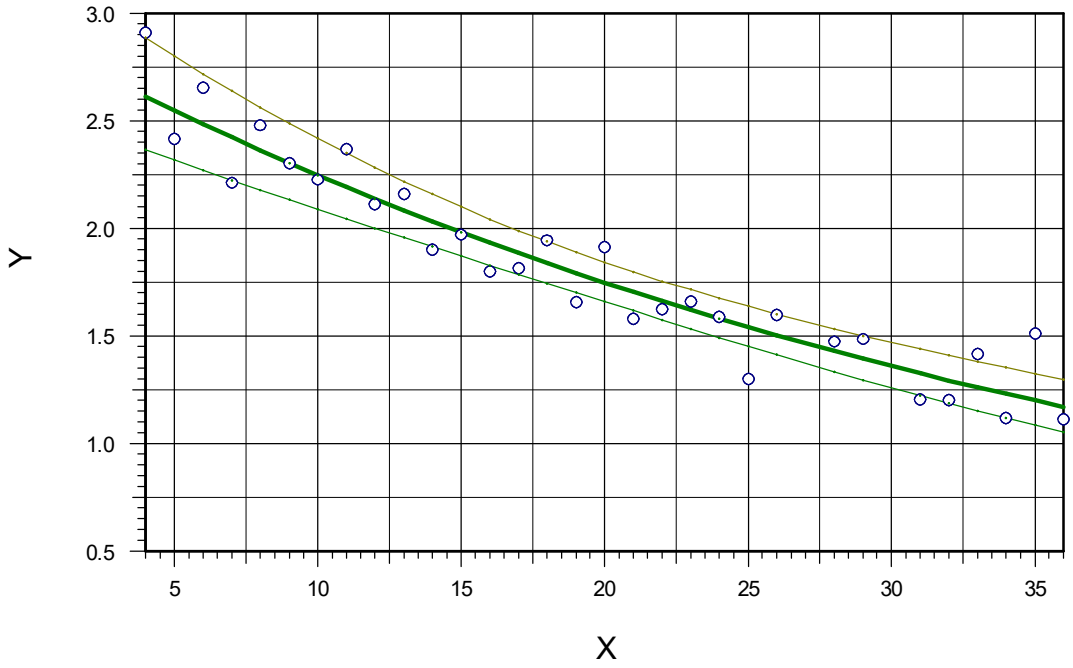
A nonlinear curve is for example $Y = a\,e^{b\,x}$. The standard deviation is here:

$$s^2 = \sum_{i=1}^{n} \left(Y_i - a\, e^{b\, x_i} \right)^2$$

C is calculated like by the linear regression. The confidence interval is adequate:

$Y_{unten} = a\, e^{b\, x - C}$ and $Y_{oben} = a\, e^{b\, x + C}$

For example:



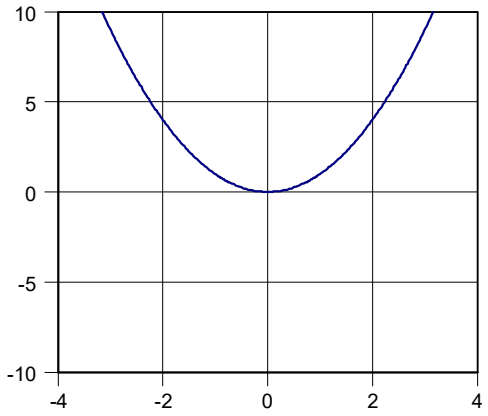
Regression types

Under the button ***Regression*** in the dialogue window ***Diagram types*** find the following represented functions, where it is up to 7 degrees possible for polynomial:

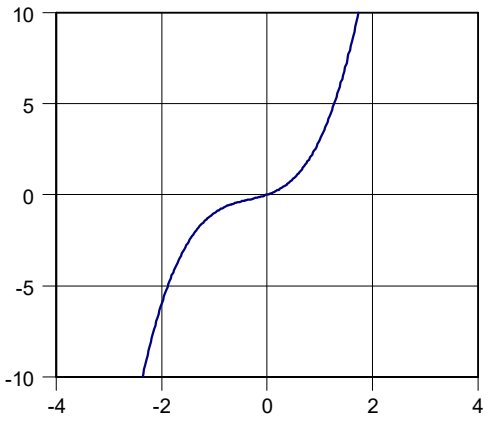
$Y = a\, x\text{^}b$	Straight line in a double logarithm scale
$Y = a + b \cdot x$	Simple straight line
$Y = a + b \cdot x + c \cdot x^2$	
$Y = a + b \cdot x + c \cdot x^2 + d \cdot x^3$	
$Y = a + b \cdot x + c \cdot x^2 + \ldots$	Polynomial up till 7th grade
$Y = a \cdot e\text{^}(b \cdot x)$	
$Y = a \cdot e\text{^}(b/x)$	
$Y = a + b/x$	
$Y = a + b \cdot log(x)$	Straight line in a single logarithm scale

To find the right function choice the following examples of the most important types are shown below (coefficients -1 …. +1):

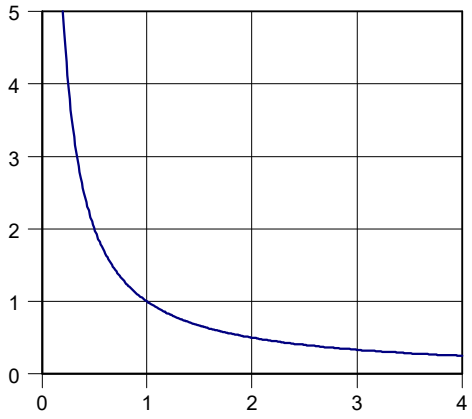
$Y = x^2$



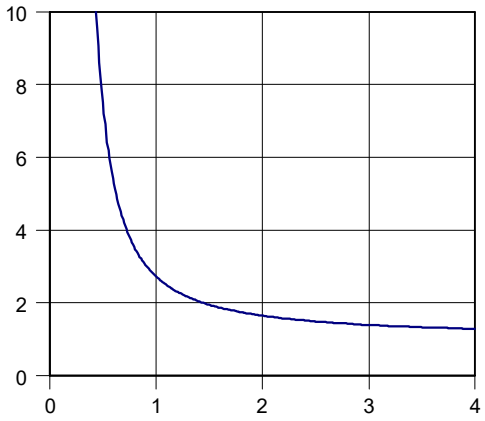
$Y = x + x^2 + x^3$



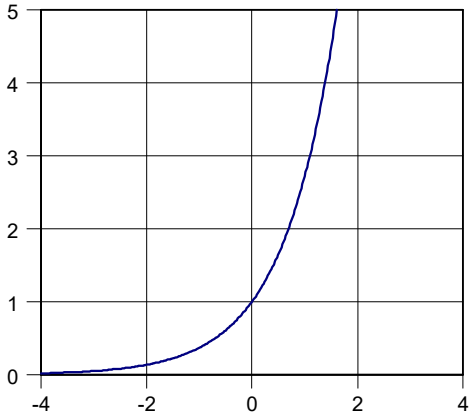
$Y = \frac{1}{x}$



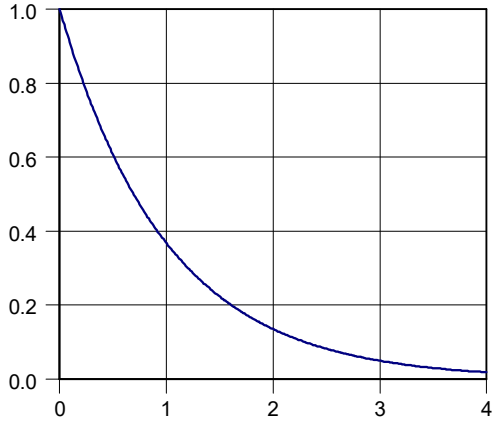
$Y = e^{\frac{1}{x}}$



$Y = e^x$



$Y = e^{-x}$



Courses which have a maximum or a minimum happen frequently. A typically function with a minimum in point 0 is $Y=X^2$. If there are data points which goes not through the 0-point, there must be an offset like $Y=X^2+b$. A regression of a parabola determines this offset b automatically. If the minimum is on the right or on the left of the Y-axis the parabola fails. The x-data column has to be moved to the y-axis necessarily around the value of the moving.

For 3D-charts with two independent variables x and z the following basic functions are available:

$$Y = a + b \cdot x + c \cdot z$$
$$Y = a + b \cdot x + c \cdot z^2$$
$$Y = a + b \cdot x^2 + c \cdot z$$
$$Y = a + b\, x^2 + c \cdot z^2$$

The functions produced after the regression with concrete coefficients are in the Formula interpreter and can be changed afterwards. Perhaps this makes sense if single coefficients from other experiences are known. In this case there is no longer connection to the previous found coefficients

Multiple Regression

One uses a multiple regression if more than one independent factor x is available. The simple linear model is:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + ...$$

It is presupposed that the features are normal distributed and linear. E.g., not linear parameters can be realized in most cases by remodeling or by using squared terms:

$$y = b_0 + b_1 \boldsymbol{x_1} + b_2 \boldsymbol{x_1^2} + b_3 x_2 + ...$$

In case of tabular values this means that one adds the column to x with the values in a new column copied and squared. E.g., a combination two influences which represents an interaction also can be carried out:

$$y = b_0 + b_1 \boldsymbol{x_1} + b_2 \boldsymbol{x_1 x_2} + b_3 x_2 + ...$$

The corresponding table columns for x then have to be inserted in a new column as a product x². Further conversions are possible to reach the linear model. In matrix form the model equation is:

$$\hat{y} = b\,X$$

with $\hat{y}$ = vector of the results from the parameter set
X = matrix of the actual parameter values
b = vector of the coefficients

$$y = \begin{bmatrix} y_1 \\ y_2 \\ .. \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & .. & x_{z1} \\ 1 & x_{12} & .. & x_{z2} \\ .. & .. & .. & .. \\ 1 & x_{1n} & .. & x_{zn} \end{bmatrix} \quad b = \begin{bmatrix} b_o \\ b_1 \\ .. \\ b_z \end{bmatrix}$$

Hint: 1st column represents in X the constant

The sought-after vector b with the coefficients determines about the matrix operation

$$b = \left(X^T X\right)^{-1} X^T y$$

Example: Interaction model is given:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2$$

The individual steps of the equation $b = \left(X^T X\right)^{-1} X^T y$ arise as follows

	experiment:		results Y
V_1	-1	-1	3
V_2	1	-1	5
V_3	-1	1	7
V_4	1	1	11
V_5	0	0	6

$X' = X^T X$ with $X = \begin{bmatrix} 1 & x_{11} & .. & x_{z1} \\ 1 & x_{12} & .. & x_{z2} \\ .. & .. & .. & .. \\ 1 & x_{1n} & .. & x_{zn} \end{bmatrix}$

$z+1$ columns and n rows

The respective cells are calculated after each other:

$$x'_{j,i} = \sum_{k=1}^{n} x_{k,i}^{(T)} \, x_{j,k}$$ (1st index = column, 2nd index = row)

The first column represents the constant b_o. The following columns are the factors x_1 and x_2 and the last column is the product of x_1 and x_2 (interaction).

$$X = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \qquad X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & 0 \\ -1 & -1 & 1 & 1 & 0 \\ 1 & -1 & -1 & 1 & 0 \end{bmatrix}$$

etc. cells

j=1 i=1
$x'_{1,1} = (1)\cdot(1) + (1)\cdot(1) + (1)\cdot(1) + (1)\cdot(1) + (1)\cdot(1) = 5$

j = 2 i = 2
$x'_{2,2} = (-1)\cdot(-1) + (1)\cdot(1) + (-1)\cdot(-1) + (1)\cdot(1) + (0)\cdot(0) = 4$

this yields to:

$$X' = X^T X = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

and the revers matrix is:

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 1/5 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 \end{bmatrix}$$

and via the intermediate step

$$X^T y = \begin{bmatrix} 32 \\ 6 \\ 10 \\ 2 \end{bmatrix}$$

one gets the result for the sought-after coefficients:

$$b = \left(X^T X\right)^{-1} X^T y = \begin{bmatrix} 6{,}4 \\ 1{,}5 \\ 2{,}5 \\ 0{,}5 \end{bmatrix}$$

So, the equation of the beginning is:

$$y = 6{,}4 + 1{,}5 x_1 + 2{,}5 x_2 + 0{,}5 x_1 x_2$$

Categorical Factors

Categorical or qualitative factors whose variations are indicated in the form of textual names must be brought in suitable number form. One uses -1 and +1 for two attitudes in a column. If the categorical factor is e.g., a component of supplier A and supplier B, then A gets the value -1 and B the value 1. As of every broader feature (variation) an additional column is laid out.

	F [B]	F [C]	F [D]
A	-1	-1	-1
B	1	0	0
C	0	1	0
D	0	0	1

The attitude A of the generally mentioned factor F represents the basic level. The corresponding line therefore contains -1 everywhere. The other variations have one in their column 1.
Partial correlations of r have construction caused test plans with categorical factors r = 0.5 or more greatly.

Repetitions – Sample size

Through repetitions of the experiment, you want to make sure not to overlook certain effects. So also, the type 2 error beside the type 1 error is relevant. The effect must be significant higher than the scatter. The ratio must have at least the amount of the quantile of the type 1 error with $1-\alpha/2$ and additional of the quantile of the type 2 error with $1-\beta$. Thus one can write approximately by using the normal distribution:

$$\frac{\Delta}{s_\Delta} \approx u_{1-\frac{\alpha}{2}} + u_{1-\beta}$$

or better because of having only at spot check (sample):

$$\frac{\Delta}{s_\Delta} = t_{1-\alpha/2} + t_{1-\beta} \quad \Rightarrow \quad \left(\frac{\Delta}{s_\Delta}\right)^2 = \left(t_{1-\alpha/2} + t_{1-\beta}\right)^2$$

The variance of the effect is

$$s_\Delta^2 = \frac{4}{N} s^2 = \frac{4}{n_{plan}\,(n_w+1)} s^2$$

N = Complete number of experiments
n_{plan} = Combinations the DoE e.g. 2^{p-1}
n_w = Number of repetitions

(Factor 4 => 2-levels experimental design -> half the number of tests for effects and the variances of the differences of two mean values are twice as large as the variance of a mean value).

$$\frac{\Delta^2}{4\,s^2}\, n_{plan}\,(n_w+1) = \left(t_{1-\alpha/2} + t_{1-\beta}\right)^2$$

Thus, the required number of repetitions is:

$$n_w = \frac{1}{n_{plan}} \frac{s^2}{\Delta^2} 4\left(t_{1-\alpha/2} + t_{1-\beta}\right)^2 - 1$$

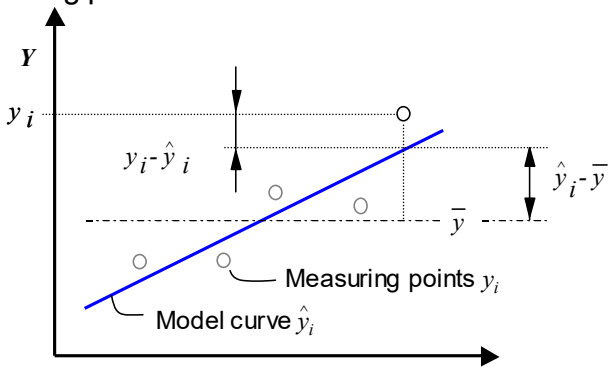
Assuming that the number of degrees of freedom in the model is about DF = 10 (10 attempts more than the model would have been necessary), we obtain for the mean quotation: $4\left(t_{1-\alpha/2} + t_{1-\beta}\right)^2 = 4\left(t_{0,975} + t_{0,90}\right)^2 = 4\,(\,2{,}23 + 1{,}37\,)^2 \approx 52$

If one calculates with β = 20%, the quotation will be 39. At D-Optimal experiments 3-5 additional tests are recommended. This is the first number of DF. After the step-wise-regression some terms will be exclude from the model, so that we can estimate DF ≈ 10. Kleppmann /3/ calculates with the factor 60, so the final equation will be:

$$n_w \approx \frac{60}{n_{plan}} \frac{s^2}{\Delta^2} - 1$$

Analyses of Variance (Model ANOVA)

For assessment of the regression model the most important index is the coefficient of determination R² and then adjusted coefficient of determination R²adj.
The closer R² is to the value 1, the better the model y is described through x. The smaller R² is the values scatter is higher and there is not the slightest connection to y.
The following picture shows the connection between measuring and the model for one factor



$$SS_{Total} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 \quad SS_{Reg} = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 \quad SS_{Res} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

$$SS_{Total} = SS_{Reg} + SS_{Res}$$

$$R^2 = \frac{SS_{Reg}}{SS_{Total}} = 1 - \frac{SS_{Res}}{SS_{Total}} \qquad 0 \le R^2 \le 1$$

One frequently also finds the adjusted coefficient of determination R ²adj. The corresponding degrees of freedom are taken into account

$$R^2_{adj} = 1 - \frac{SS_{Res} / DF_{Res}}{SS_{Total} / DF_{Total}} = 1 - \frac{MS_{Res}}{MS_{Total}}$$

MS : Variance
DF_{Reg} : Degrees of Freedom of regression -> number of X-variables in model $DF_{Reg} = z - 1$
(z = number of model-terms x_1, x_2, x_3, $x_1 \cdot x_2$, x_1^2 ….)
DF_{Res} : Degrees of Freedom of the residuals $DF_{Res} = n - z - 1$
(n = number of experiments)
DF_{Total} : Degrees of Freedom total = n

For great data sizes are like A and B brought closer. The smaller the data size gets, the bigger the deviation is. R² overestimates the declared amount of deviation considerably at a small number of degrees of freedom from time to time. Great differences between R^2 and R^2adj indicate unnecessary terms in the model.

Prediction Measure Q²

The Prediction measure is the fraction of variation of the response that can be predicted by the model.

In principle R^2 rises up with more coefficients in the model because these then can adapt to the test points always better ($SSres$ decreases). R^2 isn't suitable to recognize whether the model is over-determined. For this the Q^2 measure has been defined:

$$Q^2 = 1 - \frac{\sum\left(y_i - \hat{\hat{y}}_i\right)^2}{\sum\left(y_i - \bar{y}_i\right)^2}$$

with $\hat{\hat{y}}_i$ = model prediction for not measured points

Q^2 also can get negative if the point is bigger than the denominator.

Hints:

R² and Q² is small

The customization of the model is bad. This can have several causes:

- Outliers
- Wrong test order
- Bad reproducibility

Corrective: Checking the measurements for plausibility. Perhaps carrying out the tests once again.
Bad test plan, possible carry out a new plan for one.

R² high and Q² very small

The model offers a good description, is, however, unstable. Tendency toward the over-determination
There are too many terms or interactions taken into account. The model should be reduced. The terms with the smallest effects should be deleted from the model, but be careful with significant interactions.

- There are dominant outliers
- One response must be transformed
- The investigations should be going on

Note:
- In case of lean experiments (screening plans), often the Q^2 is worse than the model is.
- In case of many repetitions, the Q^2 is better than the model is. Therefore, it should be analyzed much more the lack of fit.

Lack of Fit

Some further information can be analyzed from the residual. SS_{res} is put together out:

$$SS_{res} = SS_{LoF} + SS_{p.e.}$$

SS_{LoF} is the Lack of Fit, with the degrees of Freedom $DF_{LoF} = n - z - DF_{p.e.} - 1$
$SS_{p.e.}$ is the pure error determined from repetitions.

$$SS_{p.e.} = \sum_{j=1}^{r}\sum_{k=1}^{r_j}\left(Y_{j,k} - \bar{Y}_j\right)^2 \quad \text{with the Degrees of Freedom} \quad DF_{p.e.} = \sum_{j=1}^{r}\left(r_j - 1\right)$$

Is SS_{res} and $SS_{p.e}$ known, the equation for the Lack of Fit is:

$$SS_{LoF} = SS_{res} - SS_{p.e}$$

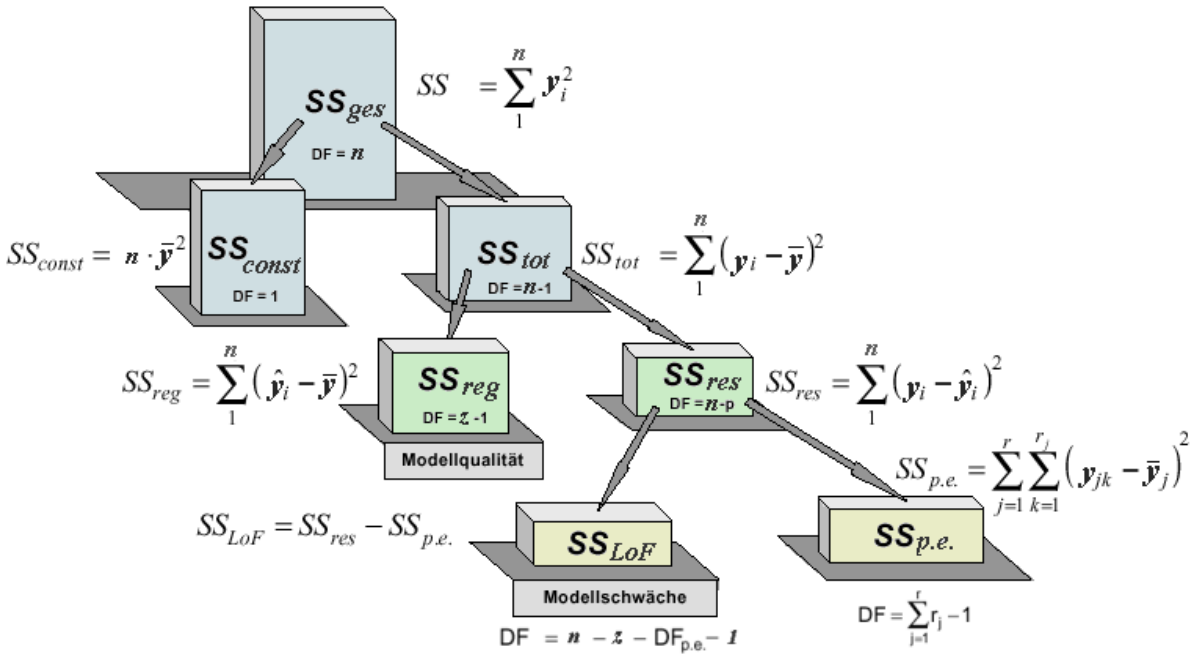
The quotient of the variances is then the Lack of Fit:

$$\frac{MS_{LoF}}{MS_{p.e.}} = \frac{SS_{LoF} / DF_{LoF}}{SS_{p.e.} / DF_{p.e.}} > F_{DF_{Lof}, DF_{p.e.}, \gamma}$$

The result is to compare to a critical F-worth (γ=confidence interval). Obviously if this is bigger than the model terms are contained too little.

Analyses of Variance overview

The following picture shows an overview to the total Analyses of Variance:



Reproducibility

The Reproducibility is described through the following equation:

$$\text{Reproducibility} = 1 - \frac{MS_{p.e.}}{MS_{total}}$$

This is a relative indicator which says as good we are able to reproduce the tests. This indicator can only be determined with repetitions of tests.

Test of the coefficient of determination

As you described at the beginning is the regression result all the better the nearer the coefficient of determination is due to 1. The question is worth as of which value under 1 the deviation by chance or already is only significant. To this one builds the null hypothesis: All regression coefficients are 0., i.e., no connection between y and x etc. insists. A weighted F value is calculated as test quantity:

$$F_{pr} = \frac{R^2(n-z-1)}{(1-R^2)z}$$

with n experiments and z = number of model terms x_1, x_2, x_3, x_1, x_2, x_1^2 etc. As the result is significantly the regulation becomes the F-distribution with the degrees of freedom to

$$f1 = z\,, \quad f2 = n - z - 1$$

used. According to the significance standard, e.g., 5% or 1%, the regression result is all the better with respect to the correlation coefficient, the nearer the value of the F-distribution is due to 0 and the null hypothesis must be rejected.

The corresponding statistical basics you find in the statistical-literature.

Test of the regression coefficients, the p-value

To determine the significance of a factor, frequently the so-called p-value is used. At first the hypothesis is defined that a coefficient of a factor b=0. Then the p-value is the probability to reject the hypothesis mistakenly. This probability is determined via the t-distribution:

$$t = \frac{b}{s_b}$$

b = coefficient from the multiple regression
s_b = deviation of the coefficient

With using the double value of t because of the two-way test and the degrees of freedom $f = n - z - 1$ (n = count of experiments, z = count of model terms x_1, x_2, x_3, $x_1 \cdot x_2$, x_1^2 etc.). With the index j for each factor t is defined with:

$$t_j = \frac{b_j}{s_{b\,j}}$$

The spread of the regression coefficient is determined through:

$$s_{b\,j} = \sqrt{s^2\, X''_{j,j}}$$

in which s is the standard deviation of the complete model. s is calculated through the sum of squares between the model and the measured values

$$s^2 = \frac{1}{n-z-1} \sum_{i=1}^{n} \left(Y_i - b_o - \sum_{j=1}^{z} x_{j,i}\, b_j \right)^2$$

with b_o = constant term of the model.

X'' is calculated through:

$$X'' = (X^T X)^{-1} \quad with \quad X = \begin{bmatrix} 1 & x_{11} & .. & x_{z1} \\ 1 & x_{12} & .. & x_{z2} \\ .. & .. & .. & .. \\ 1 & x_{1n} & .. & x_{zn} \end{bmatrix}$$

The greater the t-worth is the smaller the p-value becomes. Usually the significance level is 5%, that means if there is a p-value smaller than 0.05 the coefficient is significant.

Test of the coefficient of determination

As you described at the beginning is the regression result all the better the nearer the coefficient of determination is due to 1. The question is worth as of which value under 1 the deviation by chance or already is only significant. To this one builds the null hypothesis: All regression coefficients are 0., i.e., no connection between y and x etc. insists. A weighted F value is calculated as test quantity:

$$F_{pr} = \frac{R^2 (n-z-1)}{(1-R^2) z}$$

with n of experiments and z = number of model terms x_1, x_2, x_3, x_1, x_2, x_1^2 etc. As the result is significantly the regulation becomes the F-distribution with the degrees of freedom to

$$f1 = z\,, \quad f2 = n - z - 1$$

used. According to the significance standard, e.g., 5% or 1%, the regression result is all the better with respect to the correlation coefficient, the nearer the value of the F-distribution is due to 0 and the null hypothesis must be rejected.

Standard deviation of the model RMS

The so-called RMS-Error (Root mean squared error) represents the standard deviation of the complete model. It is calculated through:

$$RMS = \sqrt{\frac{SS_{Res}}{n-z-1}} \quad \text{with} \quad SS_{Res} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

The relative standard deviation is related to the middle data area

$$RMS / Ym$$

and is a further control criterion. This value can also analogously be seen by Taguchi to the reciprocal of the not squared signal-to-noise ratio (without the pre-factor 10 lied)

Confidence interval for the regression coefficient

The confidence interval for the regression coefficient is determined with the spread already introduced above:

$$b_j \pm \sqrt{s^2 X''_{j,j}}\, t_{n-z-1;1-\gamma/2}$$

Confidence interval for the response

For certain values of the factors (adjusting's) the response value can be calculated to Y about the model equation (forecast). The corresponding value has a confidence interval because of the spread of the tests and because of the simplification of the model to the reality. This can be decided on the following relation:

$$\hat{Y} \pm \sqrt{s^2\, x^T X'' x}\;\; t_{n-z-1;1-\gamma/2}$$

with $X'' = (X^T X)^{-1}$ (see above) and x for the corresponding factor adjustments

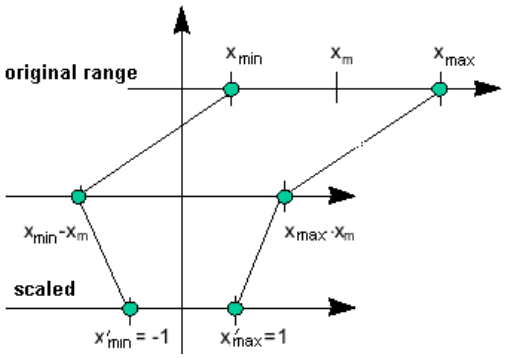
$$x = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ .. \\ x_z \end{pmatrix}$$

and γ for the confidence level, normally 5%. This form is valid under this one assumption that one parameter each are changed, the others however are fixed values (principle as in the case of the effect chart -> non simultaneous confidence interval).

Standardize to -1 ... +1

All data are transformed that the range is between -1 and 1.

$$x_n = \frac{(x - \bar{x})}{(x_{\max} - x_{\min})}$$



Through this one gets a better comparable and relative influence sizes under each other.
In addition, the multiple regression is circumstances permitting only hereby possible when the data areas lie far from each other. The standardization should be used at planned tests.

Standardize to standard deviation

At the standardized form the data values are related and put centrically to her standard deviation:

$$x_s = \frac{(x - \bar{x})}{s}$$

The standardization should be used at historical data or tests not planned since the data values can happen uneven regarding her size (not orthogonal).

The correlation matrix

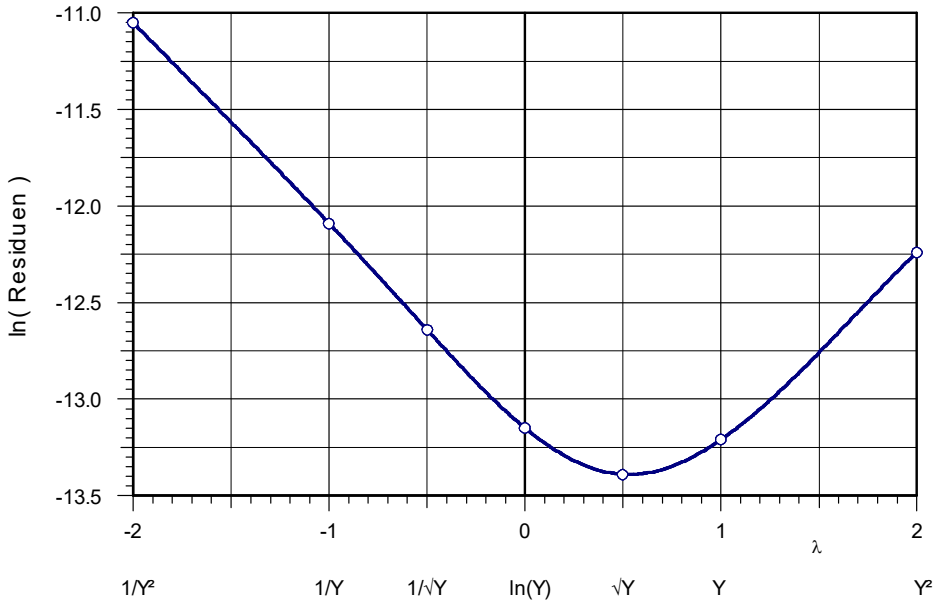
One understands by a correlation a more or less high linear dependence between two variables. The correlation between two factors or between x and y is defined through:

$$r_{xy} = \frac{\sum_{i=1}^{n} ((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

If there is a strong correlation between two x factors, in most cases one of both can be left out.

Response transformation (Box-Cox)

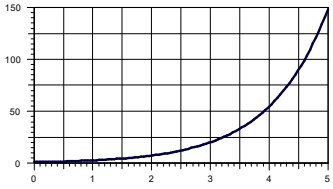
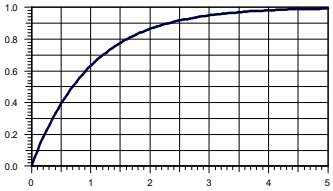
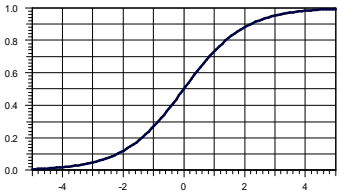
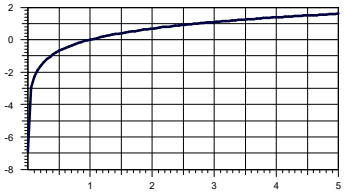
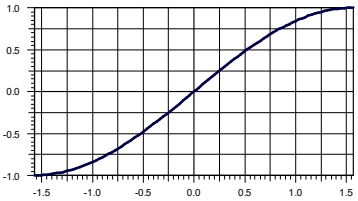
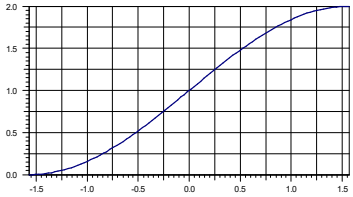
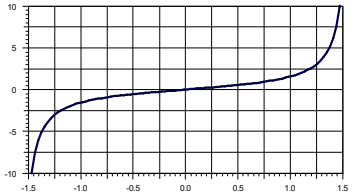
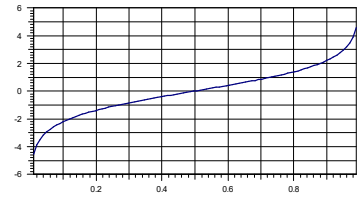
For checking a possibly necessary response transformation the so-called ***Box-Cox***-transformation is used.



One after another the response is transformed according to the functions displayed below and the residues (SSr) are determined.

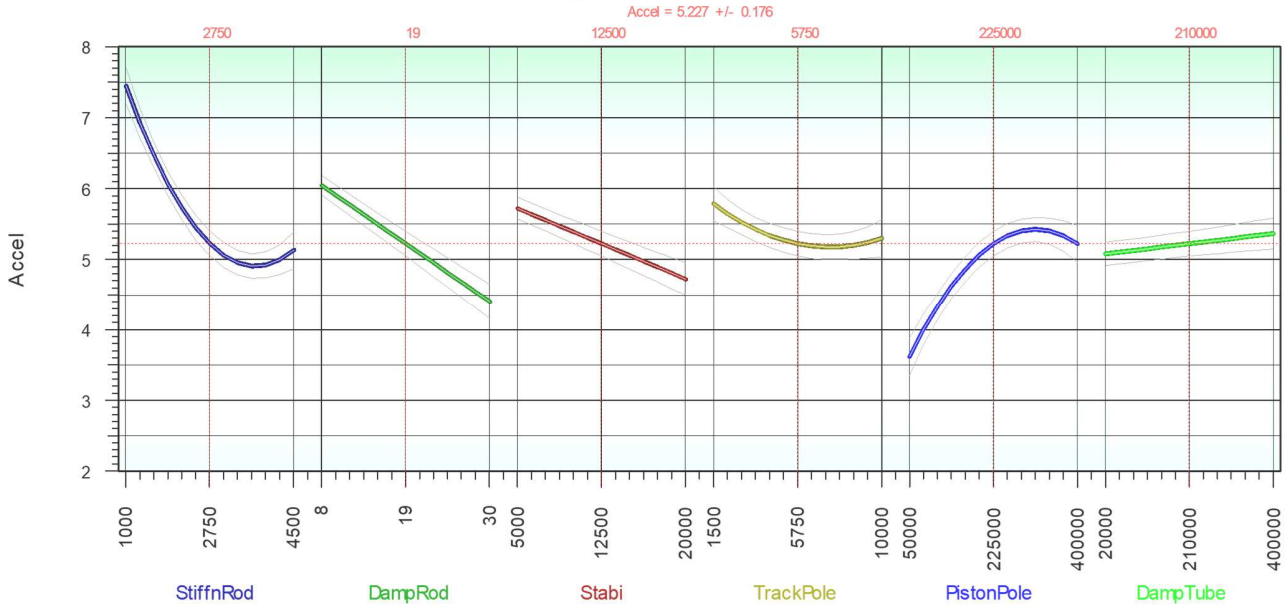
$$Y^{(\lambda)} = \begin{cases} \lambda^{-1}\,\overline{Y}^{1-\lambda}\,(Y^{\lambda}-1) & if\ \lambda \neq 0 \\ \overline{Y}\ln(Y) & if\ \lambda = 0 \end{cases}$$

The smaller the residues and therefore the deviations from the model to the measured data, the better is the transformation to be chosen. This has to be adjusted under the category data, as mentioned in the beginning. It must be pointed out that after the transformation single significances can be changed. Therefore, on the side coefficients it has to be checked, if the model has to be corrected. The Box-Cox-transformation can just be executed, if a target factor-transformation has not yet been chosen.

	Transformation	Inverse function	Example for a'=1, b'=1 c'=0
1	$Y' = a' e^{b'Y} + c'$	$Y = \frac{1}{b'} \ln\left(\frac{Y' - c'}{a'}\right)$	
2	$Y' = a'\left(1 - e^{-b'Y}\right) + c'$	$Y = \frac{1}{b'} \ln\left(\frac{1}{1 - (Y' - c)/a'}\right)$	
3	$Y' = a'\left(1 - \frac{b'}{e^{c'Y} + 1}\right)$	$Y = \frac{1}{c'} \ln\left(\frac{b'}{1 - Y'/a'} - 1\right)$	
4	$Y' = a' \ln(b'Y + c')$	$Y = \frac{1}{b'}\left(e^{\left(\frac{Y'}{a}\right)} - c'\right)$	
5	$Y' = a' \, Sin(b'Y + c')$	$Y = \frac{1}{b'}\left(ArcSin\left(\frac{Y'}{a'}\right) - c'\right)$	
6	$Y' = a'(1 + Sin(b'Y + c'))$	$Y = \frac{1}{b'}\left(ArcSin\left(\frac{Y'}{a'} - 1\right) - c'\right)$	
7	$Y' = a' \, Tan(b'Y + c')$	$Y = \frac{1}{b'}\left(ArcTan\left(\frac{Y'}{a'}\right) - c'\right)$	
8	$Y' = \ln\left(\frac{Y}{1 - Y}\right)$	$Y = \frac{1}{1 + e^{-Y'}}$	

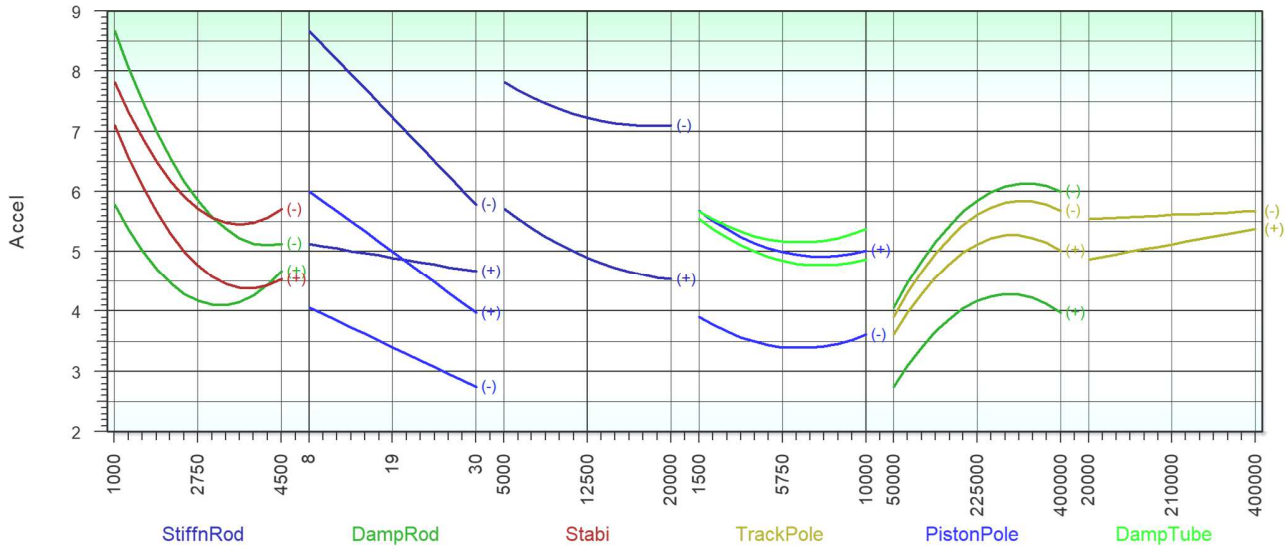
Statistical charts for multiple regression

One of the most important diagrams is the ***Curve-diagram***. Here all runs are depicted for the actual values of factors, marked by vertical red lines.



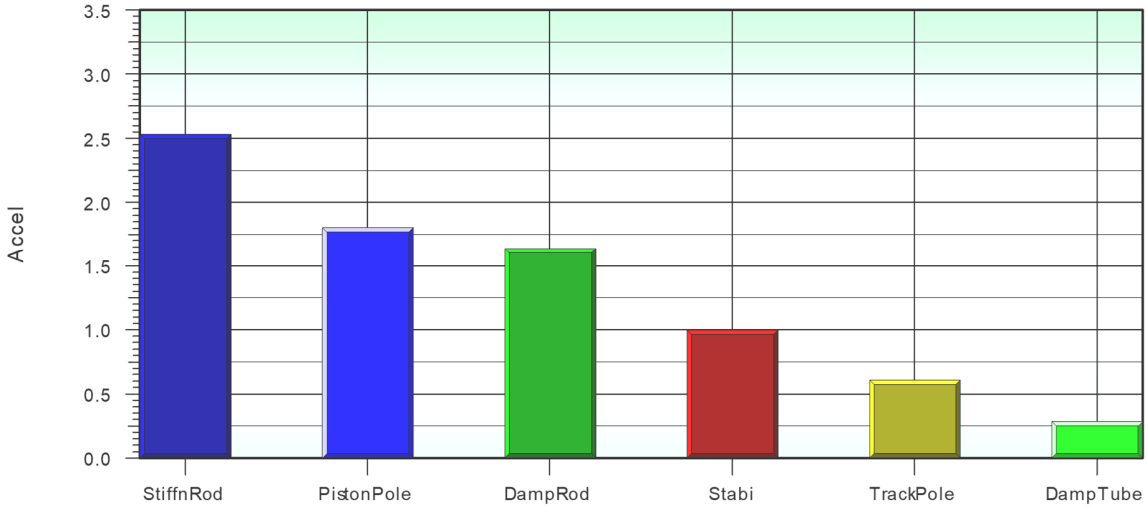
The respective adjustments can be changed by moving these red lines in the graphic with the mouse. By interactions also the other curve linearity's are changed. The horizontal red line always shows the corresponding result value of the target factor. In addition, at indication of a lower or upper limit a blue horizontal line each does exist. The advantage of this depiction is that the math. model is visualized here directly and the gradients are a measure for influences. 15 curves in maximum can be depicted. Thereby the sequence in the list of independent factors under the category model is standard and can be changed there.

The diagram ***Interaction-Chart*** resembles the curve diagram.

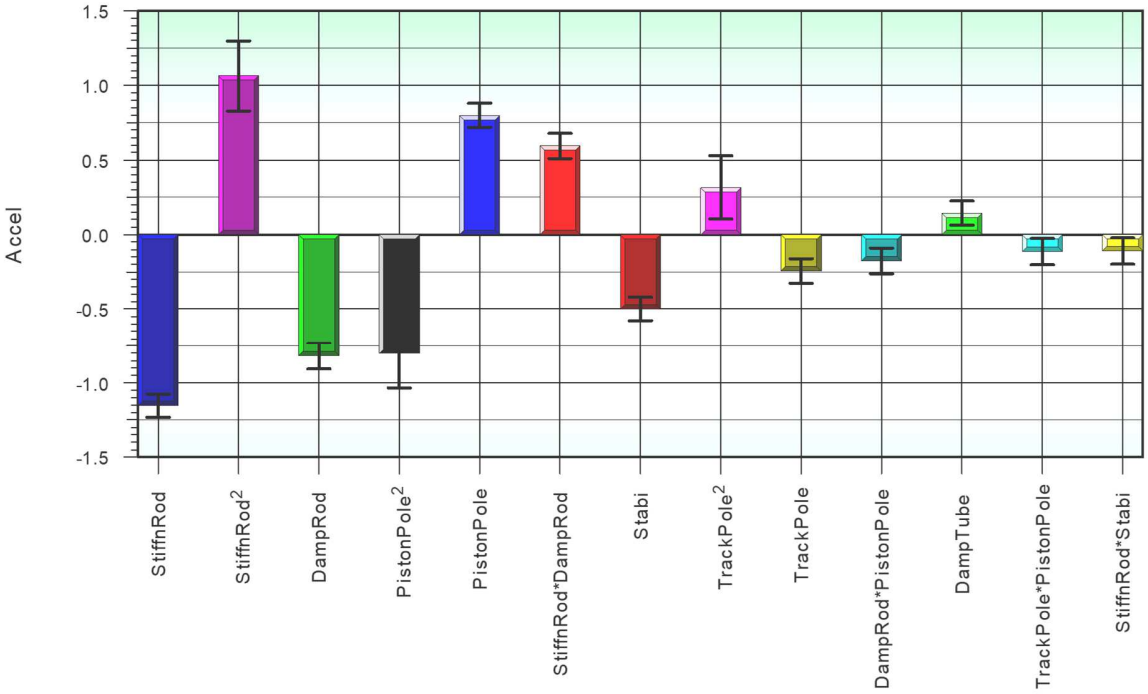


The diagram interactions resemble the curve diagram. The respective curses of curve are represented in pairs here. Every curve couple stands for the respective factors stands with his color with that one of these in an interaction (see color of factor names below the scale). The factor StiffnRod has e.g., an interaction with DampRod. A line about StiffnRod with the identification (+) stands and one with the identification for the upper one (-) for the lower attitude of the factor DampRod. The assignment is possible over the colors. Interactions which aren't significant and taken out of the model aren't represented. So, the complete connection is easily comprehensible in a look.

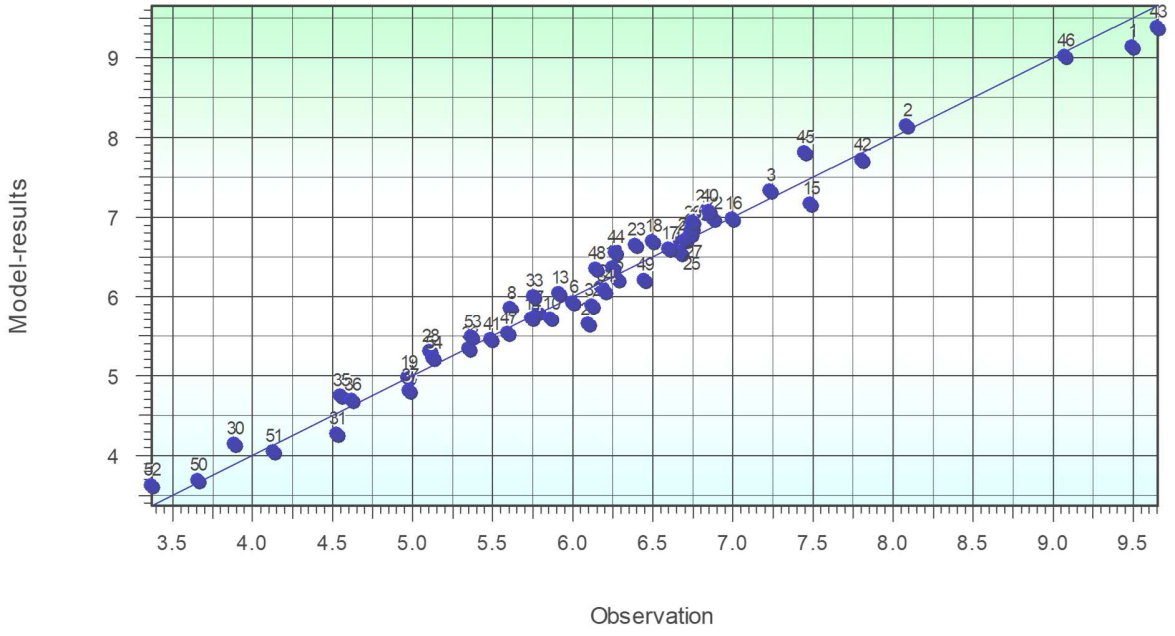
The so-called ***Effects***, which are depicted in an own diagram, are respectively built from the top and lowest point of curves. Therefore, they are dependent of the respectively actual factor adjustments. The effects are depicted as histogram sorted by their absolute size. Here you can recognize directly, where are the largest improvement potentials.



In the ***Pareto-chart*** all model terms are listed, whereas here the 95%-scatter areas are present. Besides this the algebraic sign have been taken into account.
Depending on the number of model terms, however the graphic can be complex

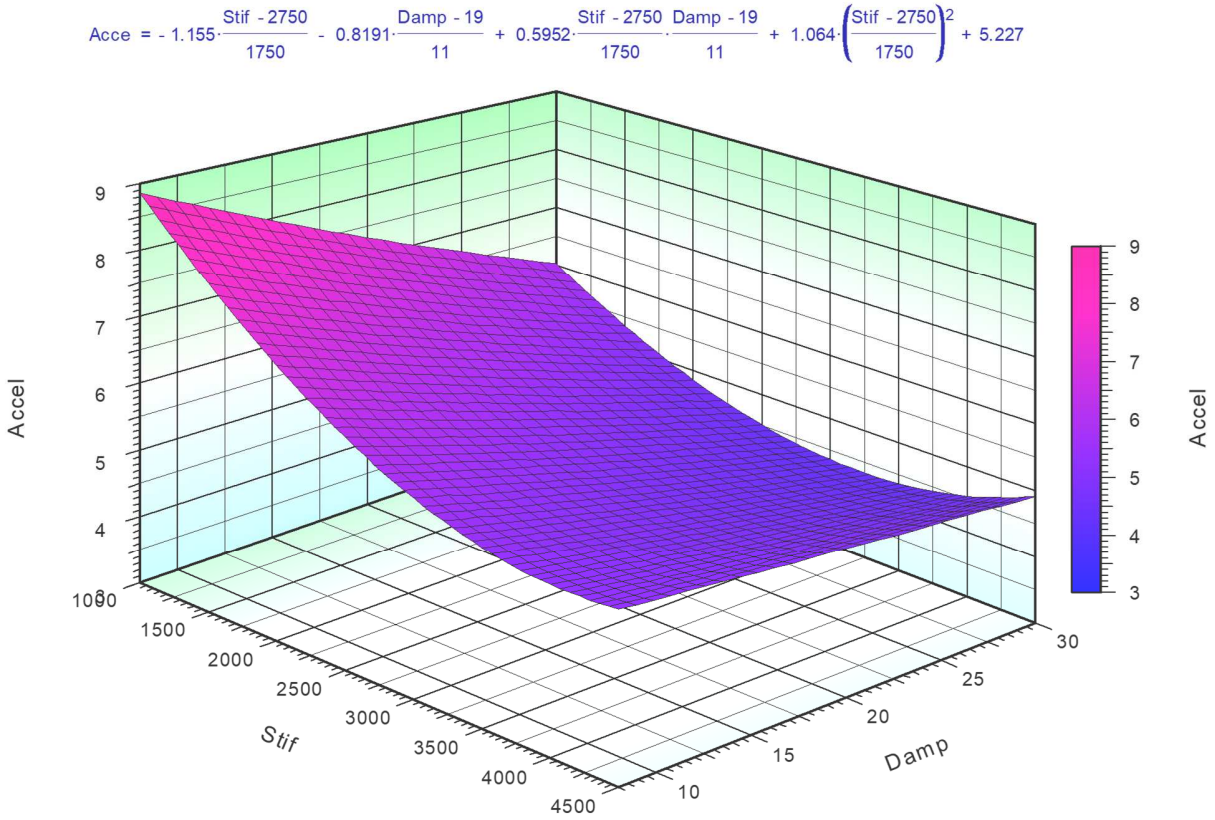


(disadvantage compared to the effect diagram). A further important graphic is the ***Model versus Observations***, whereas here ***Outliers*** are displayed, where the respectively point is red instead of blue.



The better the model and the stability index, the more exact are the model values at the observations, resp. at the measure value. It would be ideal that all points would lie on the 45°-line. The deviations of every point of this line are called ***Residues***. Because of the method of the smallest error squares, the residues should be normal distributed accordingly. They can be depicted in a further diagram:

Especially for depiction of influence of one or two factors to the response a ***2D*** or ***3D-Diagram*** can be chosen.



It makes sense to use the factors, where an interaction exists. The diagram is created via the so-called formula interpreter. Therefore, the both variables (factors) are indicated shortened. The relative long formula over the diagram partly exists because of the re-conversion of standardization, on which the factors refer to. Those again you find in the tabular overview at the beginning. Alternative the diagram type can also be another one, e.g., level-curve diagram. This corresponds to the 3D-view above.

The diagram type is selected under the menu point of the main window ***Diagram/Diagram-type***. After this diagram selection there is no longer an internal reference to the multiple regression. The diagram is seen as independent and is not actualized at modification of factors and so on.

Regulation of outliers

For the regulation of outliers, one looks at the residua of the respective points, i.e., the deviations of the observations (measurements) to the model values. When this deviation is regarded as an outlier the test after Grubbs is recommended. The hypothesis is: xr is an outlier. xr stands for the values of the residua, sr for the standard deviation of the residua

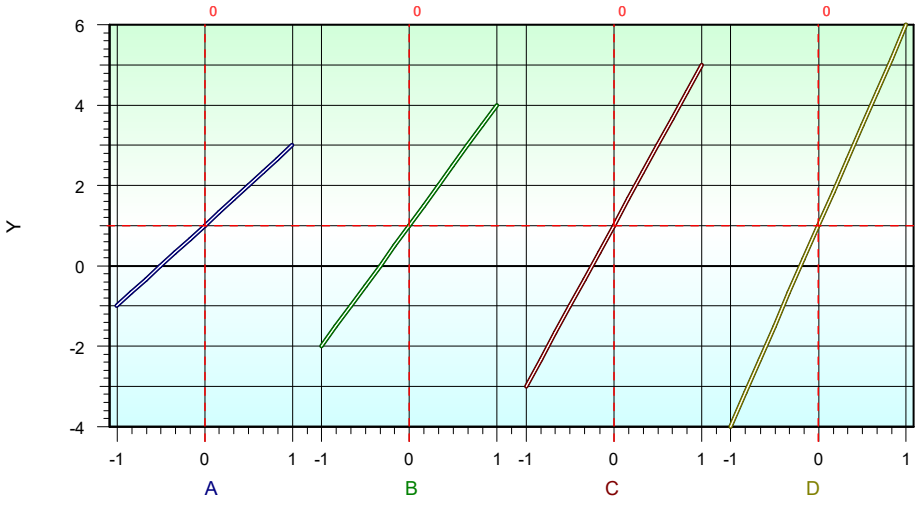
$$T_i = \left| \frac{\bar{x}_{r-} x_{r,i}}{s_r} \right| > T_{n;1-\alpha}$$

$T_{n,1-\alpha}$ is the critical worth of the Grubbs-Test after the following table:

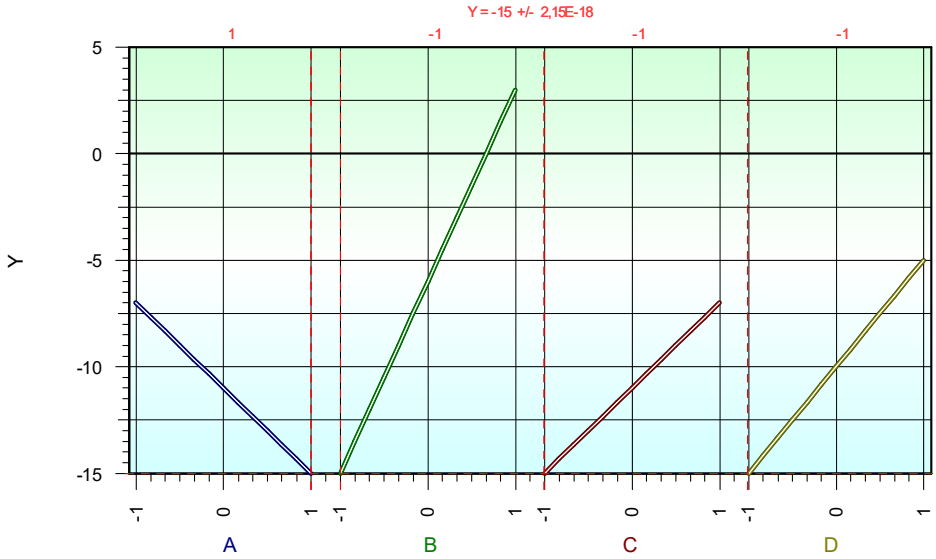
n	$T_{n,0,95}$	$T_{n,0,99}$
3	1,15	1,16
4	1,46	1,49
5	1,67	1,75
6	1,82	1,94
7	1.94	2,10
8	2,03	2,22
9	2,11	2,32
10	2,18	2,41
12	2,29	2,55
15	2,41	2,71
20	2,56	2,88
30	2,75	3,10
40	2,87	3,24
50	2,96	3,34
100	3,21	3,60

Optimization

One understands by an optimization of regression models finding the right adjusting's of all factors for a minimum, maxima or a predefined set point of the response variable.
Example: For the model $Y = 1 + 2 \cdot A + 3 \cdot B + 4 \cdot C + 5 \cdot D + 6 \cdot A \cdot B$ the minimum should be found.



The attitudes of -1 are obviously the best points for all factors. Y has the value-7. Due to the interaction the considerably better minimum is the result, however with Y= -15 by 1;-1;-1;-1



At the search for the best point all mutual attitudes must be checked because of a possible turning back of the gradients.

At the search for an optimal attitude for several response values a conflict can be appear if the best points lie in an opposite direction. A compromise must be found here. One works for it with a so-called fulfilment degree which yields a summarized value for all response values. The result is the corresponding "wish function". At first a plausible significant model is determined and an optimum of each model is found. It can already happen that some factors are not significant for all response variables. After that the optimization of all responses is together carried out via the degree of performance η:

$$\eta = \sum_{i=1}^{m} \left(\left(\frac{Opt_i - Y_{i,j}}{(Max_i - Min_i)} \right)^2 \cdot \delta_i \right)$$

with

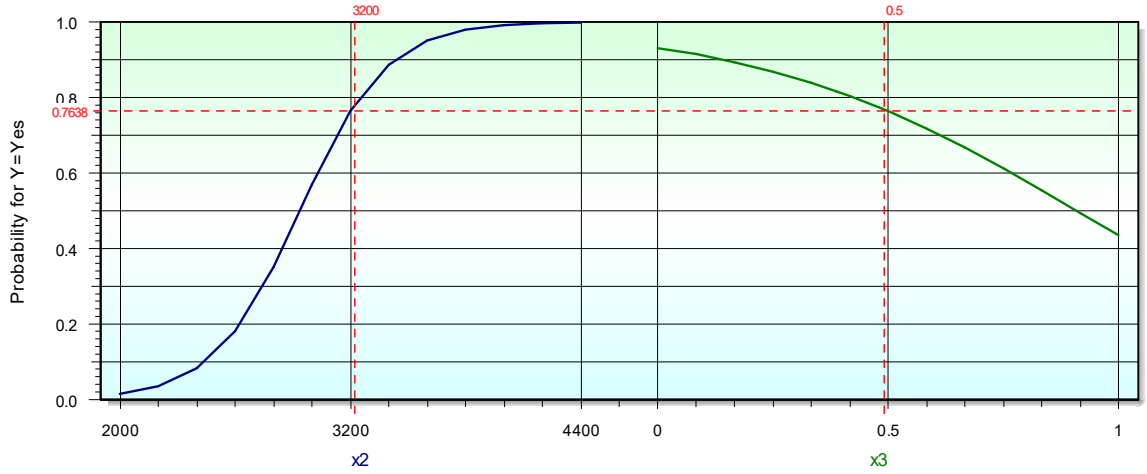
m	:	number of response variables
max/min	:	the respectively greatest and smallest Y value
$Y_{i,j}$	:	current model response for every response value at the continuous variation steps *j*
δ_i	:	weighting factor for every response variable

If certain response values have maybe a higher importance than other, this can be taken into account by a weighting factor δ.

Discrete Regression

$$LH = \prod_{i=1}^{n} \hat{p}_i^{y_i} \cdot (1-\hat{p}_i)^{1-y_i}$$

The Multiple Regression requires steady target values. However, it also can happen that the target value has qualitative character or only 2 expressions (e.g., component has a rip or is o.k.). One usually uses the so-called Discrete Regression for this way of the evaluation. The coefficient of the model is carried out the determination via the Maximum Likelihood-Method. This is in the equal dialogue window as the multiple regression treats. There are some unusual features and restrictions, though. The result describes the probability that the target value takes a certain expression. Therefore, it is to fix data in addition to which this probability applies (here expression 1) in the category.
Being the so-called pseudo-R^2 indicated instead of the certainty measure R^2 in the category of regression. LL consists this of the lying Likelihoods, short. Another indicator is the deviation $D = -2\ LL$. Since in the discrete regression probabilities are treated here does not exist any residual respectively the Sum of Squares. So instead of the ANOVA a combination of the identification values is represented. For this reason, the choice of the graphics does not contain any diagram types which represent residua either.
The Box Cox transformation is not here necessary because the transformation of the target value is already provided tightly on "Logits". The curve diagram contains typical S curves, because for the discrete regression probabilities under 0 and over 1 are not possible.



A special feature of the logistical regression is the evaluation of the *groupings.* The factors are. groups summarized here in classes. The number of response value expressions is counted (*H* event). One divides this number by the group quantity (*H observations),* one gets the observation probability (*P observations),* the probabilities found out with these from the model (*P expected)* compared and tested against a critical χ-value.

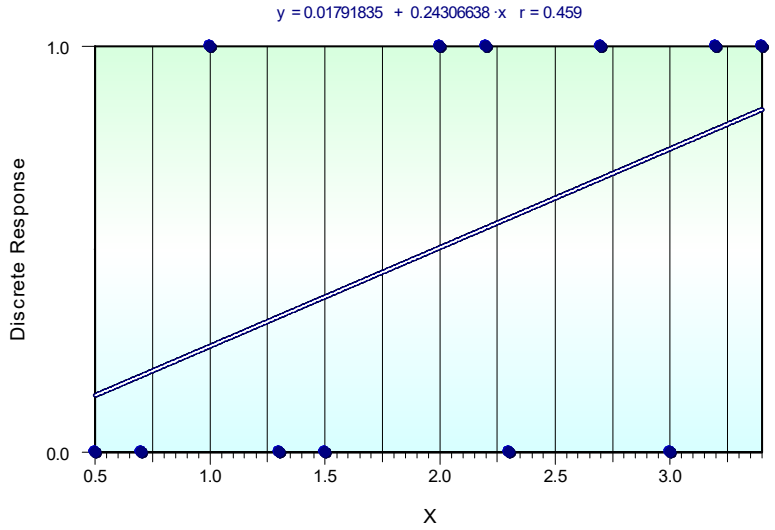
Discrete regression bases

One understands an evaluation by a discrete regression with target values which do not have any steady measurement but qualitative character. The result of an examination could be judged only "well" or "badly", as rip available or not, for example. These state-

ments represent the undermost level of the determinable. It should always be aim to receive the "dissolution" as best as possible, i.e., at least one graduation like a beginning rip, rip by center, rip almost complete and ragged. The evaluation with the standard multiple regression is still possible. The graduation has to be defined with as equal distances as possible.
Furthermore, if only 2 expressions are possible (bad/good or black/white) the following procedure can be applied. For example, the data is given:

x	0,5	0,7	1	1,3	1,5	2	2,2	2,3	2,7	3	3,2	3,4
y	0	0	1	0	0	1	1	0	1	0	1	1

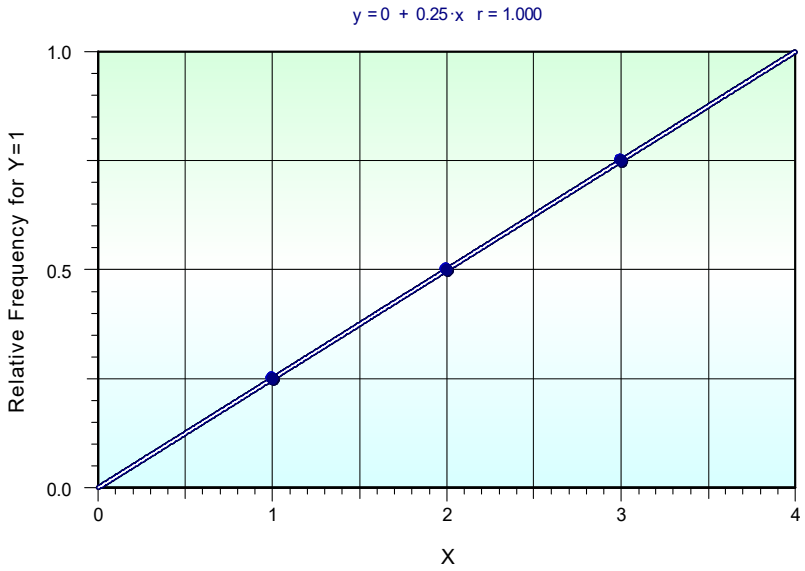
this one for the not satisfactory following regression leads (straight line approximation):



It makes more sense to represent the probabilities here instead of the direct representation of the target value that a "condition" enters. One almost combines x areas to come "onto countable events" to this (classification). The table then becomes:

x (original)	0,5	0,7	1	1,3	1,5	2	2,2	2,3	2,7	3	3,2	3,4
x- groupe (class)	1,0				2,0				3,0			
y	0	0	1	0	0	1	1	0	1	0	1	1
n_i = count (y=1)	1				2				3			
count/groupe size	1/4 = 0,25				2/4 = 0,5				3/4 = 0,75			

The x values are assigned to the groups of 1, 2 and 3 (according to a centric classification, here on integer numbers). The number is y = 1 counted (how it is "good" and "bad" at concepts to fix on what counting refers e.g., open "badly") within these groups now. From this the relative frequencies can be calculated per group. If one represents these, then a substantially better relation arises:



This is bought by a diminution of the x information, i.e., for this evaluation considerably more observations are used than at steady measurands. Originally these makes 12 information's in the previous example stand, 3 at the disposal only what is a corresponding disadvantage. Under circumstances too few degrees of freedom are entitled at the evaluation for the regulation of possible interactions at the disposal. Since it is pure observations here but usually (not around planned tests), however, sufficient data are as a rule also available.

Estimators are the formation of the relative frequencies for the probability P simultaneously, it becomes y =1. It is valid:

$$p_i = \frac{n_i}{n_{group}}$$

n_i = number of *y=1* (cannot be 0, usually n_{groupe} >= 5)

For n_i < 0 and n_i > 4 nonsensical probabilities of P < 0 and P > 1 give up, though. Therefore, suitable transformations are necessary. A transformation frequently used for this problem definition is the so-called Logit model:

$$y' = \ln\left(\frac{p}{1-p}\right)$$

respectively

$$b_o + b_1 x_1 + ...b_z x_z = \ln\left(\frac{p}{1-p}\right)$$

The expression $P/(1\text{-}P)$ represents *odds* and the meaning has admission probability/counter-probability. One also speaks here about Logits. A little strange, it is the dealing with

odds and the interpretation one is horse bets, then, because the odds correspond to the quotas here. It is important to notice that the logistical regression treats not probabilities but probability conditions.

To remove the low limit of the domain in addition, the *odds* become in addition logarithm. The inverse function is needed for the inverse function after the regulation of the model parameters on probabilities also here:

$$\hat{p} = \frac{1}{1 + e^{-\hat{y}}}$$

This also is described as a "logistical" distribution function. The limits $P = 0$ and $P = 1$ about the Logit are not portrayable. The number ni per group should not be 0 anyway.
With steady target values the prerequisite for the method of the smallest error squares for the estimate of the sought-after coefficients b is that the error deviations have an identical variance at the regression. This is not the case here. Therefore, a weighted regression must be used. To this an estimator is needed for the variance. The already established relation became a determination of the coefficients at <u>not</u> weighted regression till now

$$\hat{B} = \left(X^T X\right)^{-1} X^T Y$$

used. At the logistical regression there is the problem that the variances of the model errors are not constant. Through this the variances of the model estimators cannot be minimized about the method of the smallest error squares. However, the problem can be removed by a weighted regression. Be the variances of every observation needed to this, these through

$$\hat{s}_i^2 = \hat{p}_i (1 - \hat{p}_i)$$

you define. The estimators for the regression coefficients then determine themselves through:

$$\hat{B} = \left(X^T \delta X\right)^{-1} X^T \delta Y'$$

with
$\delta = diag\,(s_1^2, s_2^2 \ldots\ldots, s_n^2)$

Y' is the vector of the corresponding Logits. A new problem arises, however. The estimators determine itself only from the result of the calculation. So, an iterative calculation must be carried out.

Another possibility for the regulation of the model parameters is the maximum Likelihood, short ml method. The basic concept is relatively simple. The parameters are chosen so that the valued variables are the most similar to the observations in the data set (Likelihood). The similarity is, described by the so-called Likelihood function this one the Likelihoods of all cases of the data set consists of the product:

$$LH = \prod_{i=1}^{n} \hat{p}_i^{y_i} \cdot (1 - \hat{p}_i)^{1-y_i}$$

Yi dates from the n observations, $\hat{p}_i$

from the model. The coefficients of the model are to search so now that LH gets maximum. It is like a probability, can accept a value between 0 and 1 since a little similar to the likelihood of a single case. Likelihood the product of many numbers between 0 and 1 gets minute, though, therefore becomes also here LH is logarithmized and is made it this LL lied short:

$$\ln(LH) = LL = \sum_{i=1}^{n} y_i \cdot \ln(\hat{p}_i) + (1 - y_i) \cdot \ln(1 - \hat{p}_i)$$

There is no analytical solution for the two variants. The coefficients also must be determined iteratively in which at first one chooses an arbitrary start value. With these the Logits and the first estimated values of the probabilities $\hat{p}_i$ can being certainly. The product of the LH function or the sum is charged to *t*he LL with that for every data series. The same must repeated as long, as no greater LL-value can be found
It is the most important advantage of the maximum likelihood method, that for the regulation of the coefficients no group formation of the data is required (can contain 0 events, where Logits are not calculable).

A dimension for the quality of the found solution is the deviation:

$$D = -2\,LL$$

The omen is changed since the logarithmic value between 0 and 1 is always negative. One gets in addition one (χ^2-distributed value which means how badly the model describes the data through this with the factor 2. Therefore, it is all the better the smaller this value is.
At the normal multiple regression, the certainty measure R^2 is primarily indicated for the quality of the model. There is no direct correspondence, however, a pseudo-R^2 was defined by McFadden:

$$R_{MF}^2 = \frac{LL_0 - LL_1}{LL_0} = 1 - \frac{LL_1}{LL_0}$$

LLo : Log-Likelihood of the model, this is only the constant $y' = b_o$
LL1 : Log-Likelihood of the concrete model $y' = b_{o+}\, b_1 x_1 +$

cannot reach the value 1, as a rule, values from 0,2 to 0,4 are already regarded as a good model customization.

For the assessment of the significance of the individual coefficients (factors) the deviation test is recommended. It is checked whether the model shows with the respective factor compared with this without just this significant difference. For the check of a factor the difference of the deviation is formed:

$$\Delta D_F = -2\,LL_1 - (-2\,LL_F) = -2(LL_1 - LL_F)$$

in which the index $_F$ stands for the model without the factor to be looked at compared with the exit model with the index $_1$ (see pseudo-R^2).
With the χ^2-distribution as well and the *degrees of freedom df* = 1, the p-Value = 1 - α can be

determined.
The complete model also can analogously be tested (see pseudo-R^2 in turn) compared with the "zero model" to this. The difference deviance is:

$$\Delta D_G = -2\left(LL_1 - LL_0\right)$$

df = with the degree of freedom: z = number of factors, interactions etc.

Relatively to the power of computation effort means one this approach, however, because the ml iteration must be carried out for every factor to be checked. As an alternative to it the woods test frequently mentioned can be used. This is like the t test at the normal regression. The test quantity is for every factor:

$$\chi_j^2 = \left(\frac{b_j}{s_{b,j}}\right)^2$$

with

$$s_{b,j} = \sqrt{X^*_{j,j}}$$

and

$$X^* = \left(X^T \delta X\right)^{-1}$$

the already established diagonal matrix was and in which (from the variances of every observation series.

Poisson Regression

The Poisson distribution describes countable events that occur in a defined time interval. This can be, for example the number of error messages within a week. Poisson regression is treated in connection with the so-called generalized linear regression model (GLM).

The goal is to determine a model for countable events based on the Poisson distribution, which is well suited for this purpose.

The Poisson density function (probability that the number of errors y occurs) is defined as:

$$g(x) = \frac{\lambda^y}{y!}\, e^{-\lambda}$$

y : number of events or errors integer
λ : Poisson- or Error rate

Normally, the number of errors is declared as x. Since in the following x is required for the influence parameters, y should be used here. As with lifetime models, the relationship for the Poisson rate can be best represented logarithmically:

$$\ln(\lambda) = b_o + b_1 x_1 + b_2 x_2 + \cdots + b_z x_z$$

b : coefficients of the model
x : influence parameters

therefore, it is:

$$\lambda = e^{b_o + b_1 x_1 + b_2 x_2 + \cdots + b_z x_z}$$

Assuming that the observed events are y_i Poisson-distributed, the expected value is $E[y_i] = \lambda_i$. Each observation must be independent of each other and random. The best estimator for the coefficients b is given if the product of all probabilities (likelihood ⇨ L) of each i-th observation is a maximum. The following applies to this:

$$L(b_o, b_1, b_2, \dots b_z) = \prod_{i=1}^{n} \frac{{\lambda_i}^{y_i}}{y_i!} e^{-\lambda_i}$$

The determination of the coefficients cannot be done analytically, as in multiple regression, because y is both in the exponent and in the denominator with faculty. This makes it clear that the result must differ from the method of the least square estimation.

However, the differences are small and the coefficients determined herewith can be used as a starting condition of the iterative calculation. The so-called Newton-Raphson method is most often mentioned for this iterative calculation. The logarithmic form can also be used to search for the maximum probabilities, resulting in the so-called log likelihood function LL:

$$LL(b_o, b_1, b_2, \dots b_z) = \sum_{i=1}^{n} y_i \ln \lambda_i - \lambda_i - \ln(y_i!)$$

The goal is to find the coefficients where the sum of the right side is a maximum. Because of this the method is well known as Maximum Likelihood Estimation MLE for short. Depending on the cancellation criterion or the number of selected iteration steps, slightly different results can arise!

The dispersion of the coefficients sb is determined by:

$$s_{b,i} = diag(\sqrt{(X^T W)^{-1}})$$

with the weight

$$w_{j,i} = x_{j,i} \cdot \mu_i \quad \text{and} \quad \mu_i = \exp(b_o + b_1 x_{1,i} + b_2 x_{2,i} + \cdots)$$

To determine whether the coefficients are significant, the p-value is obtained by

$p\text{-}value$ = 2·*DistributionNormal*(-|z|)

with

$$z = \frac{b_i}{s_{b,i}}$$

The parameter is regarded as significant if $p\text{-}value$ < 0,05, see *chapter Statistical hypothesis tests*.

To assess the overall model, the sum of the deviation squares is calculated as a so-called deviance. The model scattering is:

$$D_{(b)} = 2 \sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{e^{Xb}}\right) - (y_i - e^{Xb}) \qquad \text{with } Xb = b_o + b_1 x_{1,i} + b_2 x_{2,i} + \cdots$$

and where only the constant b_o is in the model:

$$D_{(bo)} = 2 \sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{e^{bo}}\right) - (y_i - e^{bo})$$

which results in the so-called coefficient of determination with:

$$R^2 = 1 - \frac{D_{(b)}}{D_{(bo)}}$$

Instead of the adjusted R², like at the multiple regression, here a corrected R^2_{cor} is used:

$$R^2_{kor} = R^2 - \frac{z}{D_{(bo)}}$$

Poisson-Regression with Intercept

If there are no events for certain combinations of the influencing parameters, there is a problem that ln(yi=0) is not possible. Therefore, an offset should be used, which is called intercept here and which can be set individually for each observation. It is recommended to set this intercept only for the case of yi=0 to 0.01.

Normalized Model

As with multiple regression, it is recommended that the influence parameters x are normalized:

$$x_{norm} = \frac{(x - \bar{x})}{x_{max} - x_{min}}$$

This makes the range of values between -1 .. +1 and the determination of the model becomes more stable, the significance becomes clearer. In contrast to Visual-XSel, the many statistical programs use the original value ranges as the default setting, so the coefficients differ and the models may be different.

Further characteristics

Another very frequently mentioned characteristics for likelihood-based models and thus also for Poisson regression is the Akaike Information Criterion, AIC for short, named after the Japanese Hirotsugu Akaike AIC

$$AIC = -2\,LL + 2(z + 1)$$

The smaller AIC is, the better, but AIC alone cannot be interpreted, because there is no universal limit value for a best value. AIC is therefore more used for comparisons between different models. As with the R², models with more terms, such as interaction terms, usually provide better AIC values, although z increases and is received with a factor of 2. However, these terms may only remain in the model if they are also significant (see p-value)

Another criterion is the so-called Bayesian Information Criterion, BIC for short (named after the English statistician Thomas Bayes). This key figure is very similar to the AIC and also takes into account the number of measures n:

$$BIC = -2\,LL + (z + 1) \cdot \ln(n)$$

From n > 7, BIC becomes larger than AIC and punishes more complex models more severely. However, the influence of n should not be understood in such a way that the smallest possible data sets should be used. More information is basically an advantage for modeling. Although AIC and BIC are often mentioned as important characteristics, the practical benefits seem limited due to the disadvantages mentioned.

Confidence intervals

The χ^2 distribution can be used for the confidence intervals of the model predictions. With a usual confidence range of 90%, the probability of error is α = 10%. The two-sided trust scope for the number (error) events y is:

$$\frac{1}{2}\,\chi^2_{\frac{\alpha}{2},2y} \;\leq\; y \;\leq\; \frac{1}{2}\,\chi^2_{1-\frac{\alpha}{2},2(y+1)}$$

Example wastage in a manufacturing process

In an experimental design, the wastage of components was counted as a function of temperature, pressure, time and particle grain size in the material.

The model yields the following coefficients

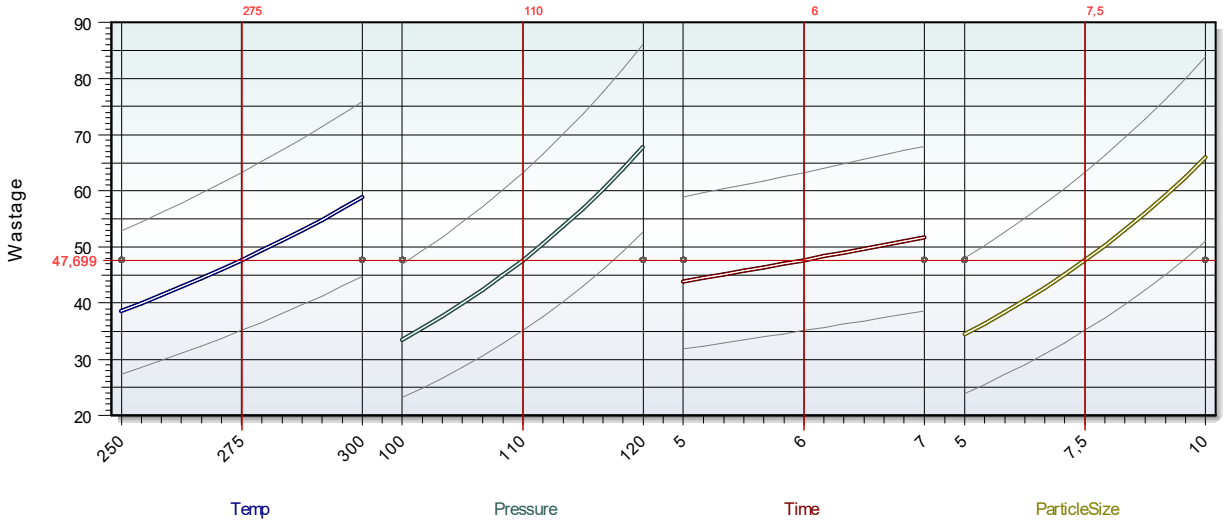
	Coefficient	p-value
Constant	-3,79657	
Temp	0,008435	0.000
Pressure	0,035221	0.000
Time	0,082689	0,015
ParticleSize	0,129531	0.000

R^2 = 0.926	DF = 11	AIC = 119.3
R^2_{kor} = 0.910		BIC = 123.1

All parameters are significant (*p-value* < 0,05).

No	Temp[C]	Pressure[t	Time[min]	ParticleSiz	Wastage
1	250	100	5	5	7
2	250	100	5	10	36
3	250	100	7	5	14
4	250	100	7	10	45
5	250	120	5	5	39
6	250	120	5	10	74
7	250	120	7	5	48
8	250	120	7	10	84
9	300	100	5	5	26
10	300	100	5	10	59
11	300	100	7	5	34
12	300	100	7	10	69
13	300	120	5	5	62
14	300	120	5	10	99
15	300	120	7	5	71
16	300	120	7	10	109

The following curve diagram shows the relationships as a continuous function, the results of which are to be rounded up or down.

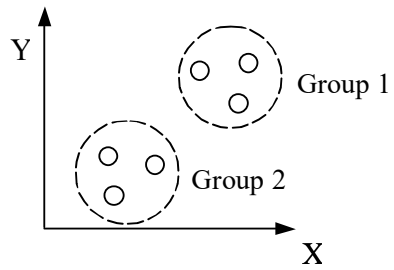


The confidence ranges are relatively strongly asymmetrical, see chapter Confidence intervals.

5. Multivariate Analyses

Cluster Analysis

One understands essentially a grouping of unordered data (e.g., measurements, image dots etc.) by a cluster analysis. For example:



The grouping is made by similarity characteristic. As a rule, these are distance data as the represented picture shows. In this case there is a high similarity if the data points have a distance as low as possible to each other.
d = degree of heterogeneous = measures for the assessment of the distances between the objects

Euclid's distance :

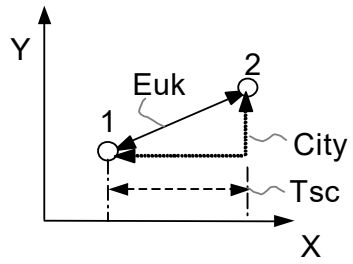
$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

City-block Distance :

$$d = |x_2 - x_1| + |y_2 - y_1|$$

Tschebyscheff distance :

$$d = \max(|x_2 - x_1|; |y_2 - y_1|)$$



There can exist similarities also in form of a correlation matrix. The higher the correlation is, the more similar the "objects" are to each other. So, a greater value is relevant here. There doesn't exist the initial data in the form of coordinates but there is a matrix where is shown a relation from each object to each other. The measurement to this is described by the correlation coefficient r. In this case the object distance is $d=1-r$ because the objects more nearly, the higher the correlation is. As an alternative to this often $d=ArcCos\ (r)$ is used. Respectively higher distances caused through this equation. The similarities cannot be related by data in

rows but with the titles and the data columns. Therefore, here has to be created first a correlation matrix before the cluster analysis.

The targets of building clusters are:

- Creating a simplified more open structure
- Data reduction
- Recognizing of connections

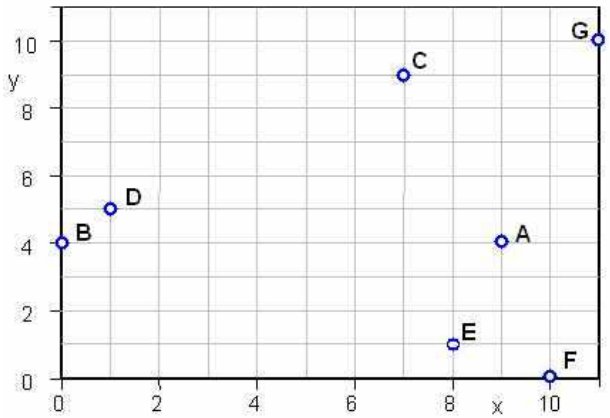
In Visual-XSel there is implemented the hierarchical agglomerative method.

The advantages are:

- No specification regarding number of clusters necessary
- Additional reduction of the clusters by "limit distance" possible
- Every run yields the same result
- Efficient algorithm to be implemented easily
- Graphic representation option of the clusters as a tree structure

The method shall be clarified at a simple example. The following objects are given with their coordinates:

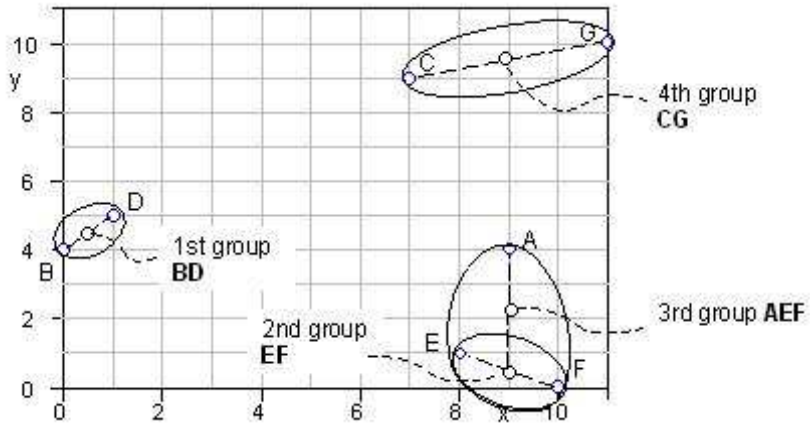
	x	y
A	9	4
B	0	4
C	7	9
D	1	5
E	8	1
F	10	0
G	11	10



Only 2 coordinates are represented here. n dimensional coordinates (columns) are possible where 3 coordinates can be visualized in a 3D-diagram.
The distance matrix arises from the coordinates. (Values = Euclid's distances):

	A	**B**	**C**	**D**	**E**	**F**	**G**
A		9,0	5,4	8,1	3,2	4,1	6,3
B	9,0		8,6	1,4	8,5	10,8	12,5
C	5,4	8,6		7,2	8,1	9,5	4,1
D	8,1	1,4	7,2		8,1	10,3	11,2
E	3,2	8,5	8,1	8,1		2,2	9,5
F	4,1	10,8	9,5	10,3	2,2		10,0
G	6,3	12,5	4,1	11,2	9,5	10,0	

The first cluster (object pair) is carried out via the smallest distance. This is between B and D with the distance of 1.4. Between these points there will be created a new center with the name BD.



The coordinates of the new group are calculated by $X_{BD} = 1/2\ (X_B+X_D)$. $Y_{BD} = 1/2\ (Y_B+Y_D)$. Correspondingly applies to the next group $X_{AEF} = 1/3\ (X_A+X_E+ X_F)$… If there exists, however, only a distance matrix, then the cluster center can be determined also about the following geometric relation:

$$d = \frac{\sqrt{2 \cdot \left(d_{AE}^2 + b_{AF}^2\right) - d_{EF}^2}}{2}$$

The results of both variants, however, do not yield exactly the same because the geometric center is calculated is here only an approximation method.
The distance of E and F amounts to 2.2 and therefore represents the 2nd group. The 3rd group is already a combination of 3 points AEF. After every run the complete table must be built up newly.
At the first summary the partner B will be deleted (values in column and line). Instead of D it will be set BD with the new distances to the remaining objects calculated with the given formula (bold values). It's better to define here BD and not DB

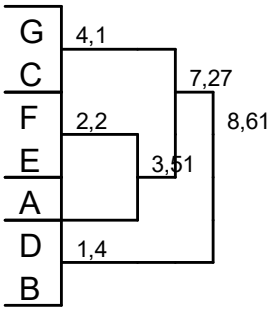
	A	**B**	**C**	**BD**	**E**	**F**	**G**
A			5,4	8,1	3,2	4,1	6,3
B							
C	5,4			7,2	8,1	9,5	4,1
BD	**8,1**		**7,2**		8,1	10,3	11,2
E	3,2		8,1	**8,1**		2,2	9,5
F	4,1		9,5	**10,3**	2,2		10,1
G	6,3		4,1	**11,2**	9,5	10,1	

The table goes down always further until 2 partners are only left. The individual steps can be clarified as a tree structure, also called **dendrogram**

The distances of the groups get longer from left to right. At the end of this algorithm, the last group will include all combinations.
Instead of the dendrogram one can have a structure list

	3,51	7,27
G C	G C	G C
F E	F E	F E
A	A	A
D B	D B	D B
4	3	2

Through direct specification or definition of the distance a desired number of clusters can be achieved. The last two summaries are not carried out.



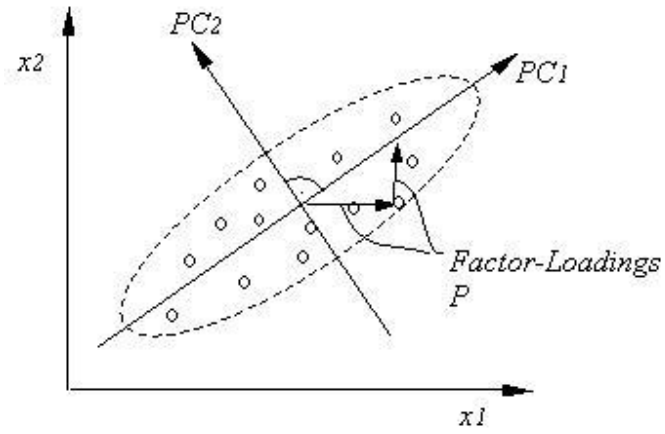
Categorial characteristics cannot be defined directly. It is necessary to transform the basis data in a numerical format first. This can be done by producing columns with worth of 1 and 0 to describe the expressions. For example, y can be transformed into the following numeric format:

Basis data		
	x	y
A	9	a
B	0	b
C	7	c
D	1	a
E	8	b
F	10	c
G	11	a

New structure				
	x	ya	yb	yc
A	9	1	0	0
B	0	0	1	0
C	7	0	0	1
D	1	1	0	0
E	8	0	1	0
F	10	0	0	1
G	11	1	0	0

Principal Component Analysis PCA

The Principal Component Analysis calculates new so-called latent variables. These are shortened called factors and represent the **Principal Components** PC. Do not mix up this name with the factors by DoE. It is the target to describe all existing variables with few factors (data reduction). With the variables *x1* and *x2* and its measurement points the principle shall be described like shown on in the following picture.



The measurement points lie in an ellipse which location depends on the correlation between the variables. A new axis system arises by moving the zero point and turning the coordinate system. The first so-called main axle rejects in the direction of the greatest spread of the standardized values *of x1* and *x2*. The second main axle stands vertically on the first one and explains the lower share of the variance. Therefore, one also describes the principal components as eigenvectors.
For the determination of the principal components so-called factor loadings P and Score values T are defined. The factor loadings describe the situation of the PC to the original coordinate system of *x1* and *x2*. The dimension of the factor loadings is -> number of components x number of variable *x*. The Score values T describe the projections on the main axles for every point. The dimension of T is -> number of component x number of measurements. The connection is in matrix notation:

$$X = T\,P^T$$

The following condition applies to the factor loadings:

$$p_1^2 + p_2^2 + \ldots p_k^2 = 1$$

The Principal Components are calculated through the Score-values t_i and the eigenvalues λ_i

$$PC_i = t_i \Big/ \sqrt{\lambda_i}$$

The eigenvalue λ_i describes, how much of the total spread of all variables is declared through the factors. The eigenvalues also serve for the decision whether factors in the model can be kept or left out. If the eigenvalue is less or equal 1, it explains less or equal of the variance

of one variable. If this is the case the factor can be left out. eigenvalues and eigenvectors yield an independent structure of each other (orthogonal).

The eigenvalues cannot be calculated directly or analytically and must be iteratively determined (eigenvalue problem). For further details we must refer to the appropriate literature.

Example: Defined are the variables x_1, x_2 and x_3. Calculated are the factor F:

x1	*x2*	*x3*	*F*
1	3	4	-1,00
2	4	3	-0,70
3	1	1	1,00
4	2	2	0,70

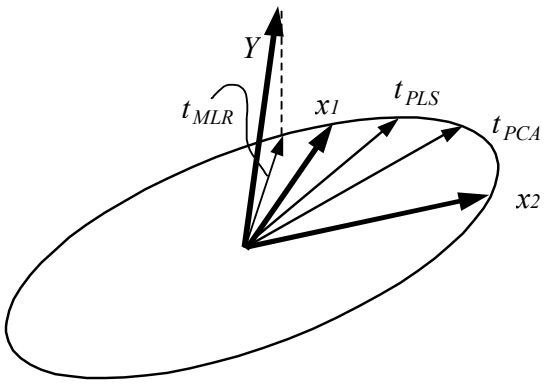
For these data a factor suffices (is λ for the second and third factor is under 1). There are also the so-called correlation loadings next to the factor loadings. These are the correlations between the factors and the original variables. If one looks at the correlations to each other, then it can be shown that the new factors correlate more highly with all exiting variables. It is just the target of the factor to reach a "description" as good as possible of all variables together.

x3: *F*	-0,958
x2: *F*	-0,881
x1: *F*	0,881
x1: *x3*	-0,800
x2: *x3*	0,800
x1: *x2*	0,600

It has to be taken into account here for the interpretation that the factor correlates with *x1* positively and with variable *x2* and *x3* negatively. A negative correlation means that the direction has turned.

Partial Least Square (PLS)

PLS was developed 1960 of the Swedish econometrist Herman Wold. PLS means: Partial Least Squares Modeling " into latent variable " . The purpose is primarily the evaluation of correlating data or the evaluation of mixture plans, where the standard method Multiple Linear Regression (MLR) isn't practicable. It is also an essential advantage of PLS that much variables can be processed. It is even possible to evaluate with less information (data rows) than variables exist. This is not possible with MLR.
The represented picture shows two variables x_1 and x_2. The main component analysis PCA with t_{PCA} lies in the "bump" of the ellipse. The greater x_1 and x_2 correlate, the longer t_{PCA} gets. If there is no correlation, the vector direction is not defined by t_{PCA} any more, because the ellipse then becomes a circle and has no more preferred direction.



The component of t_{PLS} however is then still determinable about the analysis of the covariance. This is a decisive advantage of PLS over PCA. The results, i.e., the coefficients of the variables, are then identical with the MLR method (for orthogonal data). While the MLR method provides no longer clear results or completely gets out at very correlating data, furthermore the PLS method can be used. Even if two variables have a correlation or 100%, this is still possible. Of course, the assignment of the effects is then no longer clear, in this case PLS shares the effects half to the two variables.
It is the disadvantage of the PLS method that the forecasts and R^2 are worse than at MLR. The coefficients are partly also fundamentally smaller, what causes to estimate the effects too little.

PLS is very related with PCA. Instead of the loadings (PCA) here is the weight matrix W relevant

$$X = T W^T$$

T are the so-called Scores of the components. W includes the response y, which doesn't exist in PCA. Also, here the following condition applies to the weights:

$$w_1^2 + w_2^2 + ... w_k^2 = 1$$

The regression model is defined with:

$$\hat{y} = T c^T$$

where c is the regression coefficient.

The complete algorithm (NIPALS – *Nonlinear Iterative Partial Least Square*) is shown below:

$$w' = \frac{X^T y}{y^T y}$$ *weights absolute for the standardized matrix X*

$$w = w' / \sum w'^2$$ *standardized weights*

$$t = Xw$$ *score vector*

$$= \frac{\sum_{j=1}^{z} \mathrm{cov}(y, x_j) x_j}{\sum_{j=1}^{z} \mathrm{cov}(y, x_j)^2}$$ *with z = number of variables*

$$c = \frac{y^T t}{t^T t}$$ *regression coefficients between y and the components*

$$p = \frac{X^T t}{t^T t}$$ *loading-vector*

$$E = X - tp^T$$ *residual-matrix of variables*

$$f = y - tc^T$$ *residual-vector of the response*

The next components are determined by defining $X = E$ and y = f and recalculate at the beginning. Through adding more components often R^2 raises. If this is not the case, no other components are required. By using more components, it can happen that some coefficients are changed extreme. Then the model with the bigger number of components is relevant. Regarding the original variables x, the coefficients b can be calculated through:

$$b = W(P^T W)^{-1} c^T$$

Summarized characteristics:

- R^2_{PLS} is less then R^2_{MLR}
- Coefficients of PLS are less than MLR-> Errors have a less effect through this.
- PLS maximizes the covariance between the principal components and Y, MLR maximizes the correlation
- PLS is able to work with high correlations between the x variables.

PLS has got acceptance in the sectors of pharmaceutical, chemistry and spectroscopy as a standard. It is often used as a universal method for all evaluations. However, the multiple regression still has to be preferred for evaluations where the data is not too strongly correlating (e.g., from the design of experiments). The interpretation of the effects and the model is better here. At orthogonal data the coefficients of the regression models are also the same.

Estimation of the spread at PLS

In general, the spread of the coefficients b cannot be calculated for PLS via the trace of $\left(X^T X\right)^{-1}$ like by MLR-Method. If the correlation is great between the variables, the spread can be estimated only via a so-called cross validation. The disadvantage here is a not definite result and the calculation needs much computing time.

To calculate and applicate here the p-Value, like at MLR is not recommended here. For PLS and the variable selection there is much better suitable the so-called VIP-indicator

Variable selection with VIP

For using the PLS-Method and the variable selection here it is suitable to consider the VIP-indicator. VIP is an abbreviation of ***V**ariable **I**mportance in the **P**rojection*. That means how much is the influence of the variable in the projection of the scores t.
This indicator is first launched by Wold in 1993. VIP is calculated for each x_j via:

$$VIP_j = \sqrt{z \sum_{k=1}^{h} \left(\frac{y^T t_k}{t_k^T t_k} w_{jk}^2 \right) \Big/ \sum_{k=1}^{h} \left(\frac{y^T t_k}{t_k^T t_k} \right)}$$

with h = number of components,
z = number of variables x (e.g. terms)

The y-vector must be standardized first. In the literature there is described a limit for VIP between 0,8 …1. A too less value indicates, that the variable can be left out. But experiences have shown, that VIP<0.5 are not unusual for important variables

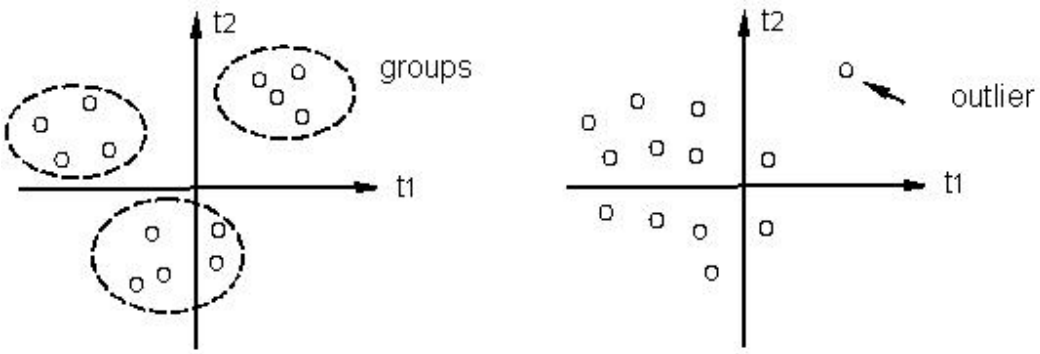
If there is the question whether a variable should be left out from the model, the coefficient size also has to be taken into account. Also, the technical connections should be considered.

PLS charts

Especially for evaluation of PLS-Analysis there are two important charts, the Score Plot and the Correlation Loading Plot. These charts can be selected under the rubric ***charts*** (after PLS data analysis via menu ***Statistics*** of the spreadsheet).

Score Plot

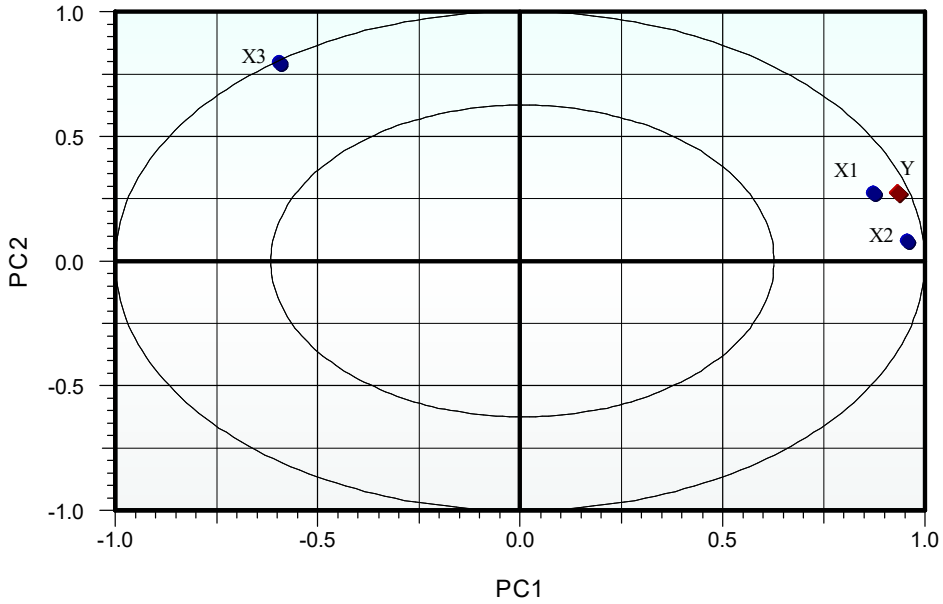
The Score plot represents every measurement point about the most important Scores t1 and t2. Possible samples and characteristics in common can be recognized. Also, outliers can be recognized.



Correlation Loading Plot

In the so-called Correlation Loading Plot the professed variances of the variables and the target value are represented indirectly on the components PC here.

The axis is scaled as correlations, so it is: professed variance = correlation^2.
Hereby the influences of the variables are shown and one recognizes which components describe the variables better. The ellipses describe 100% (outer) and 50% (inner) professed variance



The nearer the variables are to the 100% ellipse, the more important these are. In this example the component PC1 describes the variables x1, x2 and also the response y approximately alone, while the variable x3 needs both components.

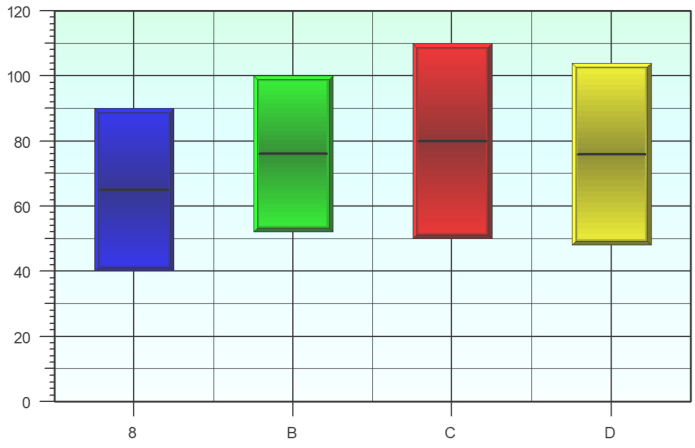
Further statistical charts

Scatter bars

In practice it frequently occurs that certain circumstances are illustrated with just one or a few measurements. If you ascertain that the result scatters more or less, in most instances the median is built. This is absolutely permitted, if the value staggers marginally after repeated measurements. But if there are larger variations, different test series are difficult to compare to each other, especially, if outliers do occur. Possibly you will get no unique compromise output. An illustration with Scatter bars will help in this case. Here an example:

A	B	C	D
90	100	110	104
50	64	52	65
70	75	72	84
40	52	50	48

In 4 test series the values respectively listed among each other have been quantified. After selection of the menu point *Statistic/Scatter bars* the following diagram results:



Please note that the titles of columns (legend) standing in the first row are used as X-axis title. The first column is also used as series and is not interpreted as reference to the x-axis like at the most other diagram types.
If the median and the Scatter s is just known from samples (resp. measurements) and a predication should be made about the totality (then infinitely many measurements should be executed), so a so-called confidence belt can be indicated, in which the true median lies with PA % probability.
If you choose maximal and minimal value in the dialog window Statistics/Scatter bars, just the maximal and minimal value of the entered data (sample) will be determined, as shown here in the example. If the number of samples would be increased, so another maximal or minimal value could be found. For this reason, it is recommended to choose one of the 3 confidence belts, which are available. Particularly here the outliers do affect not so seriously.

Activate the menu point *Options/Show data*, to type out the medians and the confidence

belts resp. the min/max-areas in the diagram. Those are also in the graphical data (table menu point *Insert/Graphic-data*).

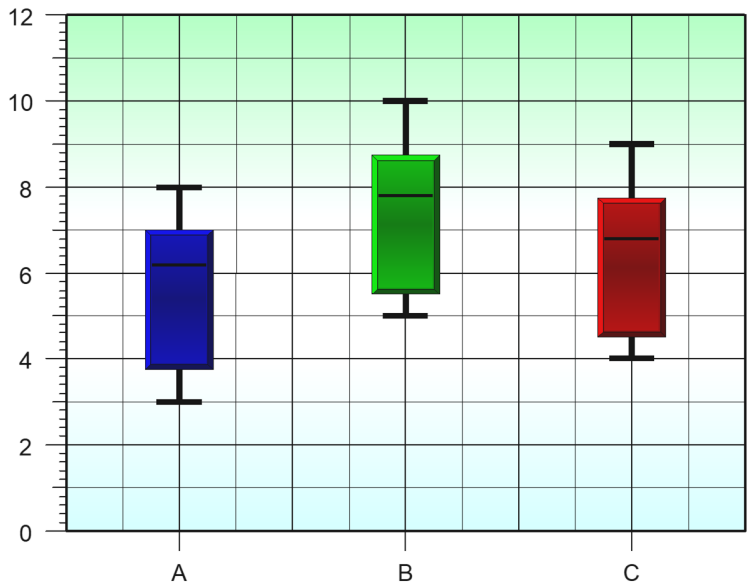
Boxplot

The boxplot is a special type of frequency scale. Here the values are depicted via the y-axis instead of the x-axis, whereas several boxplots in parallel are possible in one diagram. In the middle of the boxplot there is a line with the so-called center-value resp. median. Optional also the median can be chosen. Within the inner field there are 50% of all values. Within the outer margin lines top and down are 99% of all values. Optional also the smallest and largest occurring value can be displayed (min/max-values). If there are too little data values, the 99% areas correspond to those of the min/max-values.

In opposite to the frequency scale with Gauss curve here you get a rapid comparison about the respective status of several series.
The values of the respective series are written among each other. In the first row there is the reference to the X-axis resp. the legend for the single boxplots. An example for following table values:

A	B	C
3	5	4
6	7	6
7	8	7
7	9	8
8	10	9

After selection of diagram type boxplot vertical the following develops:



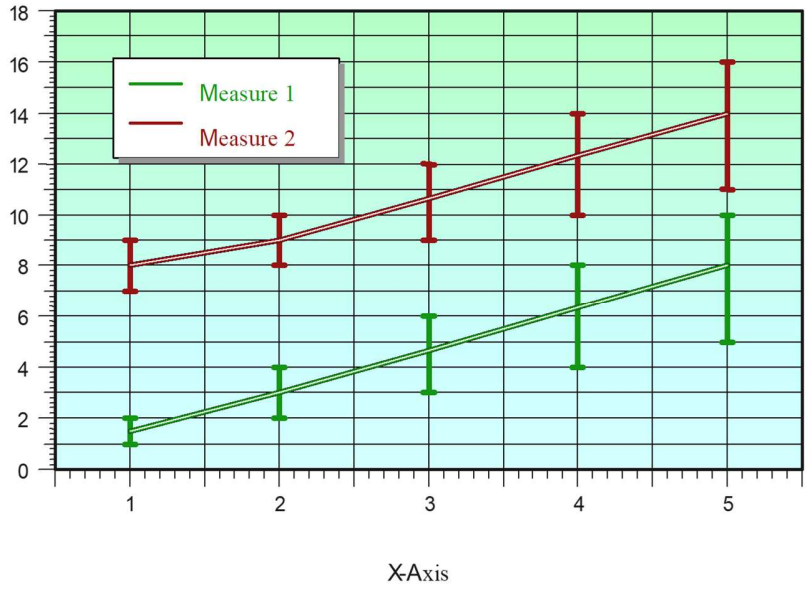
Optional the single values can be depicted as eye-catcher-points with their numerical values. See also Boxplot horizontal, Scatter bars

Median plot

In a diagram a median plot summarizes several columns of the table to a curve, which contain vertical narrow bars as min-/max values.
The curves depict the median of the cells, standing in one row. For designation of the summarized columns the legend is used, which can be found in the first row of the marked table area. The next area starts from the column of the next following legend, e.g., 2 groups with each 3 columns:

	Measure			Measure		
1	1	1.5	2	7	8	9
2	2	3	4	8	9	10
3	3	5	6	9	11	12
4	4	7	8	10	13	14
5	5	9	10	11	15	16

Those table data result this illustration as median plot:



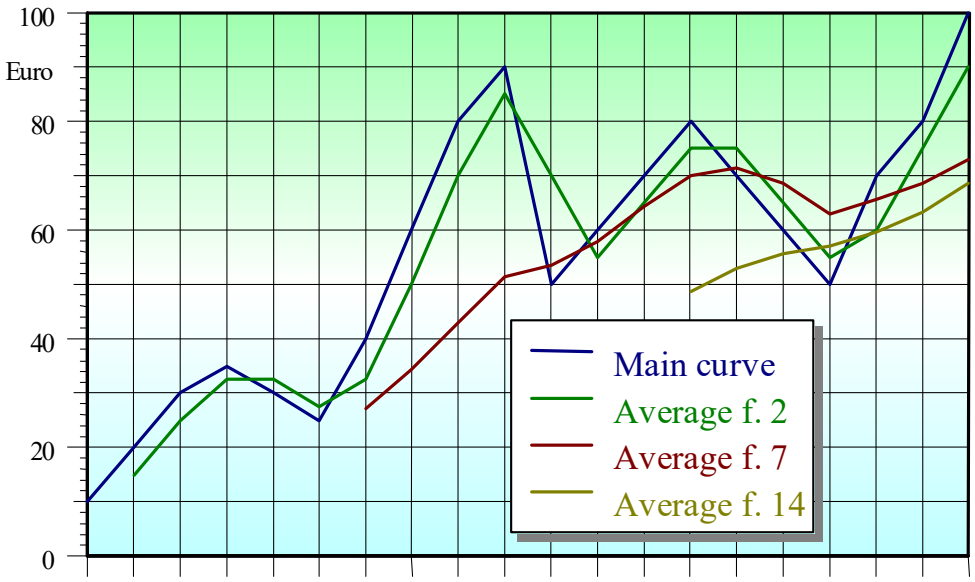
See also Group chart

Gliding average

The illustration of gliding average corresponds to the line diagram with at first one „main curve“. The reference to the X-axis stands in the first column of the table. The data of the main curve (Y-values) are in the proximate column.
Additional three other curves are created, which are built from the particular medians of the main curve. Thereby every point of these additional curves is composed from the median of several previous points of the main curve. How many points will be used for this, is fixed in the dialog window of the diagram types.
The data, developing in doing so stand in the following 3 columns, which have to be blank for this reason. Possibly existing cell-contents are overwritten. If e.g., there are the following values in the first two columns of the table:

	Main curve	Average from 2	Average from 7	Average from14
1.1.2000	10			
2.1.2000	20	15.0		
3.1.2000	30	25.0		
4.1.2000	35	32.5		
5.1.2000	30	32.5		
6.1.2000	25	27.5		
7.1.2000	40	32.5	27.1	
8.1.2000	60	50.0	34.3	
9.1.2000	80	70.0	42.9	
10.1.2000	90	85.0	51.4	
11.1.2000	50	70.0	53.6	
12.1.2000	60	55.0	57.9	
13.1.2000	70	65.0	64.3	
14.1.2000	80	75.0	70.0	48.6
15.1.2000	70	75.0	71.4	52.9
16.1.2000	60	65.0	68.6	55.7
17.1.2000	50	55.0	62.9	57.1
18.1.2000	70	60.0	65.7	59.6
19.1.2000	80	75.0	68.6	63.2
20.1.2000	100	90.0	72.9	68.6

If the median lines are ascertained from respectively 2, 7 and 14 last data points, so this diagram results, which e.g. can be used for stock quotation:

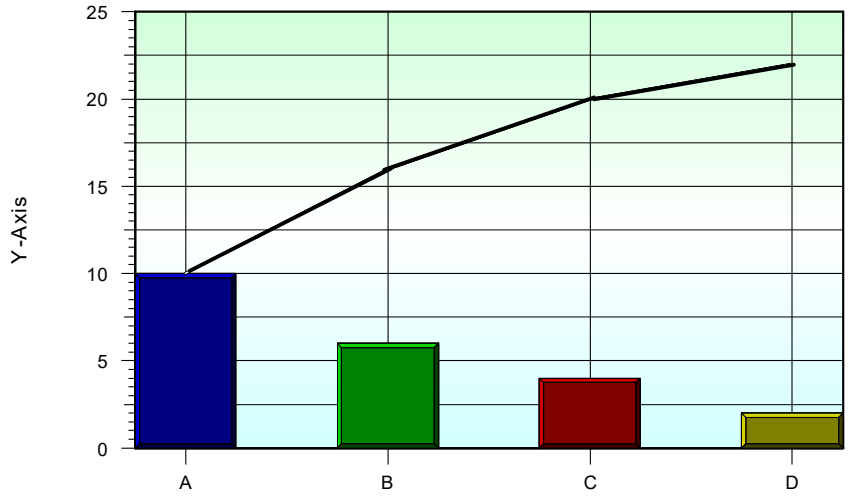


Pareto

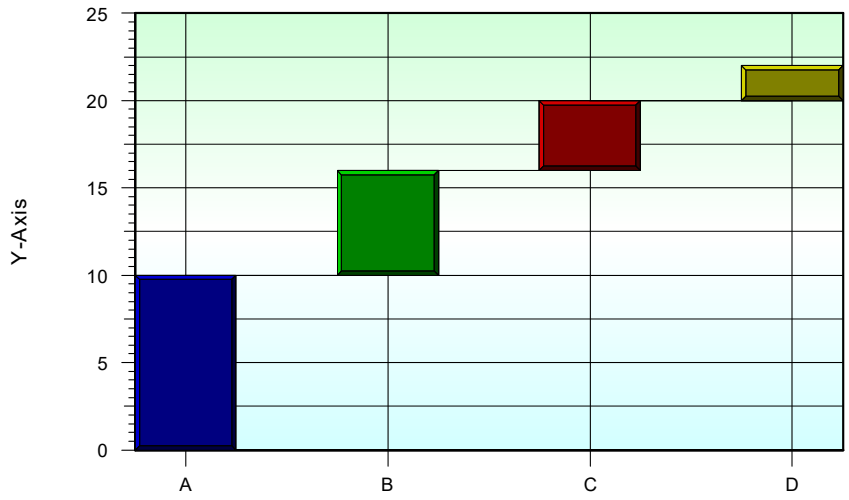
The Pareto diagram corresponds to a histogram, whereas the pursuant columns are depicted in turn, sorted according to the size. In addition the columns have different colours. The biggest value is at the beginning, the smallest at the end.

This diagram type is used e.g., to prefix the most important influence factors.

A particular form of the Pareto-chart has an additional sum-curve (cumulative values) over the bars:



The other variation is the bars represent the sum of the prior values



(Also possible in horizontal representing)

7. Capability indices

Capability indices describe the actual process as well as the achievement of a process to be expected in the future. In general, one understands by the capability index the relation of tolerance to the dispersion.
The reference is a range of ±3σ, respectively ±3s, where 99.73% are inside the specification.
In case of a production process, it concerns to the process capability index Cp.

For the consideration of a mean value displacement (divergence of the ideal process situation), the index Cpk is introduced which is normally worse than Cp (or equal in case of no displacement). As a rule, a process is capable, if $Cpk \geq 1{,}33$.

In following the relations are shown for different distribution forms:

Normal distribution

The normal distribution is to be applied if divergences to the nominal value are caused through random variations.

$$C_p = \frac{USL - LSL}{6s}$$

$$C_{pu} = \frac{\bar{x} - LSL}{3s} \quad C_{po} = \frac{USL - \bar{x}}{3s}$$

$$C_{pk} = Min\left(C_{pu}; C_{po}\right)$$

with
LSL : Lower Specification Limit
USL : Upper Specification Limit
$\bar{x}$: Mean

If the real mean and the standard deviation is known, so μ and σ must be used instead of $\bar{x}$ and s. Alternatively C_{pk} can be calculated with the following formula:

$$C_{pk} = C_p\left(1 - |z|\right)$$

and z is defined by:

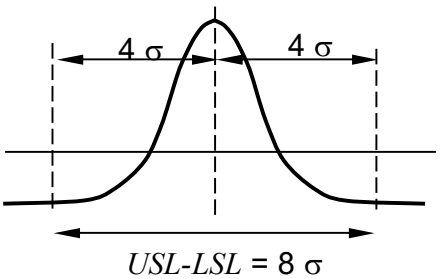
$$z = \frac{\bar{x} - (USL + LSL)/2}{(USL - LSL)/2}$$

for centric nominal value

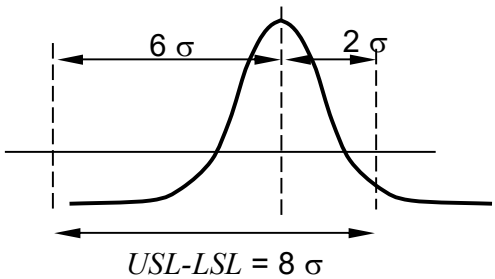
$$z = \frac{x_{soll} - \bar{x}}{(USL - LSL)/2}$$

for non-centric nominal value

Examples:



$C_p = 1{,}33 \quad C_{pu} = 1{,}33 \quad C_{po} = 1{,}33 \quad C_{pk} = 1{,}33$



$C_p = 1{,}33 \quad C_{pu} = 2{,}0 \quad C_{po} = 0{,}67 \quad C_{pk} = 0{,}67$

An application for this method is for example:

\Templates\9_Capability\ Process_Capability_Analysis_CpCpk.vxg

Lognormal-distribution

The log normal distribution is to be applied if the distribution is limited on the left unilaterally, only positive values are given and divergences to the nominal value are caused through random variations, which works multiplicative.

$$C_p = \frac{\ln(USL) - \ln(LSL)}{6\, s_{\log}}$$

$$C_{pu} = \frac{\bar{x}_{\log} - \ln(LSL)}{3\, s_{\log}} \qquad LSL > 0$$

$$C_{po} = \frac{\ln(USL) - \bar{x}_{\log}}{3\, s_{\log}}$$

$$C_{pk} = Min\left(C_{pu}; C_{po}\right)$$

$$\bar{x}_{\log} = \frac{1}{n}\left(\sum_{i=1}^{n} \ln(x_i)\right) \qquad s_{\log} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\ln(x_i) - \bar{x}_{\log}\right)^2}$$

If concrete values are not given, $\bar{x}_{\log}$ and $s_{\log}$ can be calculated with the following approximation formulas:

$$\bar{x}_{\log} \approx \ln(\bar{x}) - \frac{1}{2}\ln\left(1 + \frac{s^2}{\bar{x}^2}\right) \qquad s_{\log} \approx \ln\left(1 + \frac{s^2}{\bar{x}^2}\right)$$

An application for this method is for example:

\Templates\9_Capability\ Process_Capability_Analysis_CpCpk.vxg

Please note that all data has to be logarithmized including the limits. If the lower limit is 0, only the upper index C_{po} is valid.

Folded normal distribution 1st type

The folded normal distribution is to be applied if the distribution is limited on the left unilaterally and only positive values are given. The capability index is defined by a general equation:

$$C_{pk} = \frac{1}{3} u_{1-p}$$

p = relative fraction outside the upper specification limit, and u is the quantile of the standardized normal distribution

Instead of this alternatively the so-called percentile-method can be used, which is described in the next but one chapter. This is useful if there is only a less relative fraction p.

An application for this method is: *\Templates\9_Capability\ Process_ Capability_ Folded_Normal_Distribution.vxg*

The relative fraction outside p will be estimated via a Weibull-Distribution in the border area.

Folded normal distribution 2nd type (Rayleigh-distribution)

The type of distribution is used, e.g., for unbalance. Also here is valid the general formula:

$$C_{pk} = \frac{1}{3} u_{1-p} \quad \text{with} \quad p = e^{-\frac{\pi}{4}\left(\frac{USL}{\mu_r}\right)^2}$$

Non-parametric (distribution free) Percentile-method

For non-known distributions the so-called percentile-method is to be applied. In general, it is valid:

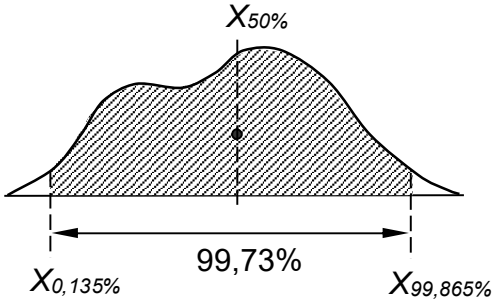
$$C_p = \frac{USL - LSL}{X_{99,865\%} - X_{0,135\%}}$$

For a normal distribution the denominator corresponds 6s. For a non-normal distribution, the relation area can be determined as it is described in the ISO / TR 12783.

Analogously for the normal distribution it is valid:

$$C_{pu} = \frac{X_{50\%} - LSL}{X_{50\%} - X_{0,135\%}} \quad \text{and} \quad C_{po} = \frac{USL - X_{50\%}}{X_{99,865\%} - X_{50\%}}$$

$$C_{pk} = Min\left(C_{pu}; C_{po}\right)$$



An application for this method is:

\Templates\9_Capability\ Process_Capability_non_normally_distributed.vxg

Distributions forms for several design characteristics

The following table shows an overview, for which design attribute which distribution has to be used:

Attribute	Symbol	Distribution
Linear measure		N
Straightness		F1
Planeness (flatness)		F1
Roundness		F1
Cylindric form		F1
Line shape		F1
Face form		F1
Roughness		F1
Unbalance		F2
Parallelism		F1
Rectangularity		F1
Angularity		F1
Position		F2
Concentricity		F2
Symmetry		F1
True running		F1/F2
Axial run-out		F1

N : Normal distribution
F1 : Folded normal 1st type
F2 : Folded normal 2nd type

Applications for capability studies:

Regarding the application one distinguishes:

- **Process Capability Study (PCS)**
- **Machine Capability Study (MCS)**
- **Measurement System Analysis (MSA)**

Process Capability Study PCS

PCS concerns to a longer time period. One use sample in fixed intervals and parts measures relevant quality characteristics of the product (min. 20 samples with n = 3-5). There must be considered influences of the machine, the material, the method, the operator and the surroundings. The process capability coefficients Cp and Cpk are used for the representation of the result. The calculation is described in the relations represented above.

Machine Capability Study MCS

MCS is carried out for a short period. Here basically the influence of the machine and the method will be analyzed. Influences of different materials, operators or environment terms are not considered and, hence, should be very steady. The formula is the same like for the Process capability study, but here the formula symbols are C_m and C_{mk}. The recommended sample size is 50, at least 20 parts. One also calls this a short time capability study. In general, this causes a higher demand in machine capability indices $C_m, C_{mk} \geq 1{,}67$.

Hint: The indices C_m, C_{mk} are no longer used in the current ISO norm. Instead of these P_p/P_{pk} or C_p/C_{pk} are used.

Measurement System Analysis with ANOVA

Measurement System Analysis investigations are the basic requirement for carrying out Capability Studies. They are intended to ensure that the used measuring equipment is suitable.
Note: With destructive tests (e.g., tensile or bend tests), a "substitute normal" must be used that is not destroyed (such as a thicker part, etc.). If there is force measurement of destroying test specimens, the test specimen can, for example, be replaced by a spring whose characteristic is in the test specimen's force/stroke range.

Procedures

Overall, a differentiation is made between the following influences:

1. **Repeatability** on a "**reference**" = constant master part
   (former process 1), pure test equipment deviation.
2. **Repeatability** on different **parts** (former process 3)
   Consideration of the value range to be measured.
3. **Reproducibility** on different parts and different **appraisers** (former process 2)
   Consideration of different appraisers.

According to VDA Volume 5 or the ISO 22514-7, measuring uncertainties are observed by means of the corresponding standard deviations that are expressed by the symbol *u* (measuring uncertainty budgets). The calculation is performed using an analysis of variance (ANOVA).
The overview below shows the most important measuring uncertainties:

Proportion	**Calculation**	**Description**
Resolution of the display	$u_{RE} = RE/\sqrt{12}$	RE Resolution of the equipment
Systematic deviation	$u_{Bi} = \lvert\bar{x}_{g-}\, x_m\rvert/\sqrt{3}$	$\bar{x}_g$ Disp. average val. of normal x_m Reference value of normal
Repeatability on normal	$u_{EVR} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_g)^2}$	x_i Meas. val. of the repetition i n Number of repetitions

From this, the influence of the instrument (MS = measuring system) is formed as an intermediate result (simplified representation considering no linearity deviation):

$$u_{MS} = \sqrt{u_{Cal}^2 + u_{BI}^2 + max\{u_{RE}^2; u_{EVR}^2\}}$$

The calibration uncertainty of the normal u_{Cal} should be considerably less than the total measuring uncertainty (recommendation of $u_{cal} \leq 0.15\ u_{MS}$). Refer to the calibration certificate for the calibration uncertainty.

$$\%Q_{MS} = 100\% \cdot \frac{k \cdot 2 \cdot u_{MS}}{TOL}$$

k = 2 VDA Standard for confidence level 95.45 %
k = 3 for confidence level 99.73 %, if the application requires, or the specialist department has corresponding normative stipulations, e.g., threaded fastener Technology.

Requirement: $\%Q_{MS} \leq 15\%$

This corresponds to the older requirement:

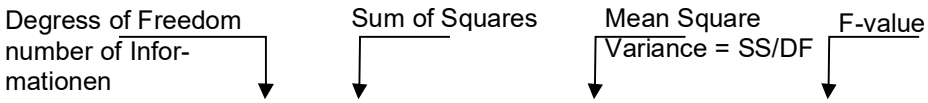
$$C_g = \frac{0{,}2 \cdot TOL}{2\ k\ s_g}\ ;\ C_{gk} = \frac{0{,}1 \cdot TOL - \lvert\bar{x}_g - x_m\rvert}{k\ s_g}$$

$\bar{x}_g$: mean of the measurements
x_m : mean of reference standard
s_g : standard deviation

In addition to the measurement uncertainties of the pure measuring system, influences from the part variation and the appraiser are also added. Overall, the measurement uncertainty of the entire measuring process is determined by:

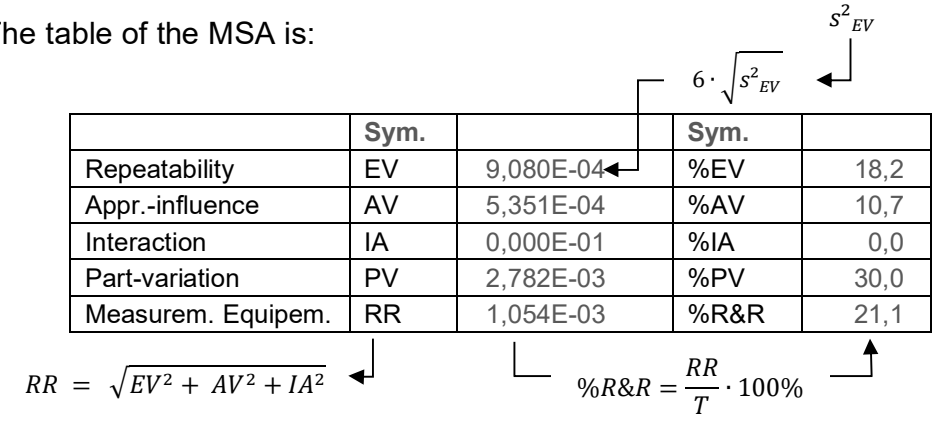
Measuring process = Measuring uncertainty of instrument +
Measuring uncertainty of equipment & appraiser

The effects are determined by ANOVA with a variance analysis (see also Chapter ANOVA). In this method the effects are a combination of parts-variation, the appraiser, and the interaction between these together. The biggest advantage of the ANOVA is the consideration of the interaction, which is why this method is preferable. To assess the effects separately, one divides the sum of the square-errors over all measurements in sub-totals and their variances. The classic representation in the Anglo-Saxon world is:



	DF	SS	MS	F
Part	9	1,181E-05	1,313E-06	71,7
Appraiser	2	3,640E-07	1,820E-07	9,9
Part*Appraiser (interact.)	18	3,293E-07	1,830E-08	0,7
Repeatability	30	7,700E-07	2,567E-08	
Total	59	1,328E-05		

The table of the MSA is:



	Sym.		Sym.	
Repeatability	EV	9,080E-04	%EV	18,2
Appr.-influence	AV	5,351E-04	%AV	10,7
Interaction	IA	0,000E-01	%IA	0,0
Part-variation	PV	2,782E-03	%PV	30,0
Measurem. Equipem.	RR	1,054E-03	%R&R	21,1

First of all, the sums of squares of the table data will be formed horizontally and vertically (Sum of Squares). With the help of the degrees of freedom DF the variance can be determined (Mean Square) and the standard-deviation of the set. The results are in each case multiplicated with the factor 6 the standard-deviations, which means that 99.73% of the parts are included. Via the F-value, which is the ratio of the sum of variances of the appraiser to the repetitions the significances can be determined (which results mostly in the p-value).

In the example one has to consider that the results are different by calculation with interaction compared.

The Visual-XSel template for this method is:
\Templates\9_Capability\ Measurement_System_Analysis_ANOVA+VDA5.vxg

The scope of the equipment and the appraiser is:

Portion	Calculation	Description
Repeatability of test object	$u_{EVO} = \sqrt{MS_{EV}}$	MS_{EV} Variance repeatability
Reproducibility of appraiser	$u_{AV} = \sqrt{MS_{AV}}$	MS_{AV} Variance of appraiser
Interaction	$u_{IA} = \sqrt{MS_{IA}}$	MS_{IA} Variance Interaction

Overall, the measuring process is determined by (simplified view):

$$u_{MP} = \sqrt{u_{Cal}^2 + u_{BI}^2 + max\{u_{RE}^2;\ u_{EVR}^2;\ u_{EVO}^2\} + u_{AV}^2 + u_{IA}^2}$$

In a similar way to the repetition and comparability precision $\%R\&R$ reference is made to the tolerance and it yields the key figure:

$$\%Q_{MP} = 100\% \cdot \frac{k \cdot 2 \cdot u_{MP}}{USL - LSL}$$

The requirement is: $\%\boldsymbol{Q_{MP}} \leq$ 20 .. 30% (depends on company requirements)

Example

VDA 5 / ISO 22514-7

		values / 1000
Resolution of gauge	u RE	0,0289
Repeatability Master	u EVR	0,0738
Standard uncertainty (Bias)	u BI	0,0058
Repeatability test-object	u EVO	0,1513
Repeatability appraiser	u AV	0,0892
Interaction	u IA	0,0000
Uncertainty measurement	u MS	0,0740
Uncertainty process	u MP	0,1758

Measurement	%Q MS (95,45%)	**14,8**
Process	%Q MP (95,45%)	**35,2**

In this example, the requirement was met with $\%Q_{MS} \leq 15\%$, but the uncertainties from repeatability at different parts and appraiser are too high ⇨ $\%Q_{MP}$ = 35.2% > 20%. The reason can be in an incorrect measuring range, which cannot cover the variation of the parts. The appraisers should be re-instructed ("operational definition"), so that all proceed in the same way.

Overview of the methods

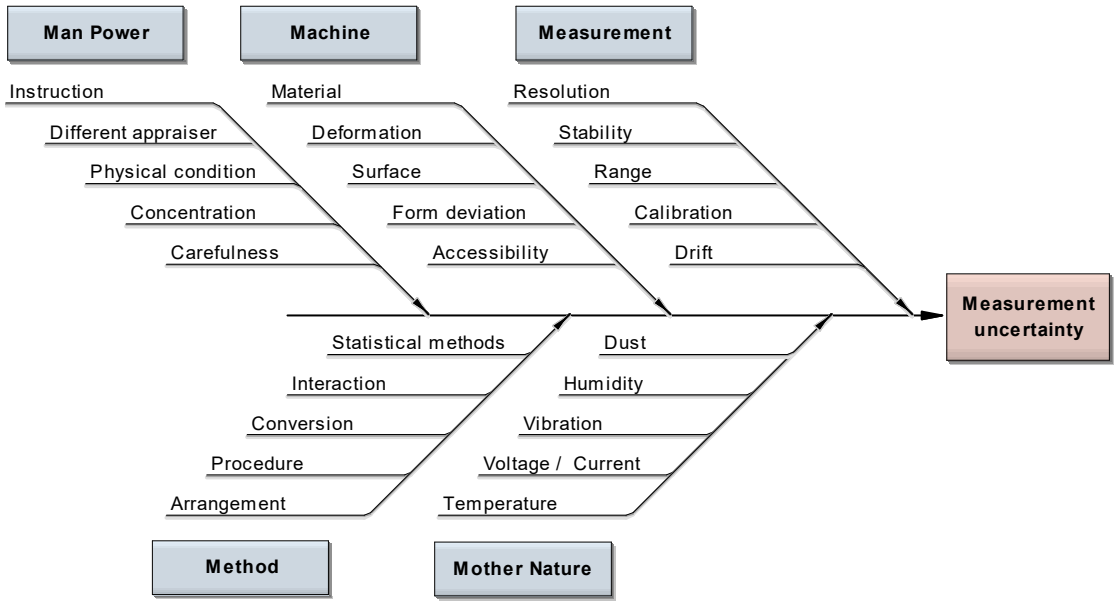
The overview below is a comparison with the old methods mentioned above:

	Influences			
	Repeatability		**Repeatability** Part to part	**Reproducibility** Part to part & Appraiser
	Testing with reference standard		Testing with various parts	Testing with various parts and various appraiser
VDA Volume 5 (ISO 22514-7)	u_{EVR}, u_{BI}	u_{RE}, u_{cal}, u_{lin}	u_{EVO}	u_{EVO}, u_{AV}, (u_{IA})
Requirement	Q_{MS} ≤ 15 %		Q_{MP} ≤ 20 %	Q_{MP} ≤ 20 %
Former classic methods	Method 1 C_g/C_{gk} ≥ 1,33		Method 3 range $\%R\&R$ ≤ 20 %	Method 2 range, mean value difference $\%R\&R$ ≤ 20 %

Other influences on measurement uncertainties

Along with the proportions of measurement uncertainties described above, there is a series of other possible influences such as stability and temperature.

$$u_{MP} = \sqrt{\ldots + u_{Lin}^2 + u_T^2 + \ldots}$$



Here too, as regards calculation further measurement uncertainties $u_{\text{influence}}$ are cumulative in accordance with the Gaussian law of error propagation.

$$u_{MP} = \sqrt{\ldots + u_{E1}^2 + u_{E2}^2 + u_{E3}^2 \ldots}$$

Especially measuring equipment holding devices and their possible deformation may have considerable influence on measurement uncertainties, see example mentioned in Ishikawa diagram. These should be quantified by tests as far as possible. If this is not possible, the percentage shares shall be considered e.g., by rigidity calculations. Furthermore, manufacturers' specifications shall be considered, e.g., in case of electronic measurement sensors.

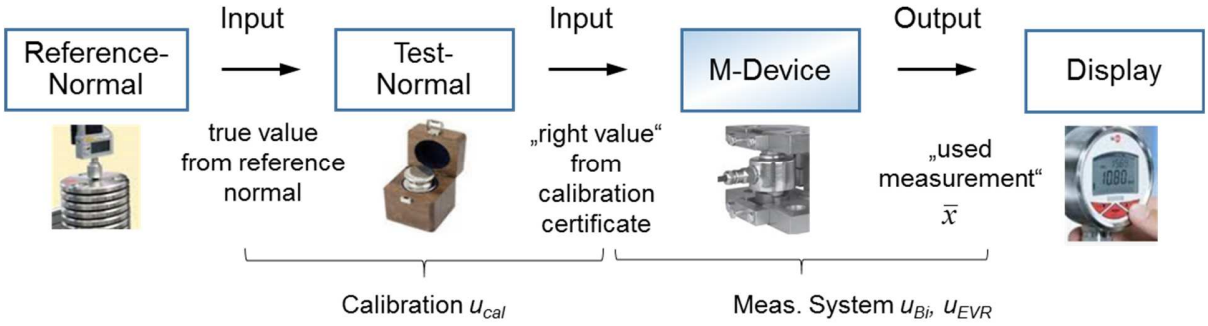
Reducing the measuring uncertainty by repetitions

In the event that the requirement is not met but no alternative measuring equipment is available, the possibility of repetitions exists. By multiple repeat measurements and averaging, it is possible to achieve a reduction in measuring uncertainty. It is possible to reduce random measuring uncertainties with m-repetitions by a factor $\sqrt{m}$. The proportion u_{EVO} then becomes

$$u_{EVO}^* = \frac{u_{EVO}}{\sqrt{m}}$$

If u_{EVO} is known from previous measurements, it is possible to determine the necessary number of repetitions to achieve the required measuring uncertainty.

Measurement chain



MSA for discrete characteristics

Discrete or attributive characteristics are measurements which can only deliver a result that is good or bad (two levels). This is often the case, e.g., in subjective observations.
In the **gauge R&R for discrete characteristics** process, to appraiser's "measure" different parts twice each. That might for example be done on parts that are either intact or faulty. If there are deviations between the results of one appraiser, or between different appraisers, these are counted. The ratio between different results and the number of parts should not be more than 10%.

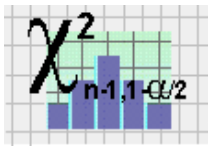
In what is called **Cohen's kappa method**, the appraiser is told to measure one part three times (result as 0 or 1). This doesn't detect deviations between appraisers, but also deviations from the actual values (reference measurements as the true status of the parts).
Score values are calculated based on the ratio between the deviations and the reference values, and must be tested against the confidence ranges from the binomial distribution. Because it relates to a reference value, this method is more meaningful than the gage R&R method. For more information, refer to MSA 4.

In what is called the **Bowker process**, three types can be taken into account, e.g., good, bad and additionally, the "non-uniform" result. At least 40 different test objects are tested by 2 appraisers, 3 times each. Each of the 40 results is allocated to three classes:

Class 1: All 3 repetitions produced the good result
Class 2: Different results within the 3 repetitions
Class 3: All 3 repetitions produced the bad result

The result in the form of a cross table is tested for symmetry using the χ^2 distribution. See VDA 5 for more information.

8. Statistical hypothesis tests



For executing the most important statistical tests the following templates including examples are available. Those are available in the directory \Templates\4_StatisticalTests:

χ^2-Test of goodness of fit

Similar to the KS-test, a sample of a population is compared to a theoretical distribution. The test statistic is determined by:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(H_B - H_E\right)^2}{H_E}$$

with k=number of classes resp. characteristics. This test value can be determined by the Visual-XSel function ***Chi²*** (see functions category statistical tests). The observed frequencies stand in column 1, the expected in column 2. If the expected frequencies for a contingency-table stands in an own table area, the function ***Chi²Contingency2*** has to be used.
There is a check of the null hypothesis: the noticed distribution H_B corresponds to the expected H_E, whereby here the absolute single frequencies are meant. In general, the χ^2-test of goodness of fit ascertains distribution irregularities. If there are small sample volumes the KS-Test rather recovers deviants from normal distribution.
This test statistic is compared to a critical value, which can be found in pertinent statistical tables, or can be specified via Visual-XSel function ***CriticalWorth_Chi²****(f, alpha,* χ^2_{kr} *)* (with alpha = 1-α). Here degree of freedom f is needed, which is determined as follows:

f = k - 1 - a

whereby a is the number of the estimated additional parameters. At assimilation to a Binomial distribution or Poisson distribution a=1, at normal distribution a=1, 2 or 3. If $\bar{x}$ and s are estimated from the categorized data, 3 degrees of freedom are needed, if $\bar{x}$ and σ are calculated directly from the original data, 2 degrees of freedom are needed and if μ and σ are known and the unknown parameter a is estimated from the original data, just 1 degree of freedom is needed.

If $\chi^2 > \chi^2_{crit}$ the null hypothesis is rejected on the level of significance α.

The example file is named *StatTest_Chi2Goodness.vxg*, which can be adjusted easily for own analysis. If another distribution than the normal distribution should be tested, this has to be exchanged accordingly in the subprogram *ExpectedValues* .

It has to be taken into consideration that the check for single frequencies<1 is inaccurate. For monitoring this own subprogram *CheckMinFrequency* has been defined, which supplies corresponding hints. However, a calculation is carried out at any rate. If there are too small single frequencies for certain characteristics, those have to be summed up manually with other values, by what different class distances develop.

See also χ^2-Homogeneity test

χ^2-Homogeneity test

In a so-called multi-field- or contingency-table with r lines and c columns frequencies are situated with characteristic M_B listed in columns and characteristic M_A listed in lines.

	M_{B1}	M_{B2}	M_{B3}	M_{B4}	M_{Bc}
M_{A1}	n_{11}	n_{21}	n_{31}	$n_{4..}$	n_{c1}
M_{A2}	n_{12}	n_{22}	n_{32}	$n_{4..}$	n_{c2}
M_{A3}	n_{13}	n_{23}	n_{33}	$n_{4..}$	n_{c3}
$M_{A..}$	$n_{1..}$	$n_{2..}$	$n_{3..}$	$n_{4..}$	n_{c4}
M_{Ar}	n_{1r}	n_{2r}	n_{3r}	n_{4r}	N_{cr}

The expectation frequencies are calculated for each field by $H_E = n_i * n_j / n$, whereby n_i = line sum, n_j = column sum and n = sum total.

It is allowed to use the test, if all expectation frequencies >=1 ! If there are smaller expectation frequencies, the table should be simplified by summarization of sub occupied fields.

Null hypothesis is: characteristic values are independent from each other or distributed homogeneously.

Test statistic is calculated by

$$\chi^2 = n\left[\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{i,j}^2}{n_i n_j} - 1\right]$$

which can be calculated by the function ***Chi²Contingency1*** (see functions category statistical tests). This test statistic is compared to a critical value, which can be found in pertinent statistical tables, or can be determined via the Visual-XSel function ***CriticalWorth_Chi²(f, alpha,*** χ^2_{kr} ***)*** (with alpha = 1-α). Here a degree of freedom f is needed, which is determined by: f = (r-1)*(c-1).

If $\chi^2 > \chi^2_{crit}$ the null hypothesis is rejected on the level of significance α .

The corresponding example can be found in the file *StatTest_Chi2Homogen.vxg* and can be adjusted for own tests. A component and its improvement measures are observed regarding its failure behavior:

	Starting con-str.	Measure 1	Measure 2
Failure after 1 week	14	22	32
Failure after 2 weeks	18	16	8
Failure after 3 weeks	8	2	2

Question is, if the measures have a temporal influence on the failure behavior. χ^2 results 17,04, the critical value χ_{kr} is for the level of significance 0,05 and the degree of freedom 4 $\chi_{4,0,95}$ = 9,46 and therefore smaller than χ^2, that means there is no independence of characteristics, a temporal influence on the failure behavior does exist (there is no influence on the null hypothesis).

See also χ^2-Test of goodness of fit und χ^2- Multi field test

χ^2- Multi field test

Several samples of a population are compared. The null hypothesis is: the mean number of errors per unit is equal to the complete population.

The so-called contingency table looks like following:

Population i	1	2	...	k
Sampling volume	n_1	n_2	...	n_k
Number of errors in a sample	x_1	x_2	...	x_k

The test statistic is determined by:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(x_i - n_i \frac{x_{ges}}{n_{ges}} \right)^2}{n_i \frac{x_{ges}}{n_{ges}}}$$

with $x_{ges} = \sum_{i=1}^{k} x_i$ and $n_{ges} = \sum_{i=1}^{k} n_i$

which also can be calculated by the visual-XSel function ***Chi²Contingency3*** (see functions

category statistical tests). χ^2 is compared to a critical value, which can be found in pertinent statistical tables, or can be determined via the Visual-XSel function ***CriticalWorht_Chi²***(*f, alpha,* χ^2_{kr}) (with alpha = 1-α). Here a degree of freedom f is needed, which is determined by: f = k-1.
The example file is *StatTest_Chi2Multifield.vxg* and can easily be adjusted for own evaluations. The same method is used for the template
StatTest_Defects_2Samples_Contingency.vxg

If $\chi^2 > \chi^2_{crit}$, the null hypothesis is rejected on the level of significance.

See also χ^2-Homogeneity test

Binomial-test

The binomial distribution describes the number of faulty unities in random checks. The number of faulty unities in the random check can be used for a monitoring of the portion of faulty unities in the total population The likelihood density is:

$$h = \binom{n}{x} p^x (1-p)^{n-x}$$

x : variable
n : number of samples
p : relative ratio of faulty parts

The two-sided confidence level for $p_o = \frac{n_{faults}}{n_{checked}}$ is:

$$p_{low} = \frac{(x+1)\,F_{f1,f2,1-\alpha/2}}{n-x+(x+1)F_{f1,f2,1-\alpha/2}} = \frac{2\,p_o\,n+u^2_{1-\alpha/2}-u_{1-\alpha/2}\sqrt{u^2_{1-\alpha/2}+4p_o\,n(1-p_o)}}{2(n+u_{1-\alpha/2})}$$

with $f1 = 2(x+1)$ and $f2 = 2(n-x)$

$$p_{up} = \frac{x}{x+(n-x+1)F_{f1,f2,1-\alpha/2}} = \frac{2\,p_o\,n+u^2_{1-\alpha/2}+u_{1-\alpha/2}\sqrt{u^2_{1-\alpha/2}+4p_o\,n(1-p_o)}}{2(n+u_{1-\alpha/2})}$$

with $f1 = 2(n-x+1)$ and $f2 = 2x$

The one-sided confidence level is:

$$p_{ob} = \frac{(x+1)F_{f1,f2,1-\alpha}}{n-x+(x+1)F_{f1,f2,1-\alpha}}$$

with $f1 = 2(x+1)$ and $f2 = 2(n-x)$

The hypothesis m events are equal p_o will be rejected if $Z > u_{1-\alpha/2}$

$$Z = \frac{m - n\,p_o}{\sqrt{n\,p_o\,(1-p_o)}}$$

As a test for a hypothesis whether a given faulty portion is within the confidence area lies is the template *StatTest_Defects_1Sample_to_Specification.vxg.*

Kolmogorov-Smirnov-Assimilation test

The Kolmogorov-Smirnov-Assimilation Test (short KS-Test) check the assimilation of an observed distribution to any expected distribution. Especially at existence of small sampling volumes the KS-Test detects rather variances from the normal distribution. In general distribution irregularities better can be proved via χ^2-Test. The KS-Test also can be used for continuous and for discrete distributions.
The null hypothesis is tested: The sample is descended from a known distribution. For each value the relative cumulative frequencies are compared and the maximum difference value is used as test statistic T_{test}.

$$T_{test} = \frac{\max\left|H_B - H_E\right|}{n}$$

This test statistic is compared to a critical value, which can be found in pertinent statistical tables, or can be determined via the Visual-XSel function ***CriticalWorth_KS****(n, alpha, T_{kr})* (with alpha = α).

If $T_{test} > T_{crit}$ the null hypothesis is rejected on the level of significance α .

The example file is called *StatTest_KolmogSmirnov_Assim.vxg*, which can easily be used for own data. In this file the number of points of a cube is checked. Of course, the same number is expected for all six sides, but there are coincidental variances. So, the maximum variance of cumulative frequencies is compared to an equal distribution. This does not exist in Visual-XSel and therefore has to be defined as an own subprogram (*DistribEqual*). If e.g. there is a test for another evaluation versus a normal distribution, the *DistribEqual* has to be exchanged to *DistribNormal* (see functions category statistical distributions).

Shapiro-Wilk test

The Shapiro-Wilk test, proposed in 1965, calculates a *W* statistic that tests whether a random sample, $x_1, x_2, \ldots, x_n$ comes from (specifically) a normal distribution (null hypothesis). The *w* statistic is calculated as follows:

$$w = \frac{b^2}{(n-1)s^2}$$

$$b = \sum_{i=1}^{n} a_i x_i$$

The null hypothesis is rejected on a significance level α if $w < w_{crit}$. The a weights and the critical w_{crit} can be found in the literature.
The alternative to this is to use a following test value T:

$$T_{test} = \gamma + \delta + \ln\left(\frac{w-\varepsilon}{1-w}\right)$$

The coefficients γ, δ and ε can be found also in the literature. The advantage of T_{test} is, that the result can be compared directly with the u-value of the normal distribution (quantile of the standardized normal distribution). If T_{test}<-1,645 the null hypothesis is rejected (significance level α=0,05).

To use this test there is available the template:

StatTest_ShapiroWilk_normal_distribution.vxg

Anderson-Darling test of normal-distribution

The Anderson Darling test checks the null hypothesis that the sample data comes from a normal-distributed population.
This test is suitable for small and big samples size and considers in particular the edge areas of the data.

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2\mathrm{i}-1)\left(\ln(\Phi(u_{\mathrm{i}})) + \ln(1-\Phi(u_{\mathrm{n+1-i}}))\right)$$

with

$$u_i = \frac{x_i - \bar{x}}{s}$$

and $\Phi(u_i)$ for the probability of the u-values of the normal-distribution.
With the help of A^2 there is defined the z-value:

$$z = A^2\left(1 + \frac{0{,}75}{n} + \frac{2{,}25}{n^2}\right)$$

The following table shows the p-value for the adequate z-value:

$z \le 0,2$	p-value = 1 − exp(−13,436 + 101,14z − 223,73 z^2)
$0,2 < z \le 0,34$	p-value = 1 − exp(−8,318 + 42,796z − 59,938 z^2)
$0,34 < z \le 0,6$	p-value = exp (0,9177 − 4, 279z − 1,38 z^2)
$0,6 < z$	p-value = exp(1,2937 − 5, 709z + 0,0186 z^2)

The p-value is the confidence level for the alternative hypothesis that the data are not normal-distributed. Therefore, the data is normal-distributed if the p-value > 0.05.

The adequate template for this method is
StatTest_Normal_Distribution_Anderson_Darling.vxg.

Literature /23/ .. /24/

t-test for two samples

This test checks the null hypothesis: The mean values of both samples are equal. From s and $\bar{x}$ of both samples the test statistic t_{test} is calculated in subprogram t_Test of the file *StatTest_t.vxg* . (The subprogram is also directly available as ***tTest*** in the selection functions category *Statistical Tests*).

$$t_{test} = \frac{\bar{x}_1 - \bar{x}_2}{s_d}$$

with

$$s_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Degree of freedom f is determined by:

$$f = \frac{1}{\frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}}$$

with $c = \frac{s_1^2}{n_1 s_d^2}$

This test statistic is compared to a critical value t_{kr}, which can be found in pertinent statistical tables, or can be determined via the function ***CriticalWorth_t****(f, alpha,* t_{kr}) (with alpha = 1 - $\alpha/2$).

If $t_{test} > t_{crit}$ the null hypothesis is rejected on the level of significance α .

Strictly speaking a F-Test should be executed before each t-test, to confirm the preconditioned equality of variances. If the null hypothesis of the variances is rejected, the t-test delivers wrong values.

The double-sided confidence range is determined by:

$$\bar{x}_1 - \bar{x}_2 - t_{f,1-\alpha/2}\, s_d \le (\mu_1 - \mu_2) \le \bar{x}_1 - \bar{x}_2 + t_{f,1-\alpha/2}\, s_d$$

See also
t-Test for Comparison of a Sample with a Default
U-Test for two Samples (distribution independent)

Test for comparison of one sample with a default value

This test checks the null hypothesis: the mean value of the sample corresponds to a default mean value μ_o.

The test statistic is determined by

$$u_{pr} = \frac{\bar{x} - \mu_o}{\sigma}\sqrt{n}$$

This test statistic is compared to the u-value at an alleged level of significance α. The u-value can be determined in the category statistical distributions (with alpha = 1 - α/2) by the function ***InvNormal****(alpha, u_α)*.

The null hypothesis is rejected if for the double-sided test-case $\mu = \mu_o$ is:

$$\left| u_{pr} \right| > u_{kr}$$

Often there is the one-sided test questioned, for example if the mean is greater than an upper or a lower limit. The table shows the possible problems:

H_o	H_1	H_o is rejected if
$\bar{x} = \mu_o$	$\bar{x} <> \mu_o$	$\left\|u_{pr}\right\| > u_{1-\alpha/2}$
$\bar{x} \geq \mu_o$	$\bar{x} < \mu_o$	$u_{pr} < -u_{1-\alpha}$
$\bar{x} \leq \mu_o$	$\bar{x} > \mu_o$	$u_{pr} > u_{1-\alpha}$

If σ is not known and has to be estimated from s, the test statistic results with

$$t_{test} = \frac{\bar{x} - \mu_o}{s}\sqrt{n}$$

The null hypothesis is rejected, if for the double-sided test-case $\mu = \mu_o$ is:

$$|t_{test}| > \ t_{crit}$$

The critical t-value can be found in pertinent statistical tables, or can be determined via the Visual-XSel function ***CriticalWorth_t****(f; alpha; t_{kr})* (with f = n-1, alpha = 1 - α/2).

See also t-Test for two samples

The adequate template for this method is
StatTest_t_1Sample.vxg

U-test for two samples

This test after Wilcoxon, Mann and Whitney tests over the order whether the median values of two spot checks are equal. It is the distribution-independent counterpart to the t-Test and insensitively against different variances. The U-Test is therefore put in if no normal-distribution can be presupposed.

To the calculation of the test value U, one brings the n_1 and n_2 to big spot checks in a common ascending order, with which is noted to each position-number, from which comes the two spot checks it. Example: Following spot checks are available:

Spot 1	Spot 2
7	3
14	5
22	6
36	10
40	17
48	18
49	20
52	39

	Spot 1	Spot 2
3		1
5		2
6		3
7	4	
10		5
14	6	
17		7
18		8
20		9
22	10	
36	11	
39		12
40	13	
48	14	
49	15	
52	16	
	$\Sigma = 89$	$\Sigma = 53$

In the common order emerges with the ranked numbers for the spot check 1 and 2 the values represented right with the position-sums R_1 and R_2. A test value can be determined for each spot check:

$$U_1 = n_1 n_2 \frac{n_1 (n_1 + 1)}{2} R_1$$

$$U_2 = n_1 n_2 \frac{n_2 (n_2 + 1)}{2} R_2$$

The in the end required test value U the smaller of the two, in this case U=$U1$=11, that is compared against a critical value of U_{crit}, is in the presentation-file
StatTest_U_Wilcoxon.vxg
In case of the fact that there are the same values one speaks of binding. In this case a middle rank is formed in each case and the rank sums are provided with a correction factor. More information will be found in the literature.

If $U < U_{crit}$, the hypothesis that the median values of the spot check are equal is rejected.

See also t-Test for Two Samples

F-test

The variances of two samples are tested.

The null hypothesis is: the samples are descended from the same population.

The test statistic is formed by:

$$F_{test} = \frac{s_1^2}{s_2^2}$$

whereby the larger variance is always in the counter, so that F_{test} >= 1. This value is compared to the critical F-value, which can be found in pertinent statistical tables, or can be determined via the Visual-XSel function ***CriticalWorth_F***(f_1, f_2, *alpha*, F_{crit}) (with alpha = 1-α/2). The degree of freedom f_1 *and* f_2 results from $f_1 = n_1 - 1$ and $f_2 = n_2 - 1$, whereby the index 1 always refers to the sample value with the larger variance.

If $F_{test} > F_{crit}$ the null hypothesis is rejected on the level of significance α .

The example file is called *StatTest_F.vxg*, which can be used for own evaluations.

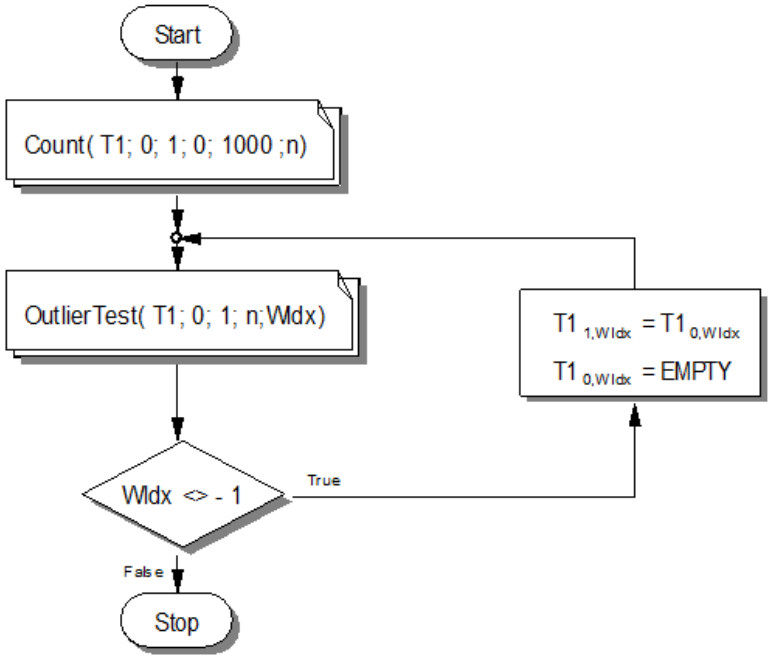
See also Rank Dispersion Test

Outlier test

With this test a series can be checked on one or several outliers. Precondition is that data are normal distributed. Sequentially this test can be repeated as long as no outlier can be determined any more. After ascertainment of an outlier this has to be removed from the series, before the next test is called. The test measurement is:

$$T_{test} = \frac{n}{(n-1)^2} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4$$

which will be compared to a critical value T_{crit} . The test is carried out within the available function *OutlierTest* in the category statistical tests. What will be supplied back is the index of the line in the matrix (resp. table), in which an outlier has been ascertained. In the example file *StatTest_Outlier.vxg* there is a program, which eliminates corresponding values from a series, before the next test is carried out.



After run of this program all outliers will be written on the right side besides the series.

Balanced simple Analysis of Variance

Expectation values of several samples (number k) with same volume n are compared. The null hypothesis is: all expectations values are equal. Precondition for the test is that $\sigma_i = \sigma$. The test statistic is formed by:

$$F_{test} = \frac{n s_{\bar{x}}^2}{s^2}$$

with

$$s_{\bar{x}}^2 = \frac{1}{k-1} \sum_{i=1}^{k} (\bar{x}_i - \bar{\bar{x}})^2$$

$$\bar{\bar{x}} = \frac{1}{k} \sum_{i=1}^{k} \bar{x}_i$$

$$s^2 = \frac{1}{k} \sum_{i=1}^{k} s_i^2$$

whereby x_i and s_i of the respective samples are assumed. This value is compared with the critical F-value, which can be found in pertinent statistical tables, or can be determined via Visual-XSel function ***CriticalWorth_F(*** f_1, f_2, *alpha,* F_{kr} *)* (with alpha = 1-α). The degree of freedom f_1 *and* f_2 results from f_1 = k-1 and f_2 = k(n-1).

If $F_{test} > F_{crit}$ the null hypothesis is rejected on the level of significance α .

The double-sided confidence belt is determined by the critical t-values (***CriticalWorht_t****(f2; alpha; t*) with alpha = 1-α/2 to

$$x_i - t_{f2,1-\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le x_i + t_{f2,1-\alpha/2}\frac{s}{\sqrt{n}}$$

The corresponding example file is *StatTest_BalVariance_Analysis.vxg,* which easily can be adjusted for own evaluations.

See also Bartlett Test

Bartlett-test

More than two populations are compared. The null hypothesis is: all variances are equal. Precondition for test application is that $n_i \ge 5$. The test statistic results to:

$$\chi^2_{pr} = \frac{1}{c}\sum_{i=1}^{k}\left[f_i \ln(\frac{s^2}{s_i^2}) \right]$$

with $f_i = n_i - 1$, $f_{ges} = \sum_{i=1}^{k} f_i$

$$s^2 = \frac{1}{f_{ges}}\sum_{i=1}^{k} f_i\, s_i^2$$

$$c = 1 + \frac{1}{3(k-1)}\left[\left(\sum_{i=1}^{k}\frac{1}{f_i} \right) - \frac{1}{f_{ges}} \right]$$

with k = number of populations = number of samples. x_i and s^2_i of the respective samples are taken for granted, respectively have to be calculated before.
χ^2_{pr} is compared to a critical value, which can be found in pertinent statistical tables, or can be determined via Visual-XSel function ***CriticalWorth_Chi²****(f, alpha,* χ^2_{crit} , with alpha = 1-α). Here a so-called degree of freedom f is needed, which is determined by: f = k-1.

If $\chi^2_{test} > \chi^2_{crit}$, the null hypothesis has to be rejected on the level of significance α.

The example file is *StatTest_Bartlett.vxg* and can easily be adjusted for own evaluations.

See also: Balanced Analysis of Variance

Rank dispersion test according to Siegel and Tukey

At this test two samples are compared to each other, where you cannot assume that they are normal distributed. The test is free of distribution. The null hypothesis H_0 is: both samples belong to a common population.
While executing the tests both samples are gathered up in a series and sorted. The smallest value gets ranking 1, the both largest values get descending rankings 2 and 3, the next smallest values 4 and 5 ascending and so on. If there is an odd number of observations, the middle observation gets no ranking so that the highest ranking always is an even number. For distinguishing which value belongs to which sample, those are indicated before (value 1 for sample 1 and value 2 for sample 2). Afterwards the total of ranking numbers for each sample is formed and issued.
The more the ranking number totals distinguish, the less it is probable that they belong to the same population.

In the following table the lower and upper limits of ranking numbers are shown

n1->	4		5		6		7		8		9		10	
n2=n1	10	26	17	38	26	52	36	69	49	87	62	109	78	132
n2=n1+1	11	29	18	42	27	57	38	74	51	93	65	115	81	139
n2=n1+2	12	32	20	45	29	61	40	79	53	99	68	121	84	146
n2=n1+3	13	35	21	49	31	65	42	84	55	105	71	127	88	152
n2=n1+4	14	38	22	53	32	70	44	89	58	110	73	134	91	159
n2=n1+5	14	42	23	57	34	74	46	94	60	116	76	140	84	166

H_0 is rejected (α=0,05 double sided resp. α=0,025 one sided) if R1 or R2 exceeds or falls below or reaches the lower respectively upper barrier.

The file *StatTest_Rank_Dispersion.vxg* is used as submission, which can be adjusted for own evaluations. For following both samples R1=134 and R2=76 have been determined. As it can be seen in the above table that R1<78 and R2>132. So, the null hypothesis has to be rejected, there is no dispersion difference.

Sample 1	Sample 2
10,1	15,3
7,3	3,6
12,6	16,5
2,4	2,9
6,1	3,3
8,5	4,2
8,8	4,9
9,4	7,3
10,1	11,7
9,8	13,1

Test of a best fit straight line

Any best fit straight line is tested on linearity and gradient. This test is a summarization of both single available submissions.

The file *StatTest_StraightLine.vxg* is used as submission, which can be adjusted for own evaluations.

Test on equal regression coefficients

Two series are tested on equal regression coefficients. There the following t-value is calculated,

$$t = \frac{|b_1 - b_2|}{\sqrt{\frac{s_{yx1}^2 (n_1 - 2) + s_{yx2}^2 (n_2 - 2)}{n_1 + n_2 - 4} \left(\frac{1}{Q_{x1}} + \frac{1}{Q_{x2}} \right)}}$$

which is compared to a critical t_{crit} on the level of significance α=5% . Both regression coefficients are equal, if $t < |t_{crit}|$.

The file *StatTest_2_RegrCoeff.vxg* is used as submission, which can be adjusted for own evaluations.

Linearity test

Tests a series on linearity. See also Test of an Equation Straight line
The data entered in the table page T1 are categorized via the function *Classify* and written in the table page T2. All occurring values within one class are entered here horizontally. Out of this matrix a F-value is calculated:

$$F = \frac{\frac{1}{k-2} \sum_{i=1}^{k} n_i \left(\overline{Y_i} - \hat{Y}_i \right)^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{i,j} - \hat{Y}_i \right)^2}$$

which will be compared on the level of significance α=5%. If $F < F_{crit\ (k-2,n-k)}$, there is a significant linearity.

The file *StatTest_Linearity.vxg* is used as submission, which can be adjusted for own evaluations.

Gradient test of a regression

It is tested, if the gradient of an equation straight line significantly differs from 0. A t-value is formed as quotient of a regression coefficient b to its variation. From this the value α is determined from the student's t-distribution and compares this to the level of significance α=5%. If α < 5%, significantly a gradient > 0 does exist.
See also Test of an Equation Straight Line, where also this test is included.

The file *StatTest_Gradient_Regression.vxg* is used as submission, which can be adjusted for own evaluations.

Independence test of p series of measurements

Test the correlation-coefficients on mutual independence. For example, this test is used to check at a multiple regression, if all variables are necessary.

Data are entered in the table page T1. First the so-called correlation matrix is formed (stands after start of program in table page T2). There the correlation coefficients r are listed in pairs in every possible combination of series of measurements. A limit R' is determined on the level of significance α via student's t- distribution:

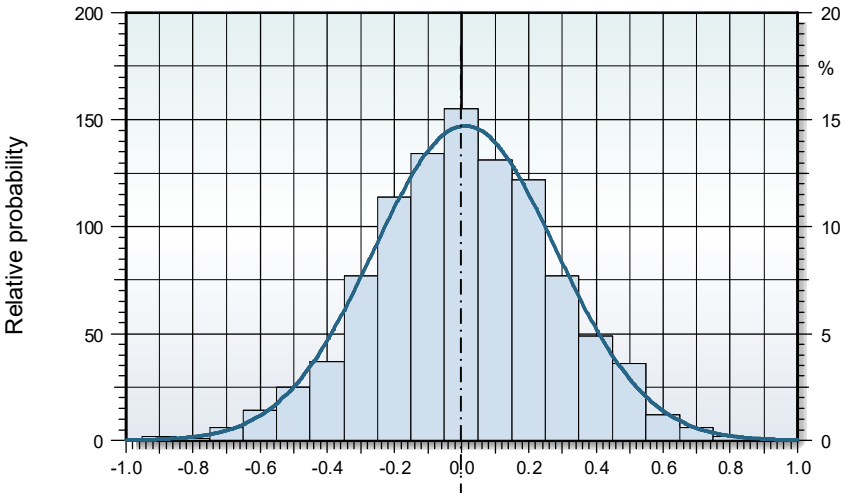
$$R' = \left| r_{\lim it} \right| \sqrt{\frac{n-2}{1-r_{\lim it}^2}}$$

and iterativly r_{limit} is calculated, which is compared to the maximum founded correlation coefficient. If $r_{max} < r_{limit}$, then on the level α=5% no pair of series of measurements significantly depends on each other.

The file *StatTest_Independence_p_Series.vxg* is used as submission, which can be adjusted for own evaluations.

9. Normal-distribution

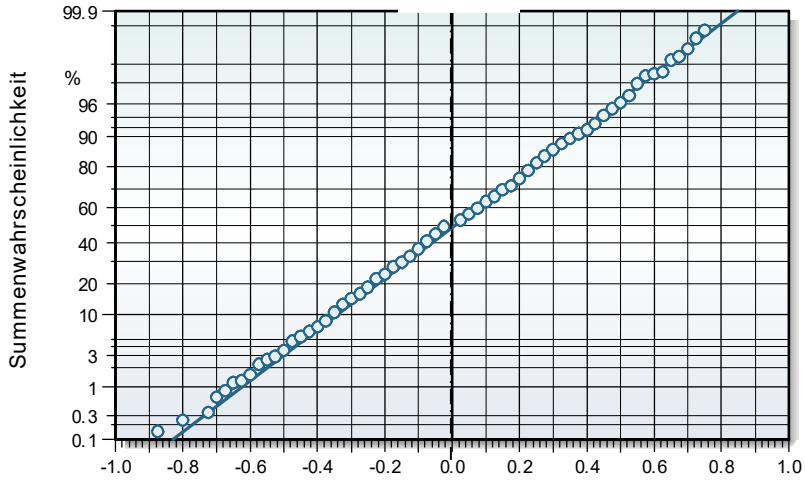
The normal distribution is the most frequently and most common form of the probability distribution. The normal-distribution is relevant when random events affect a process. Many natural, economic and engineering processes can be achieved by the normal distribution, either exactly or at least at a very good approximation (especially processes that act independently to factors in different directions).



$$h = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{s}\right)^2}$$

$$H = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{s}\right)^2} dx$$

Integral not solvable



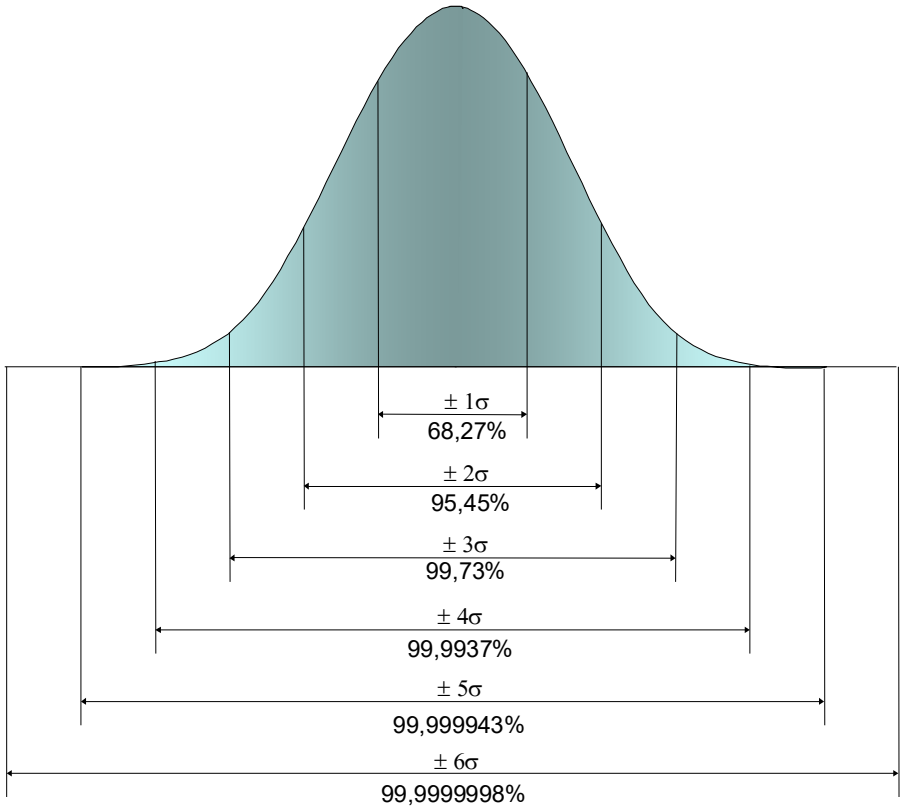
x : Variable
μ : mean
σ : Standard-deviation

In the top histogram the data is grouped in suitable classes. The Gaussian-curve represents the ideal probability-density of this histogram, if one has endless data and an absolute small class-width. In the cumulative probability chart below one can read the sum of values, which are less or equal to x. That is the area under the Gaussian curve. The mean is the probability at 50% (here at x = 0). The confidence interval of the mean value is

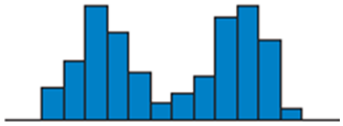
$$\bar{x} - t_{1-\alpha/2} \frac{s}{\sqrt{n}} \leq \bar{x} \leq \bar{x} + t_{1-\alpha/2} \frac{s}{\sqrt{n}}$$

$t_{1-\alpha/2}$: Quantile of the Student-distribution with the significance α, it the real standard-deviation in not known.

The slope of the straight line represents the scatter (standard-deviation) of the data. At $\bar{x} \pm s$ there is 16%, respectively 84% of the data. Both charts have their own advantages. In the histogram a mixed distribution can be detected easily if there are more than one caps. In the cumulative probability chart one can see each data-point and the deviation from the straight-line is a non-regularity to the normal-distribution.

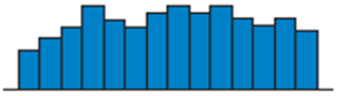


Divergences of the normal distribution



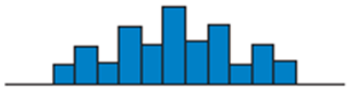
Bimodal distribution

Valley in the middle-> combination of two normal distributions.
Cause: Mixing distribution from two processes.
Remedy: Grouping and subdivision in both dimensions of influence.



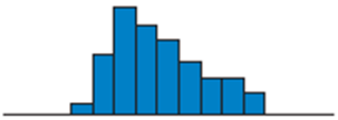
Rectangular distribution

More or less level without distinctive maximum.
Cause: Confounding of several distributions.
Remedy: Subdivision in dimensions of influence with the help of a systems analysis.



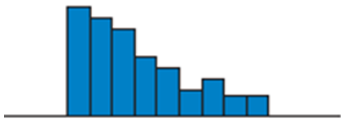
Comb-shaped distribution

Alternately big and small frequencies.
Cause: Often measuring-, rounding-error or unsuitable class width.
Remedy: Check measuring range or use smaller class width.



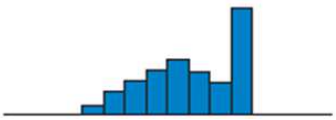
Unsymmetrical distribution

Unsymmetrical course at the short and long end.
Cause: Natural limitation on the left against 0 and high values are seldom.
Remedy: Logarithmic values causes mostly a normal distribution.



Askew distribution

Unsymmetrical unilaterally abrupt sloping side.
Cause: Cut-off distribution by sorting out.
Remedy: Check process and need of sorting out (cost).



Truncated distribution

Accumulation of a group lying on the edge.
Cause: Grouping of all measurements lying on the right.
Remedy: Clarification the classing of the data or the sorting.

10. Statistical factors

Factor	Definition	Description
DF	Degrees of Freedom	For statistical tests
N	Number of populations	e.g., production quantity
n	Sampling volume of degree of freedom or number of independent trials	In general: number of parts
f	Degree of freedom	for statistical tests
k	Number of categories	
i	Ordinal number	In general: running index
H	Frequency	Mostly in %
x_o	Reference value of population	Mostly approximated mean value
$\bar{x}$	Mean value of a sample	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Δx	Class size	In general: increment
μ	Mean value of population	
R	Range	$R = x_{max} - x_{min}$
s	Standard deviation of sample	$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$
s^2	Variance of sample	
σ	Standard deviation of population	
p	Probability of success	
b	Form parameter at Weibull	Gradient of equation straight line in Weibull-Net
t	Life cycle of variable at Weibull	route, length of use, load changes and so on
T	Characteristically service life at Weibull	For 63.2% failure frequency
w	Weighting	Number of alleged values
α	Level of significance for statistical check	The transfer parameter alpha often is alpha = 1- α resp. 1 - $\alpha/2$ for double sided tests
z	Number of variables or factors	

11. Literature

/1/ Keki Bhote
World Class Quality
American Management Association, New York 1991
ISBN 0-8144-5053-9

/2/ Georg E.P. Box, Norman R. Draper
Empirical Model Building and Response Surfaces
Wiley, New York 1987
ISBN 0-471-81033-9

/3/ Wilhelm Kleppmann
Taschenbuch Versuchsplanung
Hanser Verlag München 1998
ISBN 3-446-19271-9

/4/ H. Ulrich, G.J.B Probst
Anleitung zum ganzeitlichen Denken und Handeln
Haupt 1991

/5/ Lothar Sachs
Angewandte Statistik
Springer-Verlag Berlin 1983
ISBN 3-540-12800-X

/6/ Peterson
Grundlagen der Statistik und der statistischen Versuchsplanung
Ecomed Landsberg /Lech 1991
ISBN 3-609-75520-2

/7/ Statistik
Lehr- und Handbuch der angewandten Statistik
Hartung, Elpelt, Klösner
Oldenburg Verlag München Wien
ISBN 3-486-24984-3

/8/ Multivariate Statistik
Lehr- und Handbuch der angewandten Statistik
Hartung, Elpelt, Klösner
Oldenburg Verlag München
ISBN 3-486-21430-6

/9/ Neuro- Fuzzy-Systeme
Borgelt, Klawonn, Kruse, Nauck
Vieweg 2003, ISBN 3-528-25265-0

/10/ Optimierung vielparametriger Systeme in der Kfz-Antriebsentwicklung
Alexander Mitterer
Fortschritt-Bereicht VDI Reihe 12 Nr. 434, ISBN 3-18-343412-1

/11/ Praktische Einführung in Neuronale Netze.
Allesandro Mazzetti
Hannover: Heise, 1992

/12/ Körpereigene Drogen
Josef Zehentbauer
Artemis & Winkler Verlag, 1993, ISBN 3-7608-1935-4

/13/ Wold, H. (1966)
Nonlinear Estimation by Iterative Least Squares Procedures, in:
David, F. N. (Hrsg.), Festschrift for J. Neyman: Research Papers in Statistics,
London 1966, 411-444.

/14/ S. World, E. Johannson, M. Cocchi, 3D QSAR in Drug Desing;
Theory, Methods, and Applications, ESCOM, Leiden, Holland, 1993, pp. 523-550.

/15/ Praxisleitfaden Qualität
Walter Jahn, Lorenz Braun
Hanser Verlag 2006, ISBN-10: 3-446-40616-6

/16/ Taschenbuch der Statistik
Rinne
Verlag Harri Deutsch 2003, ISBN 3-8171-1695-0

/17/ Formelsammlung zu den statistischen Methoden des Q-Managements
DGQ-Band 11-05, Graeb
ISBN 3-410-32877-7

/18/ Multivariate Datenanalyse
Waltraud Kessler
WILEY-VCH Verlag GmbH & Co KGaA
ISBN 978-3-527-31262-7

/19/ Statistics for Experiments
Box & Hunter
Wiley Series
ISBN 0-471-09315-7

/20/ Six Sigma $^{+Lean}$ Toolset
Lunau, Roenpage, Staudter, Meran, John, Beernaert
Springer
ISBN 10-3-540-46054-3

/21/ Design for Six Sigma^{+Lean} Toolset

Lunau, Mollenhauer, Staudter, Meran, Hamalides, Roenpage, von Hugo
Springer
ISBN 978-3-540-69714-5

/22/ Design for Six Sigma
Gamweger, Jöbstl, Strohmann, Suchowerskyj
Hanser Verlag 2009
ISBN 978-3-446-41454-9

/23/ Groß, J.
A Normal Distribution Course.
Peter Lang Verlag 2004
ISBN 978-3631529348

/24/ Stephens, M.A.
Tests based on EDF statistics.
D'Agostino, R.B. and Stephens, M.A., eds.: Goodness-of-Fit Techniques. Marcel Dekker, New York 1986

/25/ Thode Jr., H.C.
Testing for Normality.
Marcel Dekker, New York 2002

/25/ Thode Jr., H.C.
Testing for Normality.
Marcel Dekker, New York 2002

/26/ Graf, Henning, Stange, Wilrich
Formeln und Tabellen der angewandten math. Statistik
Springer Verlag 1987
ISBN: 978-3-540-16901-7

/27/ Measurement System Analysis MSA
Fourth Edition 7/2010
ISBN# 978-1-60-534211-5

/28/ Measurement System Capability – Reference Manual
September 2002,
Version 2.1 D/E

12. Index

Edition 20
Curt Ronniger