

Design of Experiment & Statistics

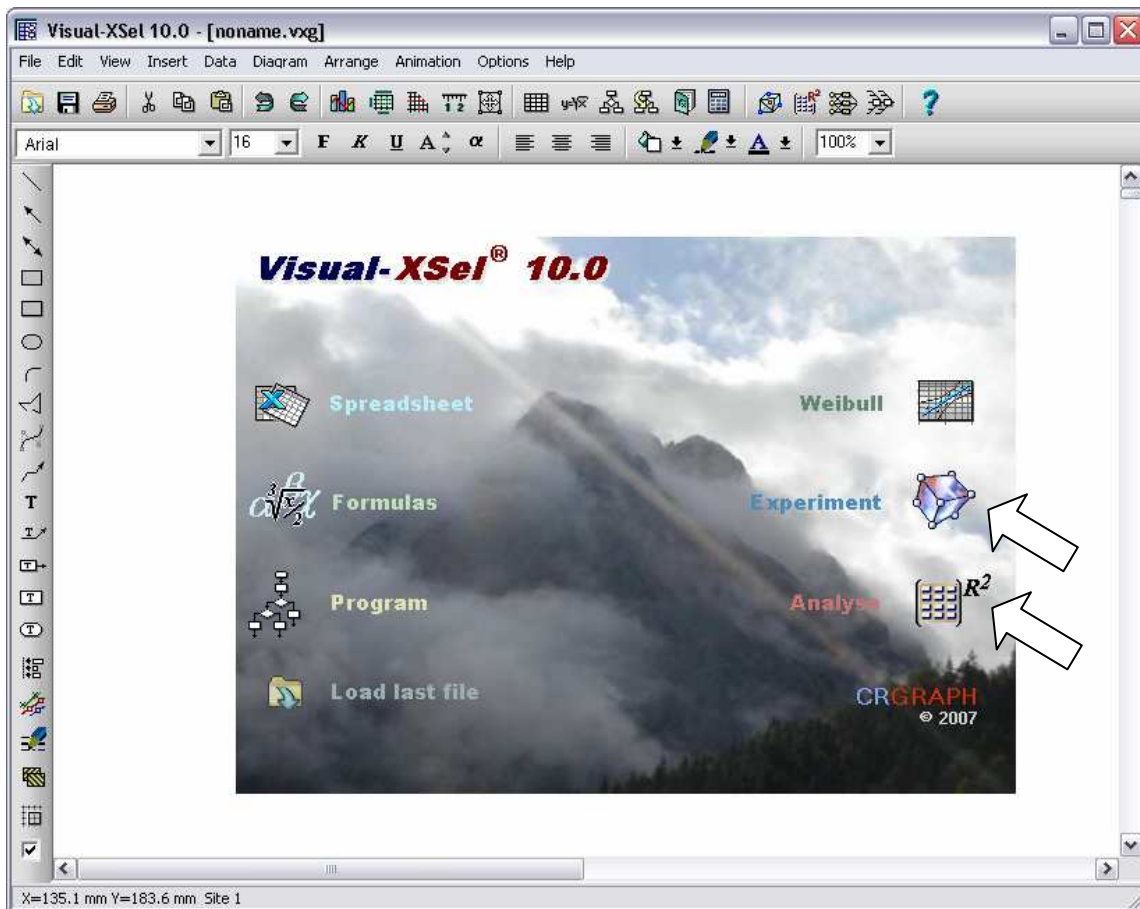
Content

Test methods	6
Components Search	7
<i>Scatter Plot (Realistic Tolerance Parallelogram(SM) Plot)</i>	8
<i>Measurement-Comparison (Isoplot^(SM))</i>	9
<i>Multi-Vari-Chart^(SM)</i>	10
Paired Comparison	11
<i>Comparison B vs. C(SM)</i>	12
Intensity-Relation-Matrix	14
Priority Matrix	15
Analysis of Variance (ANOVA)	17
ANOVA between two Series	17
ANOVA & ANOM with several Factors.....	19
Design of Experiment	23
Design	23
Full-, fractional and Taguchi Experiments	25
Plackett-Burman-Experiments.....	27
Central Composite Design	29
Box-Behnken Design	30
D-Optimal Experiments	31
Mixture Plans	32
Correlation	35
Correlation coefficient after Bravais - Pearson.....	35
Rank correlation nach Spearman.....	35
Correlation matrix.....	36
Partial Correlation Coefficient.....	36
Regression	37
General	37
Linear Regression	37
Linear regression through 0-point	38
Nonlinear regression	39
Regression types	39
Multiple Regression	42
Analyses of Variance (Model ANOVA).....	45
Prediction Measure Q^2	46
Lack of Fit.....	47
Analyses of Variance overview	47
Reproducibility	48
Test of the coefficient of determination	48
Test of the regression coefficients, the p-Value.....	48
Test of the coefficient of determination	49
Standard deviation of the model RMS	50
Confidence interval for the regression coefficient.....	50
Confidence interval for the response	50
Condition Number.....	50
Standardize to -1 ... +1	51
Standardize to standard deviation	51
The correlation matrix.....	51
Response Transformation (Box-Cox).....	52
Statistical Charts for Multiple Regression	54

Regulation of outliers	57
Optimization.....	58
One understands by an optimization of regression models finding the right adjustings of all factors for a minima, maxima or a predefined set point of the response variable.	58
If certain response values have maybe a higher importance than other, this can be taken into account by a weighting factor δ	59
Discrete Regression	60
Discrete regression bases.....	60
Multivariate Analises	66
Cluster Analysis	66
Principal Component Analysis PCA	70
Partial Least Square (PLS)	72
Estimation of the spread at PLS	73
Variable selection with VIP	74
Neural Networks	76
Topology	76
$Y' = x'$	78
Training-Algorithm.....	78
Neural Network as an alternative for multiple regression	79
Attributes of Neural Networks.....	80
Example	81
Further statistical charts.....	82
Scatter bars.....	82
Boxplot	83
Median plot	84
Gliding average.....	85
Pareto	85
Pareto	86
χ^2 -Test of Goodness of Fit	89
χ^2 -Homogeneity Test.....	90
χ^2 - Multi Field Test	91
Kolmogorov-Smirnov-Assimilation Test	92
t-Test for two Samples	92
Test for Comparison of a Sample with a Default Value.....	93
U-Test for two Samples.....	94
F-Test.....	95
Outlier Test	96
Balanced simple Analysis of Variance	97
Bartlett-Test	98
Rank Dispersion Test according to Siegel and Tukey.....	99
Test of an Best Fit Straight Line	100
Test on equal Regression Coefficients.....	100
Linearity Test.....	100
Gradient Test of a Regression	101
Independence Test of p Series of Measurements.....	101
Statistical Factors	102
Normal-distribution	103
Literature	103
Literature	104
Index	106

Software

For the methods and procedures which are shown here the software Visual-XSel[®] 10.0 Weibull 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 (like methods after Shainin). For this use the menu *File/Methods...*

The software can be downloaded via www.crgraph.com

Test methods

Under test methods there are statistical methods to understand which were developed through Shainin /1/ and 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.

To every method there are file templates to comprehend this one with easily examples of Visual-XSel®. The 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.

The following issues are treated:

- Test methodes from Shainin and others
- Taguchi strategy and experiments
- Standard experiments and D-optimal
- Variance - analysis
- Statistical diagrams and spezial charts
- Correlation and regression
- Multiple Regression (stepwise regression)
- Multivariate analyses
- Statistical tests and evaluations
- Statistical distributions
- Optimization

Templates for standard statistics and hypothesis test are provide in the subdirectory
\Statistics

Templates for Shainin and examples are provided in the subdirectory
\StatisticalMethods

Experiments and there evaluation with Multiple Regression or Neural Networks are available via the Data Analysis Guide

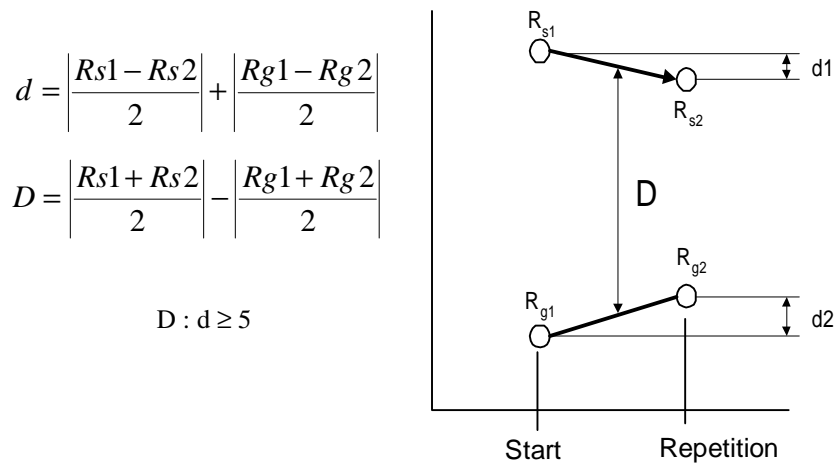
Hint: Shainin® and Red-X® and *Visual-XSel*® are Trademarks.

Components Search

With the component test e.g. an error cause at a "device" should be found. For this a device with a defective performance is necessary and one, which shows faultless characteristics.

Precondition for this method is that the devices can be disassembled non-destructively. Here the performance should not be changed considerably by the re-assembly. Before, it has to be fixed, which components resp. single modules have to be exchanged. The process can be multilevel, that means if a subassembly is found as the relevant one, this also can be demounted into its devices. So that the number of mountings is as small as possible, preferably big units should be used in the beginning.

To differentiate a substantially change compared to the not precluded straggling, first the devices must be disassembled in the beginning and reassembled (repeat trials). This should take place at least twice. From the first measuring and the repetitions after the disassembly a scatter band results (d1 and d2).

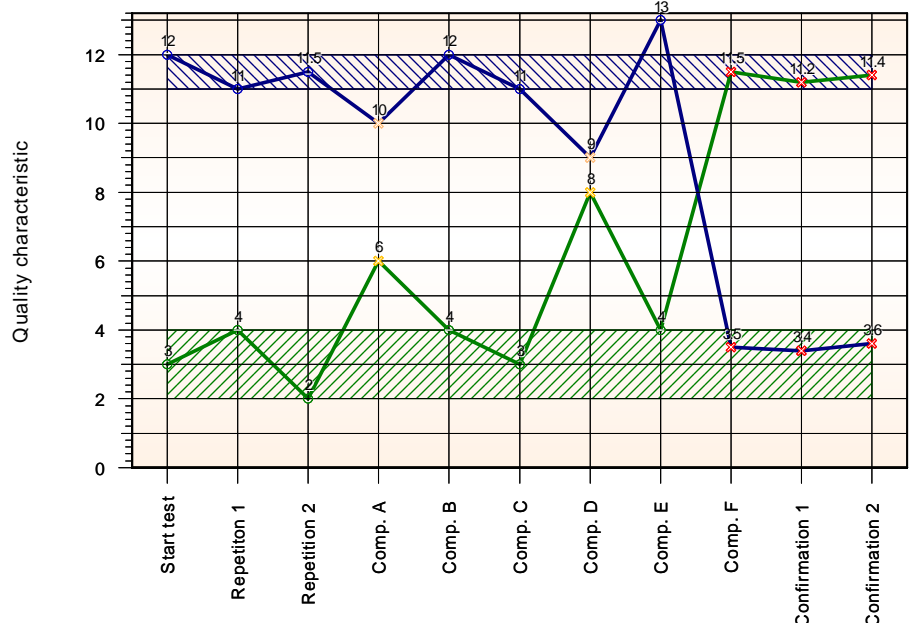


The proportion of the medium discrepancies (D) of the corresponding measurements to those within a device is not permitted to fall below the factor 5.

Now successively the exchange of the devices starts. There the devices with the probably major influence should be exchanged first. After each mounting the devices have to be reconstructed, whatever there was a variation or not. Just then the next device should be exchanged.

At the test overview the results of the good and of the worse devices are pictured as line chart and the particular changes are valued in the distribution.

In the validation trial all devices, which have changed the result, are exchanged together and the result is also entered.

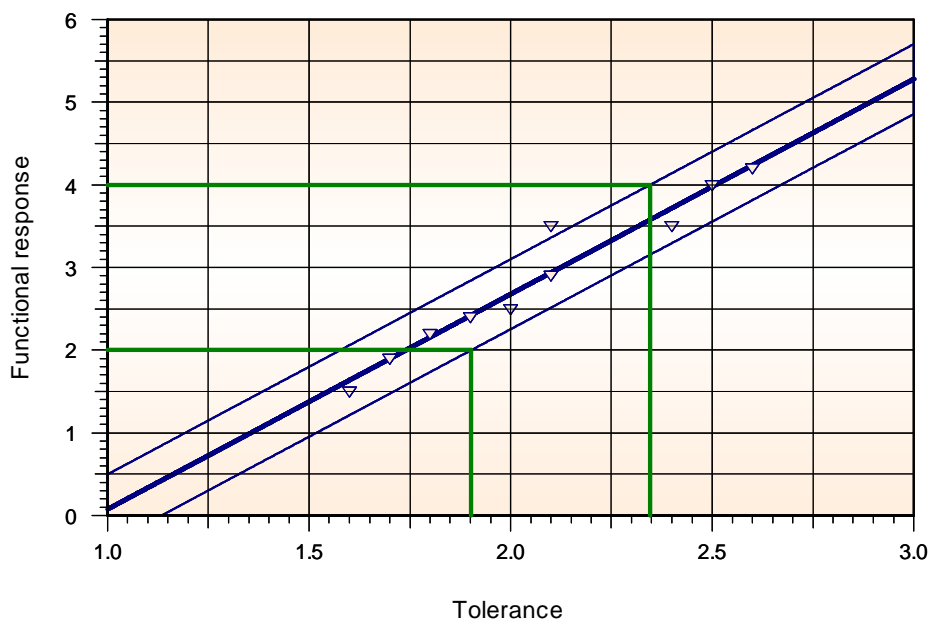


If the measured data alter in the middle, outside the scatter band, so more or less big influences are transparent, which depict a reciprocation (here comp. A and D). If the measured data are cross-over in the respectively other scatter zone, so the device resp. component with the critical influence is found.(the so called red X).

This method can be executed with help of the template file [Components_Search.vxg](#) .

Scatter Plot (Realistic Tolerance Parallelogram_(SM) Plot)

In the classical linear regression an optimal best fit straight line is set through the measured values. In dependency of the straggling of the breakpoints those are more or less far away from this straight line. If a distribution is generated for these deviations (residues), you can determine a frequency region for the number of points. As a rule the 95%-region is depicted, that means 95% of all breakpoints are in the hatched depicted parallel band around the straight line. Precondition for these considerations is of course the normal distribution of the measured values.

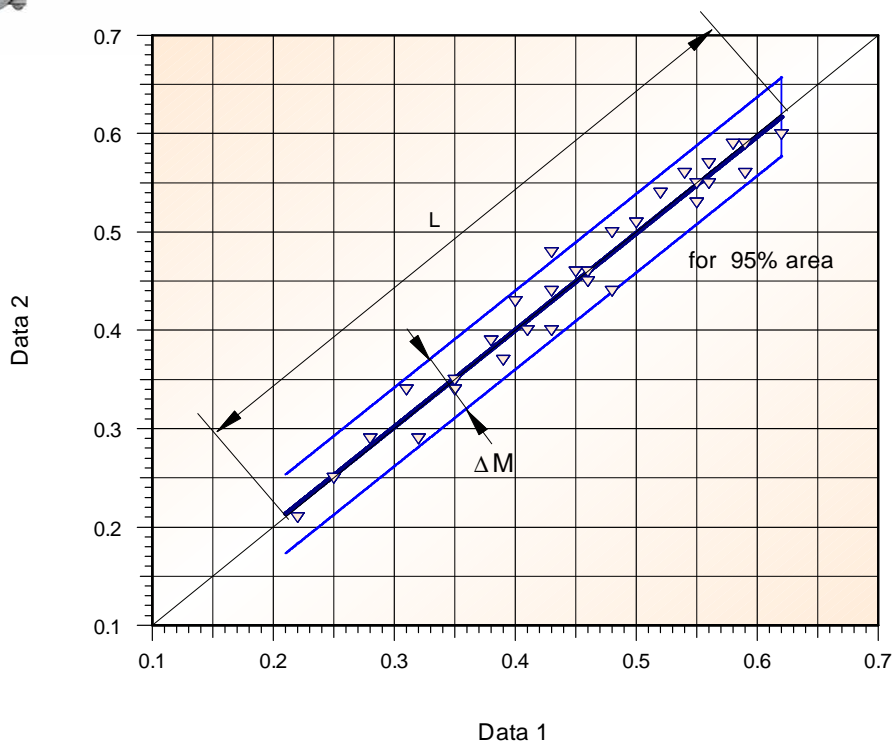
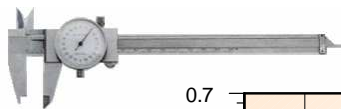


If there are requests for the target value, those can be entered (green horizontal lines) and be drawn to the respectively upper and lower frequency region. The perpendicular lines show the tolerance range on the x-axis necessary for this. This of course is narrower than the one, as if just the best fit straight line would be used, because the scatter band has to be taken into consideration.

This method can be executed with help of the template file [Scatter_Plot.vxg](#). Optional the target value region requested here can be default. After start of the program the upper and lower tolerance range and the corresponding median are issued. If the indication of the target value is open, automatically the smallest and biggest measured value is used.

Measurement-Comparison (Isoplot ^(SM))

The display format of the measurement equipment capability is very similar to the *Scatter-Plot*. But here it is a matter of comparison of two measuring methods. The results of one measurement is spread over the other. The linear regression is below a 45° line (same measure for X and Y). If a distribution for those variances of measured values (residues) is built, you can determine a frequency region for the number of points. Normally the 95%-region is depicted, that means 95% of all breakpoints are in the hatched depicted parallel band around the straight line. Precondition for those considerations is that the measured values are normal distributed. At least 30 measured values should be available.



ΔP is determined from both factors L and ΔM :

$$\Delta P = \sqrt{\frac{L^2}{2} - \frac{\Delta M^2}{2}}$$

The so called resolving power should be

$$\frac{\Delta P}{\Delta M} \geq 6$$

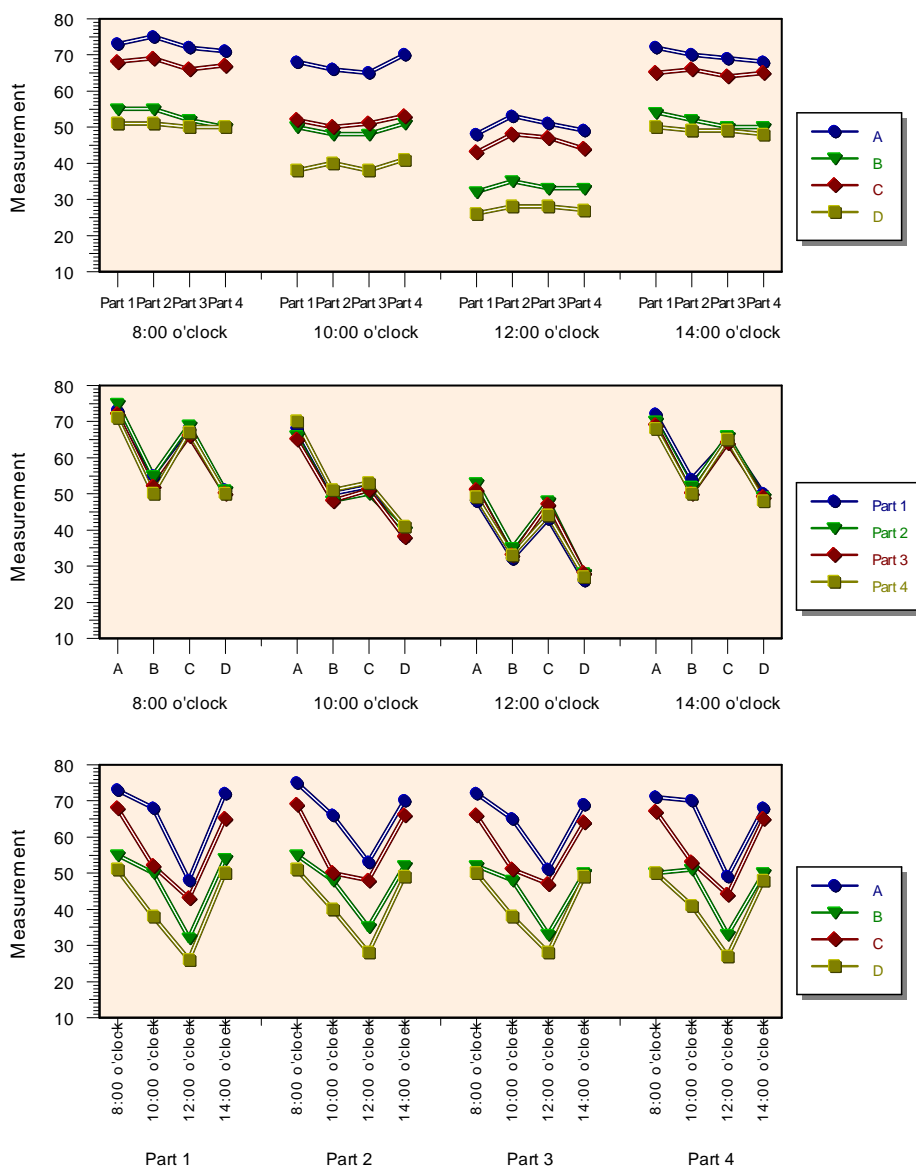
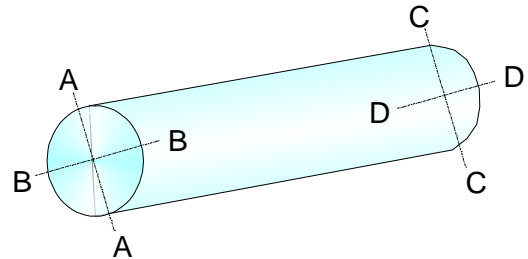
If the best fit straight line is parallel shifted to the 45°-line, there is a constant deviation. If it generates a striking other angle than 45°, this is a variable deviation. This method can be executed with help of the template file [Isoplot.vxg](#).

^(SM) Isoplot is a Service Mark of Shainin corp.

Multi-Vari-Chart^(SM)

By the Multi-Vari-Chart you can recognize, if a certain straggling pattern is position dependent, part conditional or temporal different. The multi-variation-card pictures a snap-reading method for the respective produced parts and provides an indication on systematic errors. Among position dependent you see e.g. measured values within a component, e.g. the diameter of an axle back and front. It also can be synonymous characteristics of an assembly, or measures at different positions of the production and so on.

Under part conditional you see successive parts of a production or batches. The temporal influence can extend over hours, shifts or days.



It is searched for typical patterns of line drawings, resp. bucklings. It is important to know what is expected and what are the valid straggling.

This method can be executed with help of the template file [Multi_Vari_Chart.vxg](#). There it is not necessary that always 4 measures are available.

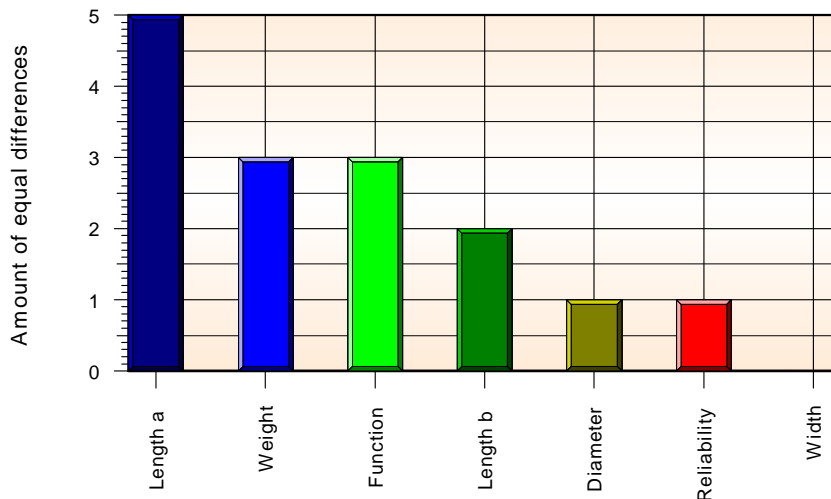
^(SM) Multi-Vari-Chart is a Service Mark of Shainin corp.

Paired Comparison

At the pairwise comparison it is a matter of the comparison of discrepancies between characteristics, regardless their quantitative size. Always pair wise so called "good" and "bad" parts are compared. Various criteria are fixed, which are entered side by side in a table.

	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5
Length a	123 125 <	122 125 <	120 126 <	119 123 <	123 124 <
Length b	23 25 <	22 25 <	20 23 <	23 21 >	23 23 =
Width	11 11 =	12 11 >	11 12 <	11 11 =	12 12 =
Diameter	1 2 <	1 2 <	2 1 >	2 1 >	2 1 >
Weight	1 2 <	1 2 <	1 0 >	0 1 <	-1 1 <
Function	3 4 <	2 4 <	5 4 >	1 2 <	3 4 <
Reliability	34.5 36.5 <	35.8 36.5 <	37 34.5 >	37 36 >	35.8 36.5 <

Then the measured values are compared to each other, if they are bigger, equal or smaller. For this the signs \leq , $=$ or $>$ have to be entered beside the particular pairs. In the evaluation the sign $<$ gets the value -1 , $=$ the value 0 and $>$ the value 1 . If you now add up the summation of rows and pictures the absolute values assorted, you get the ranking of the most important influences:



The result shows, which parameter have to be optimized to get a decisive improvement.

This method can be executed with help of the template file [Paired_Comparison.vxg](#).

Comparison B vs. C_(SM)

Often it occurs that you want to compare 2 "things" e.g. a new product against an old one. There a certain criterion is relevant, which if possible, should be described with a measured value. If you got a number of "New" and "Old" or "B versus C" parts or systems, those are described sequentially after their evaluation. If there for example ever 2 parts, the following sequence could addict:

Part	Evaluation
N	1,2
N	1,1
A	0,9
A	0,8

Here it is provided that the higher evaluation is the better one. At first the result seems to be unique after the resulted sequence. The both new-parts are before the old-parts. But the sequence could also be the same incidentally. For 2 New and 2 Old there are 6 different possibilities in total:

N	N	N	A	A	A
N	A	A	N	N	A
A	N	A	N	A	N
A	A	N	A	N	N

In general the number of possible variants is determined by:

$$\text{Varianten} = \frac{(n_{\text{neu}} + n_{\text{alt}})!}{n_{\text{neu}}! \cdot n_{\text{alt}}!}$$

So the probability of the result in the first column is 1/6 or 16.667%. With this random sample a proposition should be done about the main unit. For this you establish a null hypothesis, that New in fact is better than Old. The probability that the null hypothesis applies is 100% less the random probability of 16.667% = 83.3%. Similarly, as a significance level is fixed for statistical tests, here a limiting value of 5% should be valid. Because this exceeds the limiting value with 16.667% widely, the result of the null hypothesis is not significant. In principle the significance level should be fixed before. At a sequence, where always all new ones are before the Old ones, you just have to compare the reciprocal of number of variants against the fixed significance level. The null hypothesis that New is better than Old, always has to be dismissed, if 100%/variants > significance level.

The following table shows the corresponding sample sizes for various significance levels, where the null hypothesis should not be dismissed:

Significance-level	Scope N	Scope A
0,1%	2	43
0,1%	3	16
0,1%	4	10
0,1%	5	8
0,1%	6	6
1%	2	13
1%	3	7
1%	4	5
1%	5	4
1%	6	
5%	1	19
5%	2	5
5%	3	3
5%	4	3
5%	5	
10%	1	9
10%	2	3
10%	3	2
10%	4	
10%	5	

If a sample size of at least ever 10 is available, you can verify the hypothesis, if the ranking overlaps. E.g.:

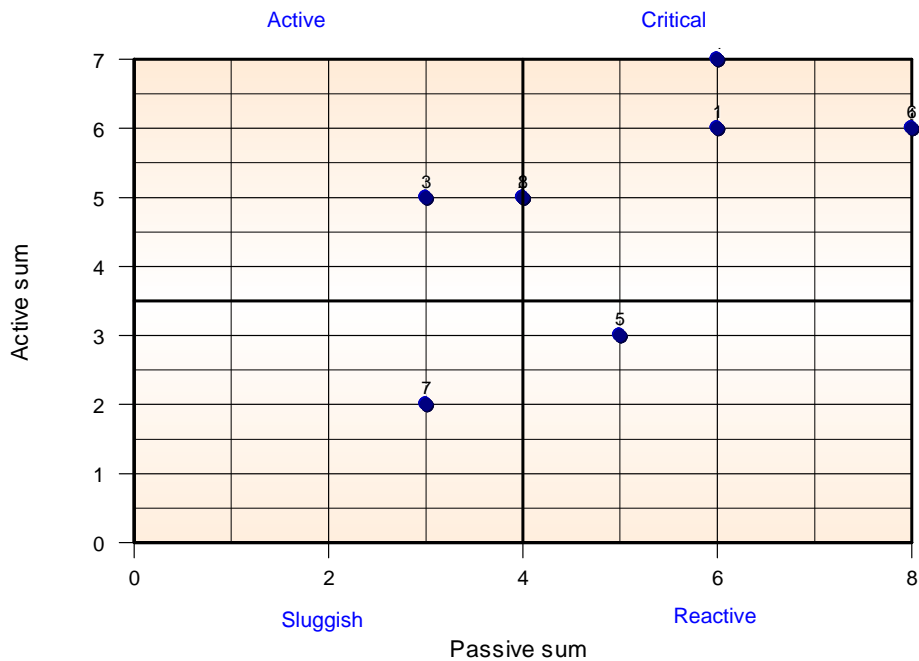
NNNANNAANNNNAAANAAAA

If the sample sizes of N and A differ, than the sample size for N must be bigger. There a proportion of those sizes from 5:4 must not be exceeded. In case of overlapping the respectively same New have to be counted at the beginning, till Old occurs an count Old starting from the end, till New occurs. In our example the number of same mentions $3+4=7$ results. This number has to be compared against the following limiting value:

Significance level	Limiting value
5%	6
1%	9
0,1%	12

If the value of same mentions is bigger than the limiting value in consideration of the significance level chosen before, so the null hypothesis, that New is better, has to be confirmed as Old. In our example this is the case for 5%, because $7 > 6$.

If measurable evaluation is not possible, but just a comparing verification, so nevertheless the test can be used, e.g. if the design of a car has to be compared to each other. The principle always is the same. A ranking New against Old is set up. The "tester" have to compare the cars independently, which car is more handsome. If one and the same car is named from all 5, then this is the more handsome one under the signifi-



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 criteria or characteristics are compared in the Priority matrix together and a ranking was formed. The result can be used also for importances of the criteria for continuing evaluations

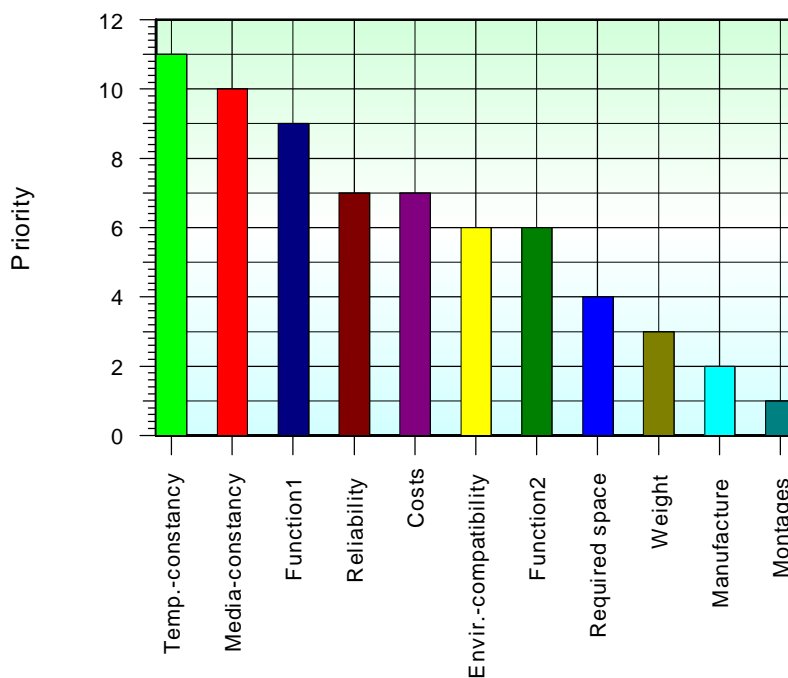
No quantitative measurements are necessary for the comparisons of the characteristics. The test is just a pair-wise comparison and an estimation of experts.

For example: The importances 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 function 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 importances 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](#)

Analysis of Variance (ANOVA)

ANOVA between two Series

In the analysis of variance a significant difference should be determined between two test series. In an example it should be determined, if the body height between Europeans and Africans is different. There are following data:

Euro-peans	Afri-cans
159	187
163	173
156	177
173	181
161	
169	

First the the square sum of variances for the average is formed, which corresponds to the so called correction factor.

$$SQA_m = \frac{(\sum Data1 + \sum Data2)^2}{n_{tot}} = CF$$

with n_{tot} = number of measures Data1 and Data2

Afterwards the total of squared variances is determined:

$$GSQ = \sum Data1_i^2 + \sum Data2_i^2 - CF$$

with the belonging degree of freedom $DF = n_{tot} - 1$

Furthermore the total of squared variances of the single data arrays have to be formed

$$SQA = \frac{(\sum Data1)^2}{n_1} + \frac{(\sum Data2)^2}{n_2} - CF$$

with degree of freedom $DF_A = 1$

The square total of the error is calculated by:

$$SQF = GSQ - SQA$$

with degree of freedom $DF_F = n_{tot} - 2$

Variances are determined accordingly:

$$V_A = \frac{SQA}{DF_A} \quad V_F = \frac{SQF}{DF_F}$$

The so called F-value is the quotient of the both variances

$$F = \frac{V_A}{V_F}$$

which is compared to a critical F-value F_{krit} on a fixed level of significance, e.g. 95%. If $F > F_{krit, DFA, DFF}$, it means that both series are significantly different.

The percental share in the total effect is calculated by

$$SQA' = SQA - DF_A \cdot V_F$$

$$A = \frac{SQA'}{GSQ} 100\%$$

and describes the average effects. The difference to 100% corresponds to the share of errors

This procedure can be used with the submission file [ANOVA_2_Series.vxg](#) in directory `\Statistics`

See also : ANOVA & ANOM with several Factors

ANOVA & ANOM with several Factors

In the analysis of variance with several factors the influences of test parameters are tested on a target size.

It should be found out, which influence do the parameter have on the test result proportional to the dispersions.

After analysis of variance it is issued by the statistical F-Test, if the parameter does have a significant influence and how great its percentage share is compared to the remaining dispersion. It is assumed that the deviations are normal distributed. Otherwise the result is not unique.

In the following example the depicted trials are executed.

	Target size	Temperature	Print	Set up time	Cleaning
1	-20	1	1	1	1
2	-10	1	2	2	2
3	-30	1	3	3	3
4	-25	2	1	2	3
5	-45	2	2	3	1
6	-65	2	3	1	2
7	-45	3	1	3	2
8	-65	3	2	1	3
9	-70	3	3	2	1

The single steps of ANOVA:

1) Formation of square total of deviations for the mean value, which is also indicated as correction factor

$$SQM = \frac{1}{n} \left(\sum_{i=1}^n Y_i \right)^2 = CF$$

2) Formation of square total of deviations of the total mean value

$$GSQ = \sum_{i=1}^n Y_i^2 - CF$$

3) Formation of total of the squared deviations regarding the factors

$$SQA_A = \frac{1}{n_{A1}} (\sum Y_{A1})^2 + \frac{1}{n_{A2}} (\sum Y_{A2})^2 + \frac{1}{n_{A3}} (\sum Y_{A3})^2 - CF$$

whereby n_{A1} , n_{A2} and n_{A3} at each case is the number of points of similar adjustments for A and in our example A stands for temperature. For B (print) counts analogous:

$$SQA_B = \frac{1}{n_{B1}} (\sum Y_{B1})^2 + \frac{1}{n_{B2}} (\sum Y_{B2})^2 + \frac{1}{n_{B2}} (\sum Y_{B2})^2 - CF$$

$$SQA_C = \frac{1}{n_{C1}} (\sum Y_{C1})^2 + \dots\dots$$

and so on.

4) Estimation of variances of single factors as quotient from the squared deviation to the degree of freedom

$$V_A = \frac{SQA_A}{DF_A} \quad V_B = \frac{SQA_B}{DF_B} \quad V_C = \dots \quad \dots \quad \dots$$

whereby DF = number of steps -1 (number of independent settings, which can still be changed starting from a step, in the example $DF_A = 2$).

5) Determination of error variance

In general at examination of experimental designs two types of errors can occur:

F1 = error within a characteristic combination, whereby this should be 0 at corresponding carefulness of execution.

F2 = error at repeating of measurements

The variance of error F2 can be estimated according to following rule: you contract the square total of factors with the least squared deviations, in our example $SQA_{C+D} = 400$. Approximately half the number of DF 's should be used. Thus the error variance results by

$$V_{F2} = \frac{SQA_{C+D}}{DF_{C+D}}$$

6) Calculation of proportion of factor variances to error variance

$$F_A = \frac{V_A}{V_{F2}} \quad F_B = \frac{V_B}{V_{F2}} \quad F_C = \dots\dots$$

7) Determination of significance of the corresponding factors

The prior determined F-value can be compared to a critical F-value. The null hypothesis is set up $F_A > F_{krit}$, there is a significant difference with x % . The critical F-value, e.g. for A you get from F-tables with degree of freedom $f_1 = DF_A = 2$ and $f_2 = DF_{C+D} = 4$ and a level of significance of 95%.

8) Percental meaning of a factor

An important result of the analysis of variance is the percental share of a factor on the target size.

This is determined e.g. for A by:

$$SQA'_A = SQA_A - DF_A V_{F2}$$

$$A_A = \frac{SQA'_A}{GSQ} 100\%$$

The percental share of F2 is determined by:

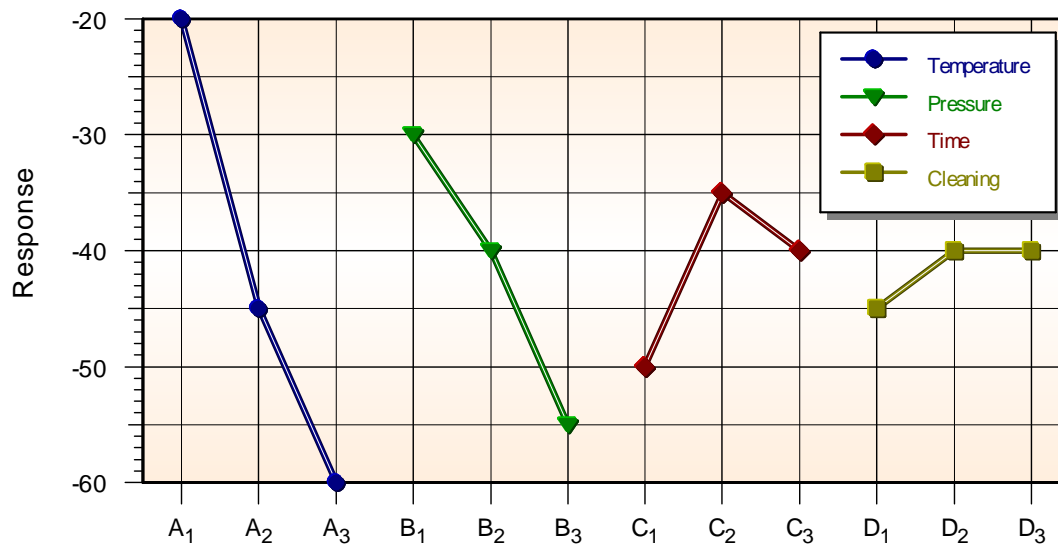
$$A_{F2} = \frac{SQA_{F2} - V_{F2}}{GSQ} 100\%$$

For the example in total there are following results:

	DF	SQA	V	F	SQ'	Percent %	F critical
Temperature	2	2450	1225	12.25	2250	59.2	6.94
Pressure	2	950	475	4.75	750	19.7	6.94
Time	2	350	175	1.75	150	3.9	6.94
Cleaning	2	50	25	0.25			6.94
Error 2	4	400	100			7.9	

Consequently the critical influence is the temperature.

In the so called **ANOM** (Analysis Of Means) the mean values of target sizes of each adjustment of each factor are depicted. For the example described in the ANOVA following description arises:



For the described procedure the submission [ANOVA_Multi.vxd](#) in the directory \Statistics has to be used.

If several measurements are used for each factor adjustment, so the lot fraction defective has not be estimated with the smallest factor shares, but can be determined directly.

First of all a square total is calculated for the error:

$$SQA_{F2} = GSQ - \sum_{i=1}^p SQA_i \quad \text{with } p = \text{number of factors}$$

and the variance is:

$$V_{F2} = \frac{SQA_{F2}}{DF_{F2}} \quad f_{F2} = (ny \cdot n - 1) - (n - 1)$$

with ny = number of repetitions, n = number of trials

The relative shares are determined analogous to the previous approach via:

$$SQA'_x = SQA_x - DF_x \cdot V_{F2}$$

$$A_x = \frac{SQA'_x}{GSQ} 100\%$$





For the ANOVA with repetitions the submission [ANOVA_Multi_Repetition.vxg](#) in directory \Statistics has to be used.

See also: ANOVA between two Series

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 $Y = b_0 + b_1 x_1 + b_2 x_2$	Factors on respectively only 2 steps, min number of tests $p + 1^*$	No nonlinearities and interactions determinable
 Change effects $Y = . . b_4 x_1 x_2 . . .$	Factors on respectively only 2 steps, min number of tests $p + p(p - 1)/2 + 1^*$	No nonlinearities determinable, but interactions
 Square $Y = . b_4 x_1^2 .$	Factors on respectively only 3 steps min number of tests $2p + p(p - 1)/2 + 1^*$	Nonlinearities recognizable. Incl . interactions
 Cubic $Y = . . b_4 x_1^2 + b_5 x_1^3 . .$	Factors on respectively only 4 steps, min number of tests $3p + p(p - 1)/2 + 1^*$	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 less number of tests like vollfactoriell, full orthogonal	Mixing of interactions Unsafe of evaluation
 Plackett Burmann	Derivation from factorial design. Very low number of tests.	Interactions are not fully confounded
 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

 <i>D-Optimal</i>	Very low number of tests, Clear regulation of interactions,	not orthogonal good evaluable
 <i>Mixture</i>	Use of factors whose sum must always amount to 100%	not orthogonal, factors de- pendent 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.

Full-, fractional and Taguchi Experiments

Full factorial

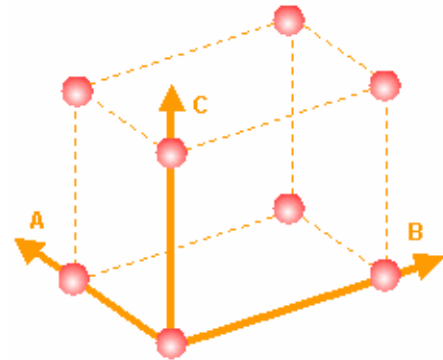
A vollfaktorieller test plan is made if all possible attitudes of the factors are combined with each other.

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.

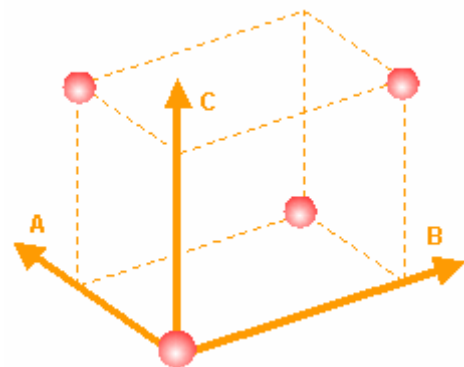
This plan is full orthogonal

Fractional

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 receive 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 build 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:



$n \backslash p$	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

	Fullfactorial	-> all interactions are evaluable
	Fractional plans	-> all two-factor interactions evaluable $\geq V$
	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 Designs

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

Resolution IV Designs

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

Resolution V Designs

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 greater 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. But there are in each case no confoundings 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.

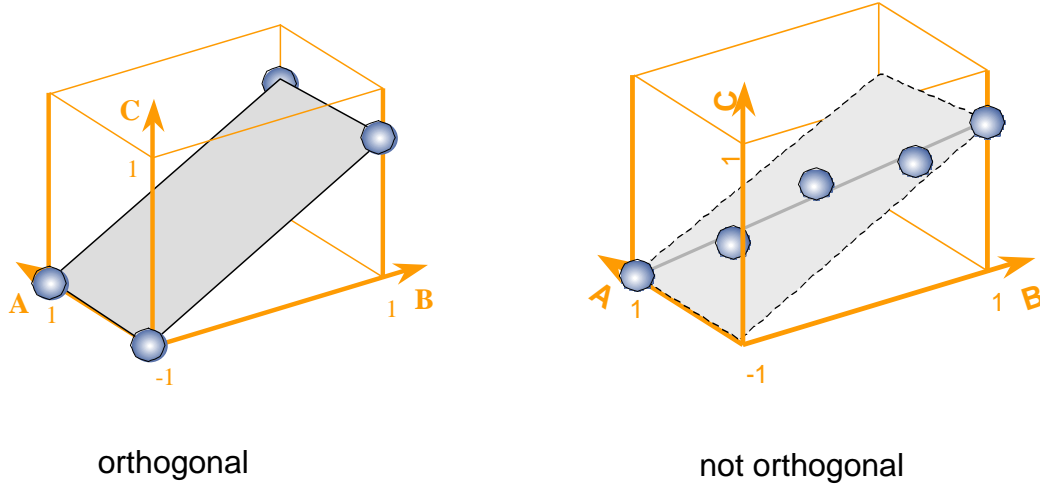
The creation of the plans occurs through the following pattern: A combination order which is repeated column for column around a line down moved is used in each case. The pattern is depending on n:

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 points, 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.



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 L_x in which x is the number of test. These plans are appropriately orthogonal. 2 examples of orthogonal combinations to Taguchi represent the following plans:

$L_4 (2^3)$

	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

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

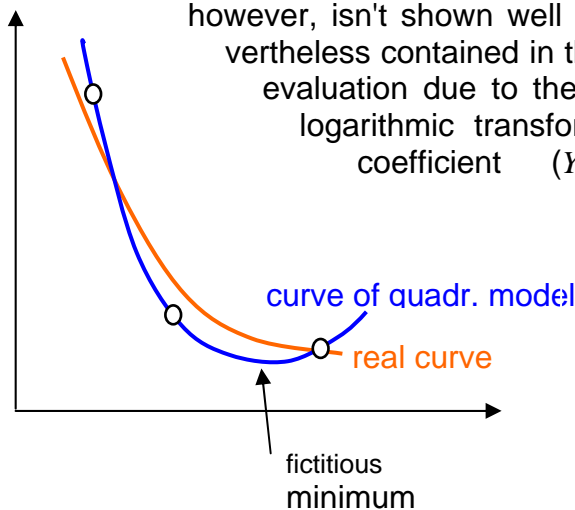
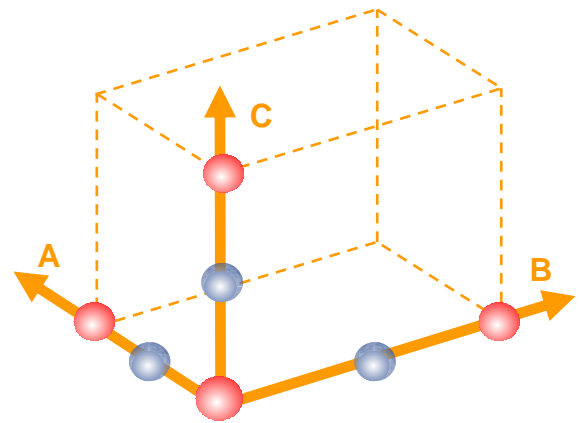
Full-factorial quadratic

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 test 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

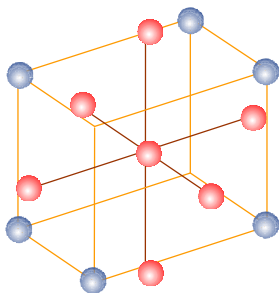
$$(Y = b_0 + b_1 \cdot \ln(x_1) + b_2 \cdot \ln(x_1)^2)$$



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 = b₀ + b₁ · ln(x₁) + b₂ · ln(x₁)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

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

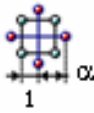
Those plans are also called Central Composite Circumscribed (CCC). Plans with $\alpha = 1$ is also as Central Composite Face (CCF) plans described

A	B	C
-1	-1	-1
-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
1,414	0	0
0	-1,414	0
0	1,414	0
0	0	-1,414
0	0	1,414
0	0	0
0	0	0
0	0	0

full factorial

star

centre

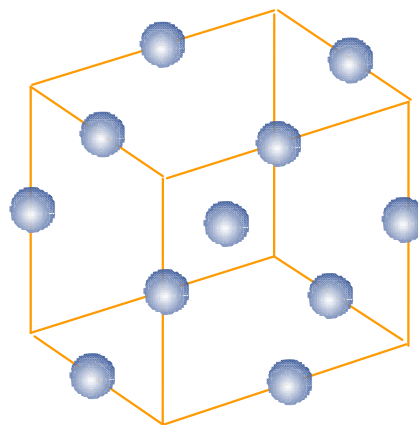


The evaluability of this type of experiment is very good, however, is even bigger than full factorial.

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. Additionally 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 can be advantageous for tests where these extreme combinations are not adjustable.

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			



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.

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 factorial design where interactions are confounded with each other partly.

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

$$p' = p \cdot (p-1) / 2$$

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:

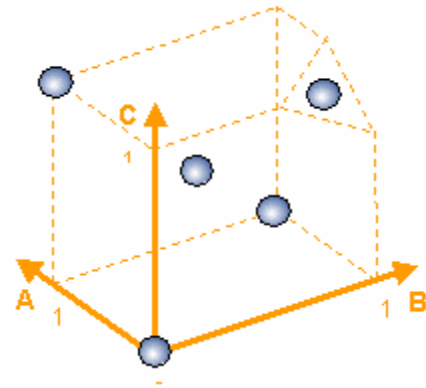
Constant	: 1
Main effects (factors)	: p
Interactions	: $p' = p \cdot (p-1) / 2$
Sum	: $p + p \cdot (p-1) / 2 + 1$

In the case of a square model there are still one time p tests (with a middle attitude). Furthermore gets approx. 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

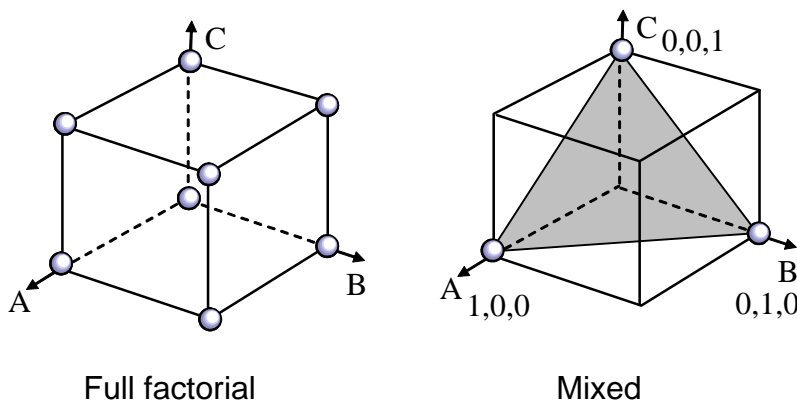
Mixture Plans

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

$$x_1 + x_2 + \dots + x_k = 1 \quad k = \text{count of components}$$

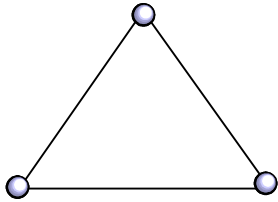
and mean that the components are dependent on each other. This e.g. must be taken into account for the respective tests and can't be treated by standard test plans (only with effort). The possible quota combinations lie in an equilateral triangle.

In most cases there are 3 components. The corresponding test plan looks like represented on the right in comparison with the "conventional" one:



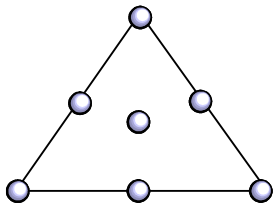
Combinations must be within the range represented grayed. At $k=4$ components = the possible combinations lie in a tetrahedron. Simplexe are called triangle, tetrahedra and the corresponding arrangements at more than 4 components, the mixture plans are therefore also described as a simplex-plans. For 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



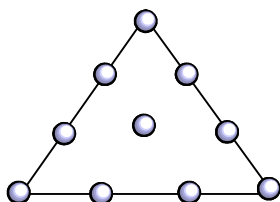
A test plan of the type grade 2 shows the following combinations (in addition with use of all components in the last line):

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

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

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 $x_1 + x_2 + \dots + x_k = 1$ is the reason, however, that some of the coefficients generally approach disappear. But the evaluation can be done via Neural Network anyway.

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:

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

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

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

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

Rank correlation nach Spearman

If the dataset is strongly non normally distributed or if there are categorical 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:

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

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 dataset (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 $r_{x_1x_2} = r_{x_2x_1}$ etc.

	x_1	x_2	x_3	..	x_n
x_1	1.0	$r_{x_2x_1}$	$r_{x_3x_1}$..	$r_{x_nx_1}$
x_2	$r_{x_1x_2}$	1.0	$r_{x_3x_2}$..	$r_{x_nx_2}$
x_3	$r_{x_1x_3}$	$r_{x_2x_3}$	1.0	..	$r_{x_nx_3}$
..	1.0	..
x_n	$r_{x_1x_n}$	$r_{x_2x_n}$	$r_{x_3x_n}$..	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 uncoverd 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:

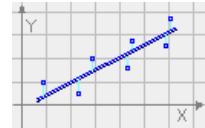
$$t_{pr} = \frac{r_{xy.z}}{\sqrt{1 - r_{xy.z}^2}} \sqrt{n - 3}$$

The hypothesis on independence is rejected, if

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

In Visual-XSel use the menu statistics in the spreadsheet.

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 differences 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 exists.

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:

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

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

$$Y_u = a + b x_i - C \quad Y_o = a + b x_i + C$$

with

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

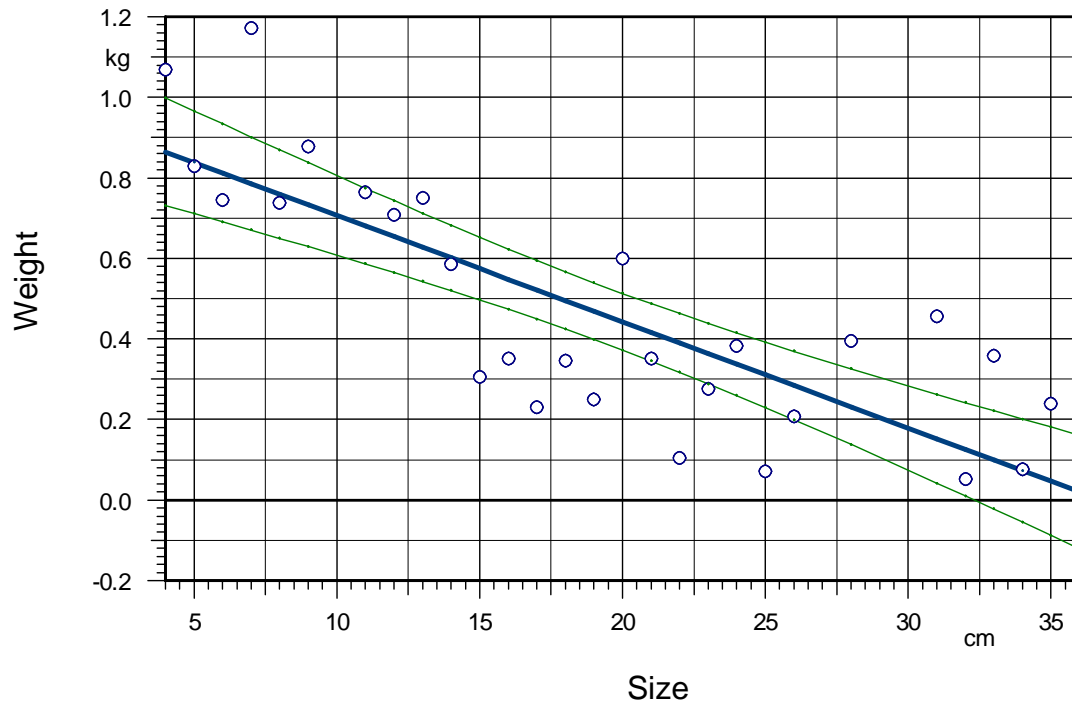
The estimated standard deviation s is calculated from the variance by the deviations of the observations to the compensation straight line:

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

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

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

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

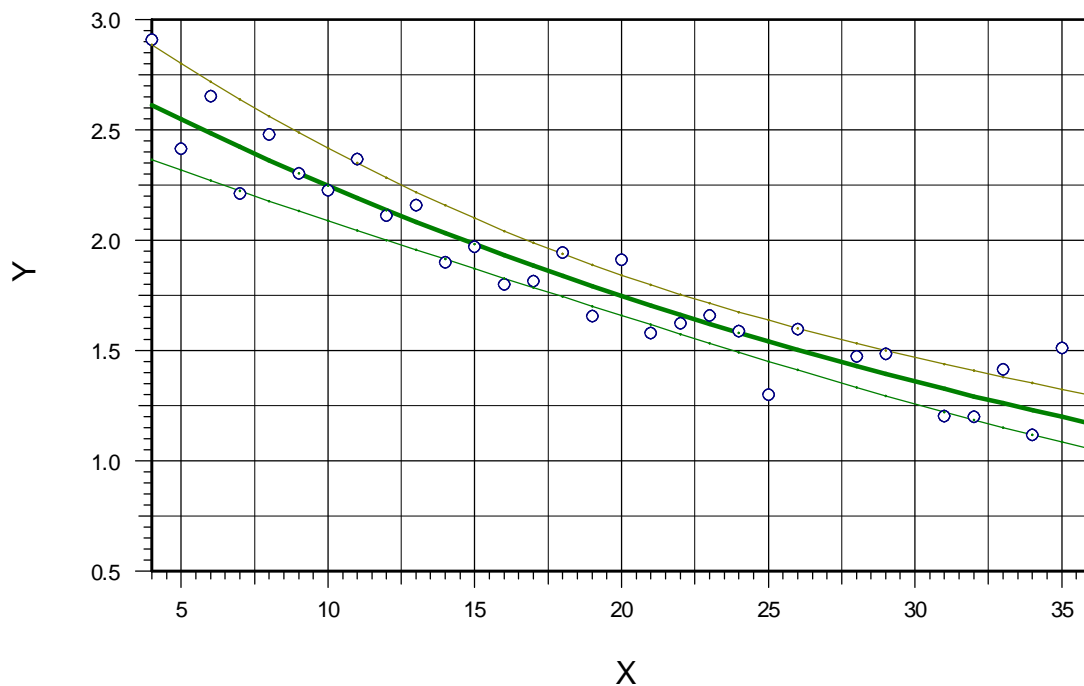
A nonlinear curve is for example $Y = a e^{bx}$. The standard deviation is here:

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

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

$$Y_{\text{unten}} = a e^{bx-C} \quad \text{and} \quad Y_{\text{oben}} = a e^{bx+C}$$

For example:



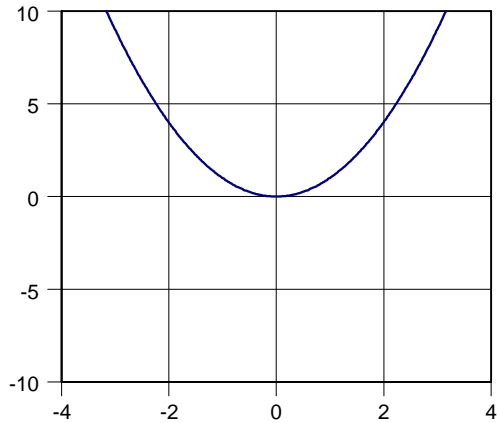
Regression types

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

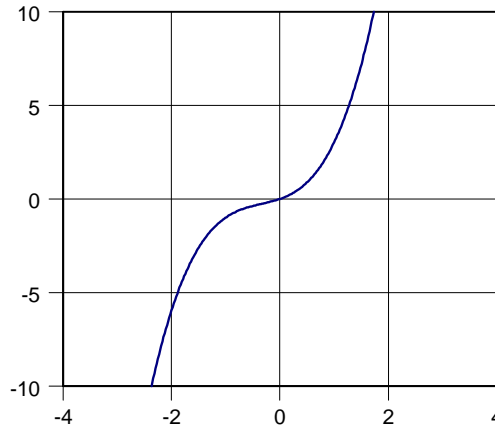
$Y = a x^b$	Straight line in a double logarithm scale
$Y = a + b \cdot x$	Simple straight line
$Y = a + b \cdot x + c \cdot x^2$	
$Y = a + b \cdot x + c \cdot x^2 + d \cdot x^3$	
$Y = a + b \cdot x + c \cdot x^2 + \dots$	Polynom up till 7th grade
$Y = a \cdot e^{(b \cdot x)}$	
$Y = a \cdot e^{(b/x)}$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

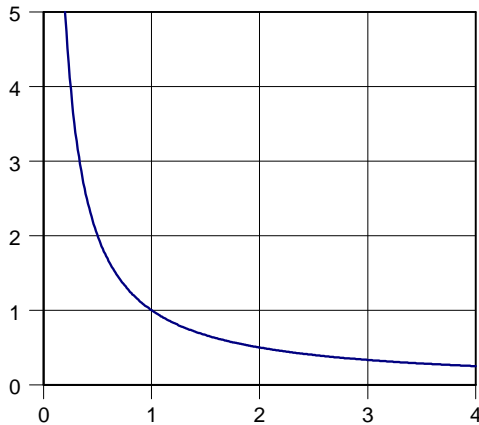
$$Y = x^2$$



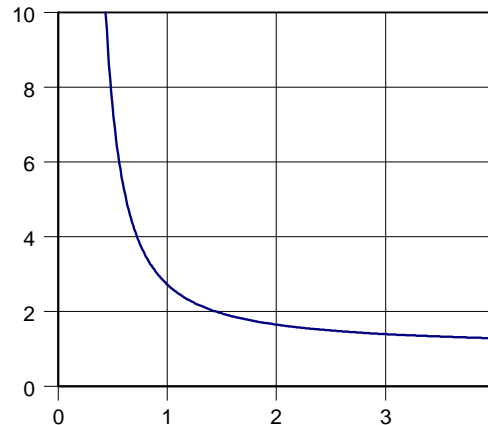
$$Y = x + x^2 + x^3$$



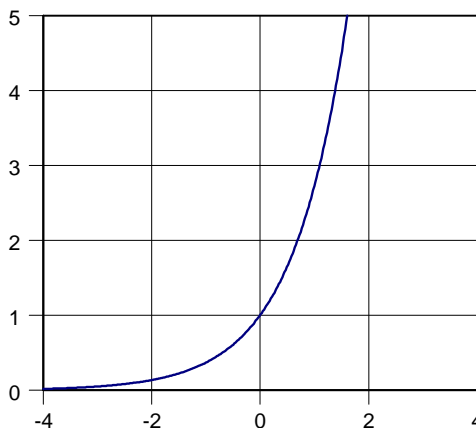
$$Y = \frac{1}{x}$$



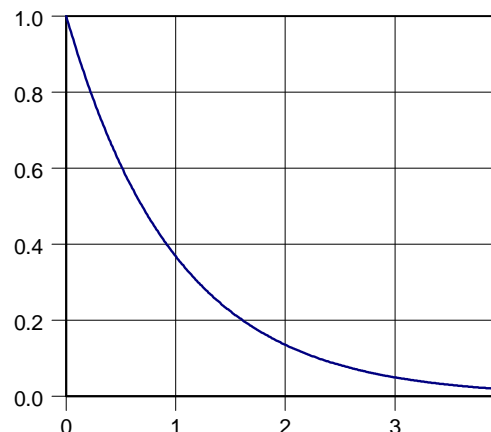
$$Y = e^{-x}$$



$$Y = e^x$$



$$Y = e^{-x}$$



Courses which have a maximum or a minimum happen frequently. An 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 pa-

rabola 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:

$$Y = a + b \cdot x + c \cdot z$$

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

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

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

The functions produced after the regression with concrete coefficients are in the Formula linterpreter 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:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \dots$$

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

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + b_3 x_2 + \dots$$

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:

$$y = b_0 + b_1 x_1 + b_2 x_1 x_2 + b_3 x_2 + \dots$$

The corresponding table columns for x then have to be inserted in a new column as a product x^2 . 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

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{z1} \\ 1 & x_{12} & \dots & x_{z2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{zn} \end{bmatrix} \quad b = \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_z \end{bmatrix}$$

Hint: 1st column represents in X the constant

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

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

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

$$b = (X^T X)^{-1} X^T y \quad \text{arise as follows}$$

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

$$X' = X^T X \quad \text{with} \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{z1} \\ 1 & x_{12} & \dots & x_{z2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{zn} \end{bmatrix} \quad z+1 \text{ columns and } n \text{ rows}$$

The respective cells are calculated after each other:

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

The first column represents the constant b_0 . 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).

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

etc cells

$$j=1 \quad i=1$$

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

$$j=2 \quad i=2$$

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

as a result yields:

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

and the revers matrix is:

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

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:

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

So the equation of the beginning is:

$$y = 6,4 + 1,5x_1 + 2,5x_2 + 0,5x_1x_2$$

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 there fore contains -1 everywhere. The other variations have one in their column 1.

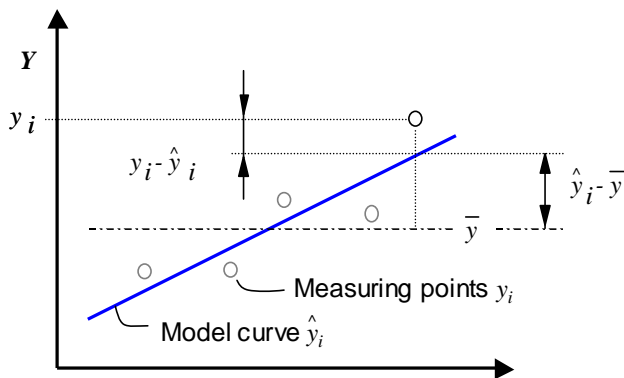
Partial correlations of r have construction caused test plans with categorical factors $r = 0.5$ or more greatly.

Analyses of Variance (Model ANOVA)

For assessment of the regression model the most important index is the coefficient of determination R^2 and then adjusted coefficient of determination R^2_{adj} .

The closer R^2 is to the value 1, the better the model y is described through x . The smaller R^2 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



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

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

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

One frequently also finds the adjusted coefficient of determination R^2_{adj} . 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 $x_1, x_2, x_3, x_1 \cdot x_2, x_1^2 \dots$)

DF_{Res} : Degrees of Freedom of the residuals $DF_{Res} = n - z - 1$
(n = Number of experiments)

DF_{Total} : Degrees of Freedom total $DF_{Total} = n$

For great data sizes are like A and B brought closer. The smaller the data size gets, the bigger the deviation is. R^2 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^2_{adj} indicate unnecessary terms in the model.

Prediction Measure Q^2

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

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

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

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

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

Hints:

R^2 and Q^2 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.

R^2 big and Q^2 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 analysed 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.

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

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

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

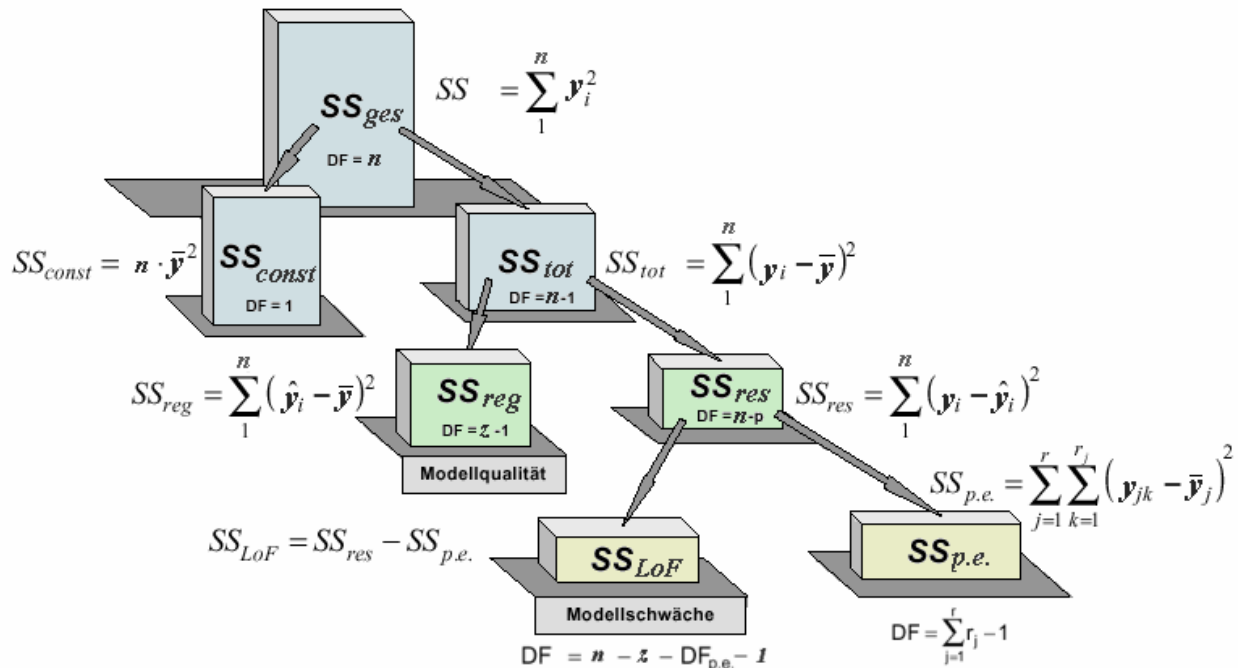
The quotient of the variances is then the Lack of Fit:

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

The result is to compare to a critical F-worth (γ =confidence interval). Obviously if this is bigger then 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. in-sists. A weighted F value is calculated as test quantity:

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

with n number of series 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.

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 fo 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

$$s^2 = \frac{1}{n - z - 1} \sum_{i=1}^n \left(Y_i - b_0 - \sum_{j=1}^z x_{j,i} b_j \right)^2$$

with b_0 = constant term of the model.

X'' is calculated through:

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

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 number of series 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:

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

The relative standard deviation is related to the middle data area

$$RMS / Y_m$$

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 (adjustings) the response value can be calculated to \hat{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

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

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).

Condition Number

The so-called Condition Number is the relationship of the greatest and smallest singular value of the matrix X eigenvalues of $X'X$. This indicator is a measure for the orthogonality of X . All full factorial and fractional factorial test plans have a Condition Number of 1 (without the column 1 with values of 1 for the constant). No central points may be existing, all points lie in the marginal area (see chapter experiments). The Condition Number > 1 is, the matrix is no longer fully orthogonally, i.e. the individual factors have a more or less big correlation under each other what is the case among others at D-optimal test

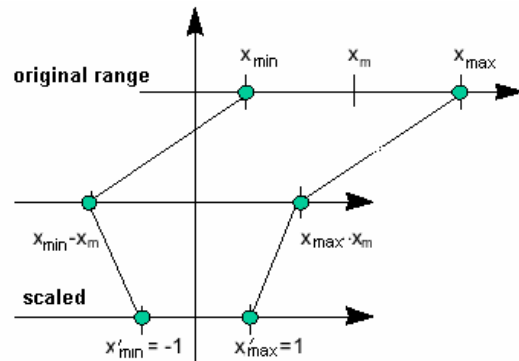
plans.

Regarding the first column with values of 1 for the constant the Condition Number will be a little bit over 1 and therefore the matrix is not full orthogonal:

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 centrally 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

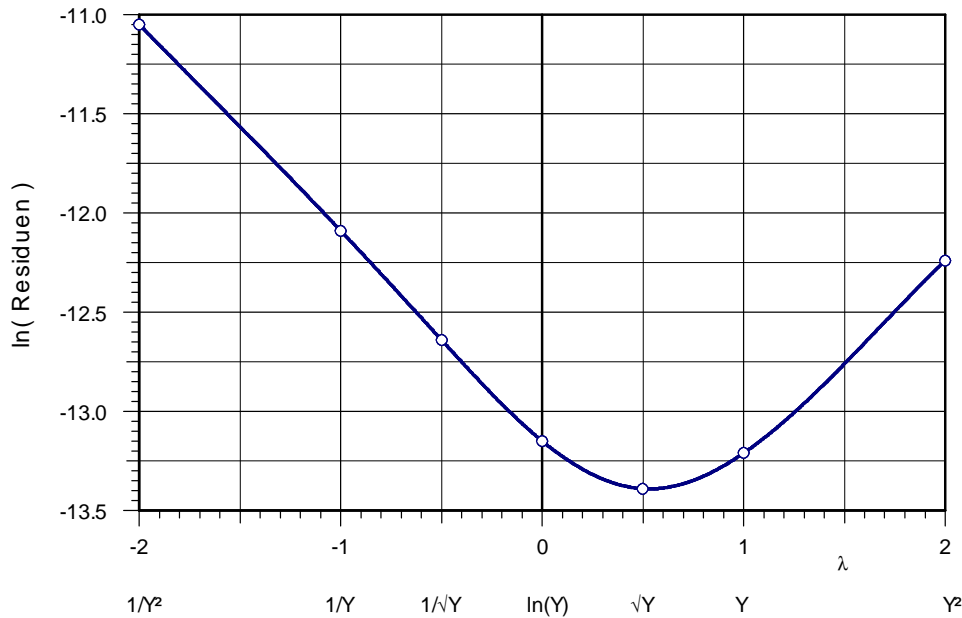
One understands by a correlation a more or less high linear dependence between two variables. The correlation between two factors or between x and y is defined through:

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

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

Response Transformation (Box-Cox)

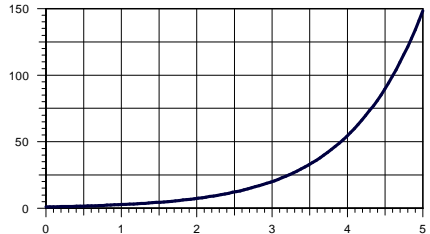
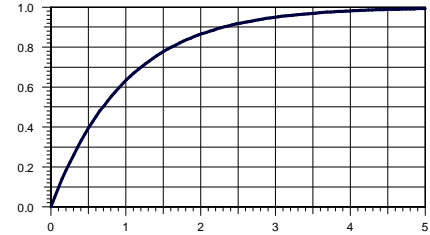
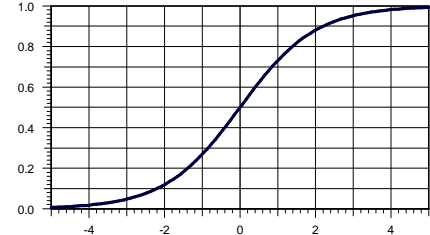
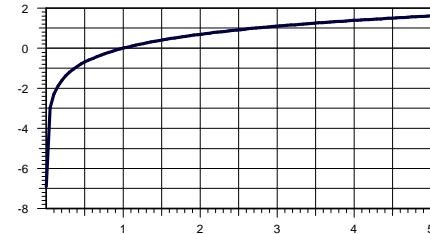
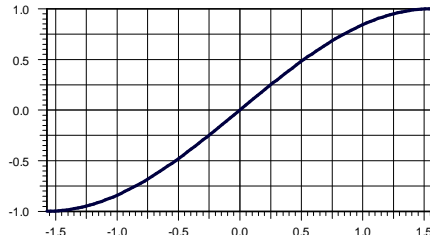
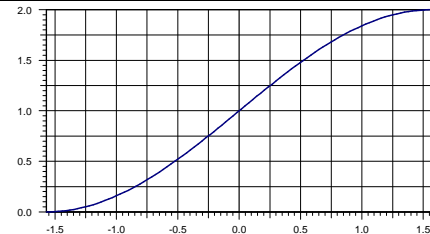
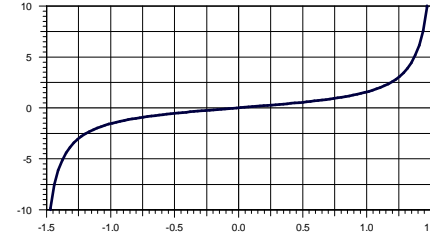
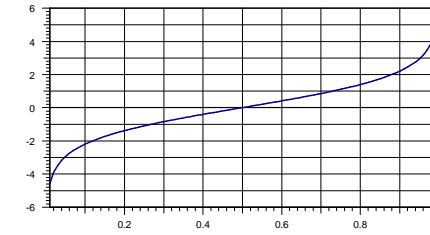
For checking a possibly necessary response transformation the so called **Box-Cox**-transformation is used.



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

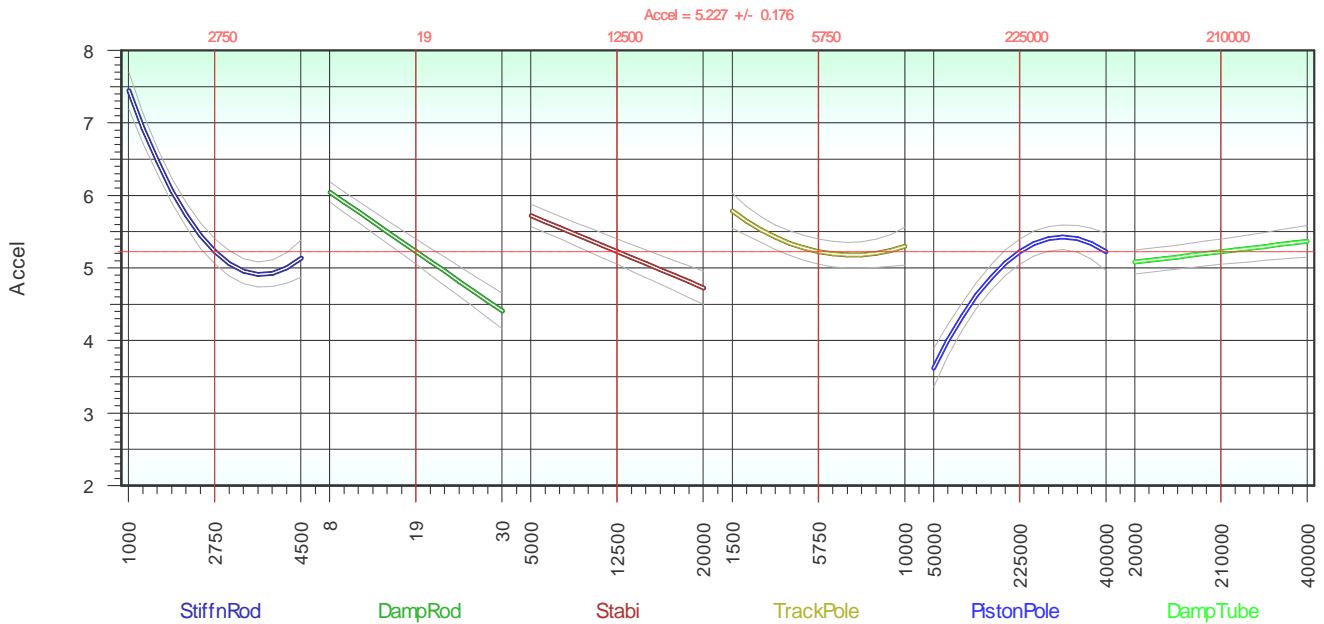
$$Y^{(\lambda)} = \begin{cases} \lambda^{-1} \bar{Y}^{1-\lambda} (Y^\lambda - 1) & \text{if } \lambda \neq 0 \\ \bar{Y} \ln(Y) & \text{if } \lambda = 0 \end{cases}$$

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	$Y' = a'(1 - e^{-b'Y}) + c'$	$Y = \frac{1}{b'} \ln\left(\frac{1}{1 - (Y'-c')/a'}\right)$	
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	$Y' = a' \text{Sin}(b'Y + c')$	$Y = \frac{1}{b'} \left(\text{ArcSin}\left(\frac{Y'}{a'}\right) - c' \right)$	
6	$Y' = a'(1 + \text{Sin}(b'Y + c'))$	$Y = \frac{1}{b'} \left(\text{ArcSin}\left(\frac{Y'}{a'} - 1\right) - c' \right)$	
7	$Y' = a' \text{Tan}(b'Y + c')$	$Y = \frac{1}{b'} \left(\text{ArcTan}\left(\frac{Y'}{a'}\right) - c' \right)$	
8	$Y' = \ln\left(\frac{Y}{1-Y}\right)$	$Y = \frac{1}{1 + e^{-Y'}}$	

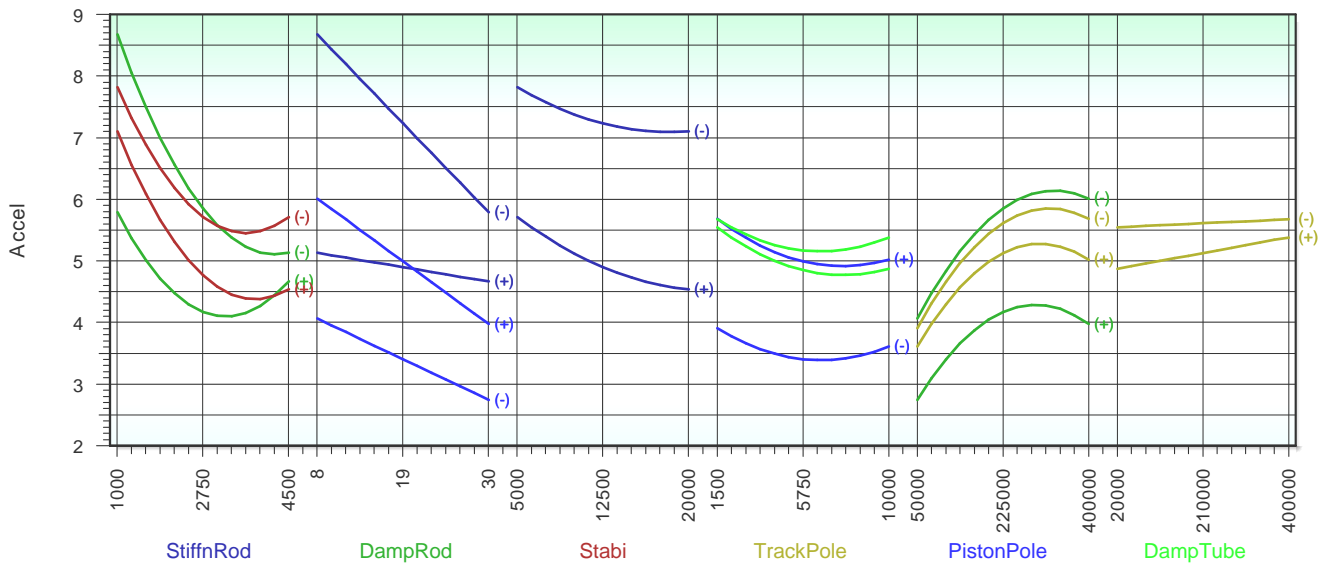
Statistical Charts for Multiple Regression

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.



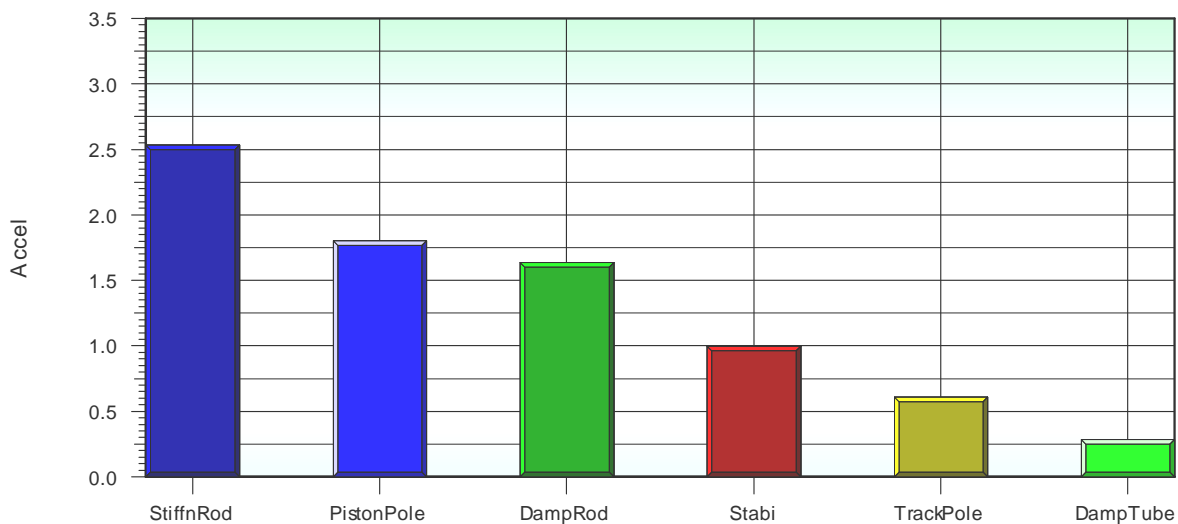
The respective adjustments can be changed by moving these red lines in the graphic with the mouse. By interactions also the other curve linearity's are changed. The horizontal red line always shows the corresponding result value of the target factor. In addition at indication of a lower or upper limit 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 *Interactions-Chart* resembles the curve diagram.

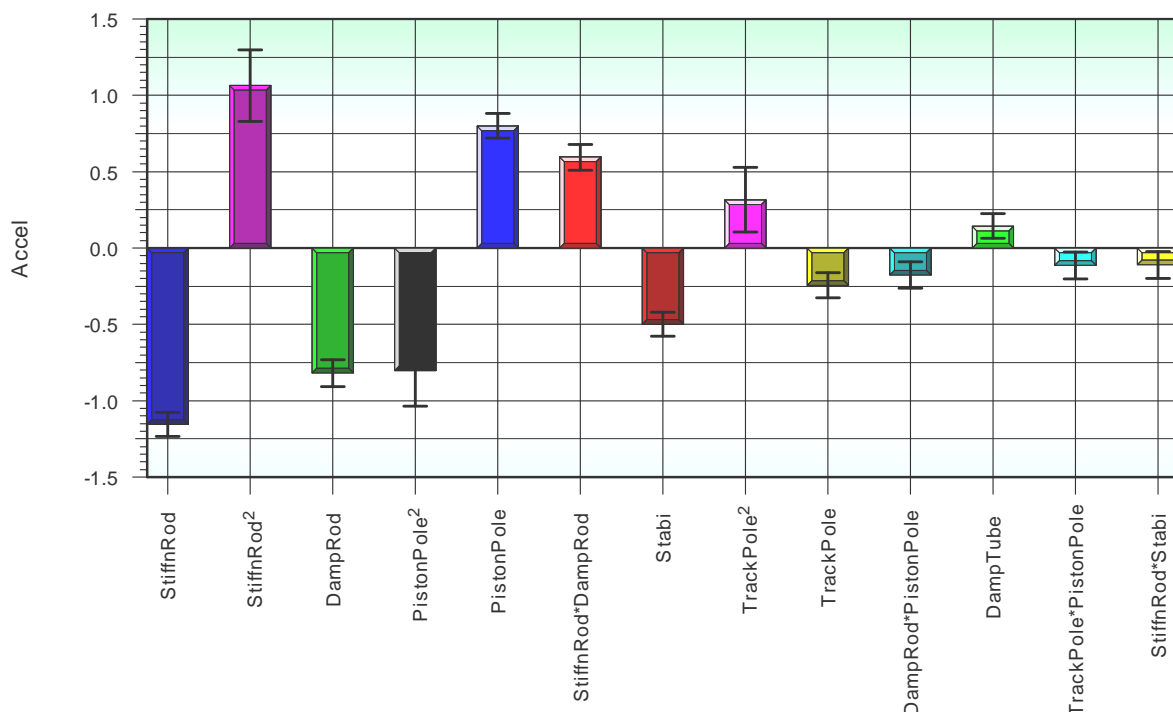


The diagram interactions resembles the curve diagram. The respective curves of curve are represented in pairs here. Every curve couple stands for the respective factors stands with his color with that one of these in an interaction (see color of factor names below the scale). The factor StiffnRod has 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.

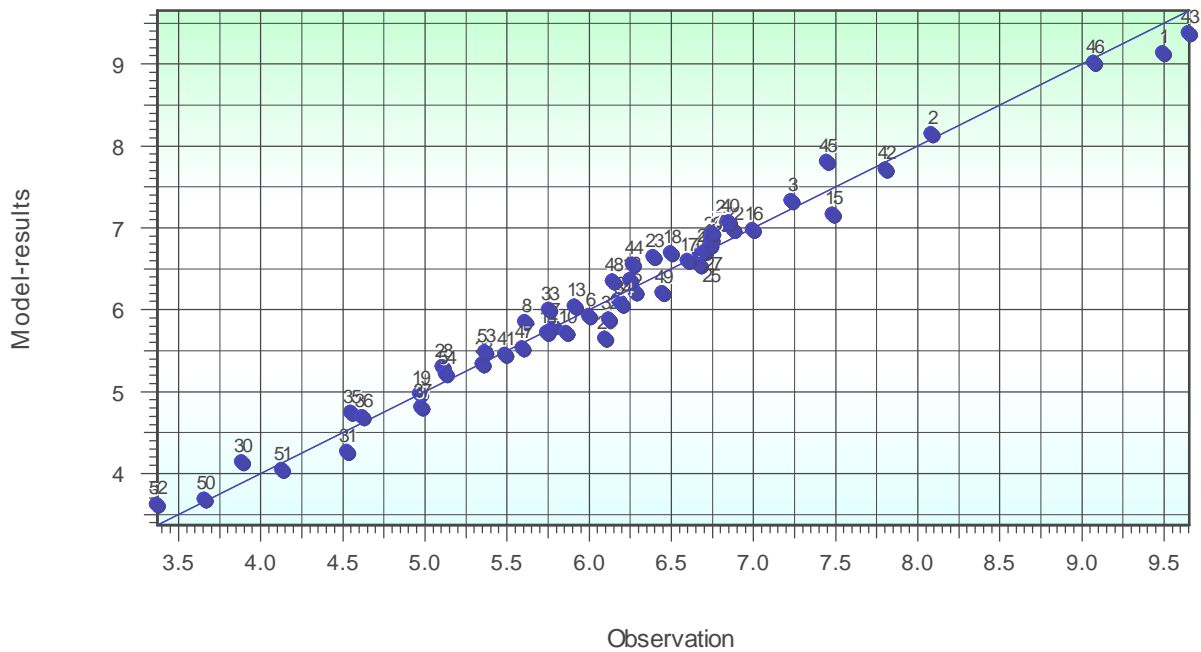
The so called *Effects*, which are depicted in an own diagram, are respectively built from the top and lowest point of curves. Therefore they are dependent of the respectively actual factor adjustments. The effects are depicted as histogram sorted by their absolute size. Here you can recognize directly,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



In the *Pareto-diagram* 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).



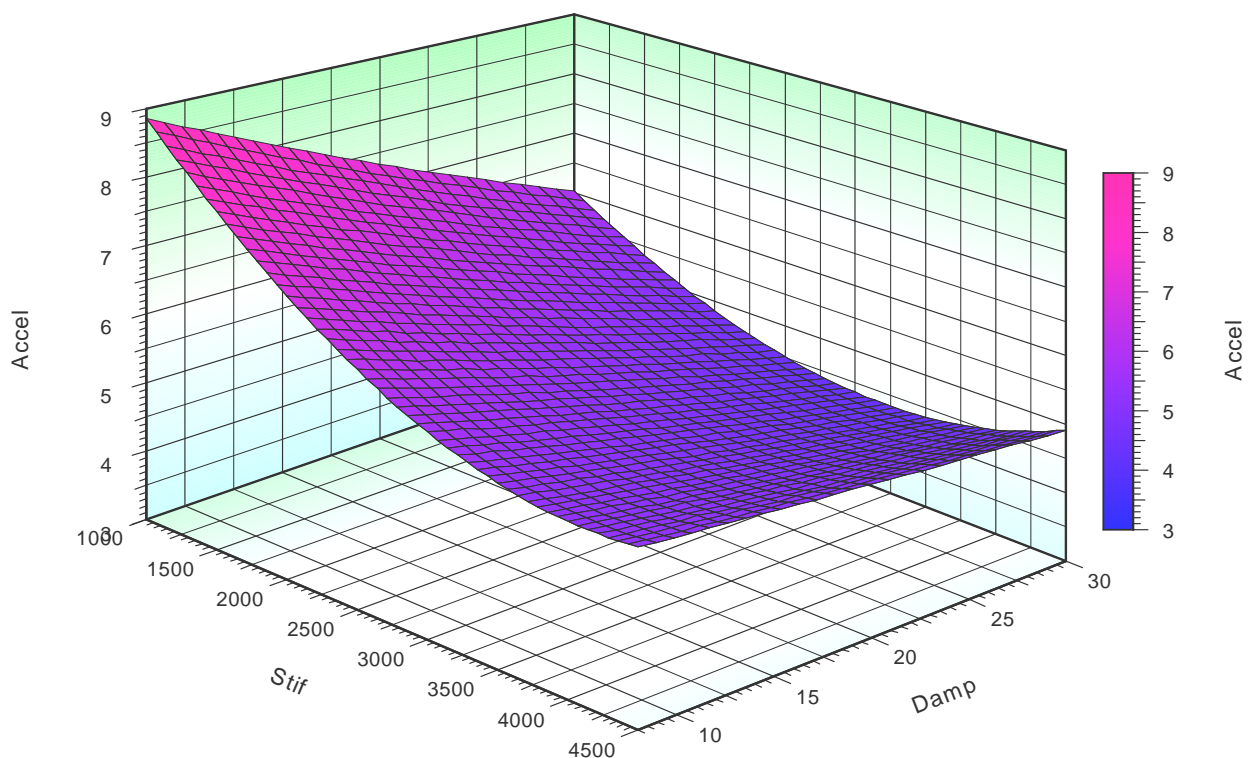
A further important graphic is the *Model versus Observations*, whereas here *Outliers* are displayed, where the respectively point is red instead of blue.



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.

$$\text{Acce} = -1.155 \cdot \frac{\text{Stif} - 2750}{1750} - 0.8191 \cdot \frac{\text{Damp} - 19}{11} + 0.5952 \cdot \frac{\text{Stif} - 2750}{1750} \cdot \frac{\text{Damp} - 19}{11} + 1.064 \cdot \left(\frac{\text{Stif} - 2750}{1750} \right)^2 + 5.227$$



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 ***Dia-gram/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 a outlier the test after Grubbs is recommended. The hypothesis is: x_r is an outlier. x_r stands for the values of the residua, s_r for the standard deviation of the residua

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

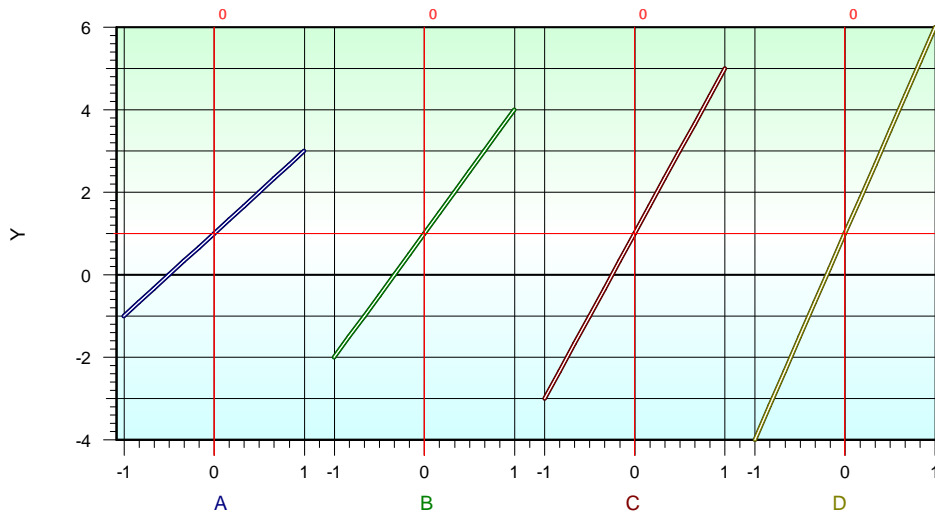
$T_{n;1-\alpha}$ is the critical worth of the Grubbs-Test after the following table:

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

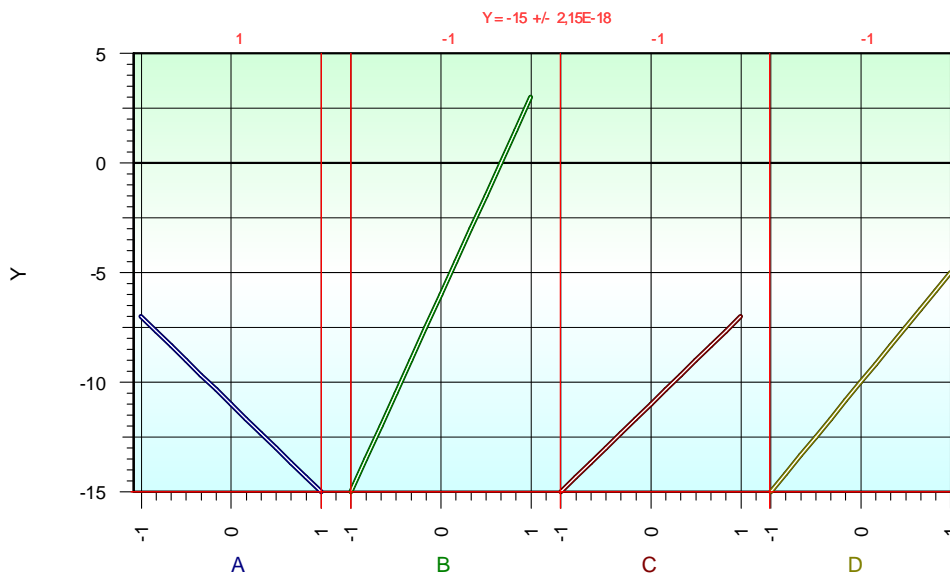
Optimization

One understands by an optimization of regression models finding the right adjustments of all factors for a minima, 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 fulfilment degree 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

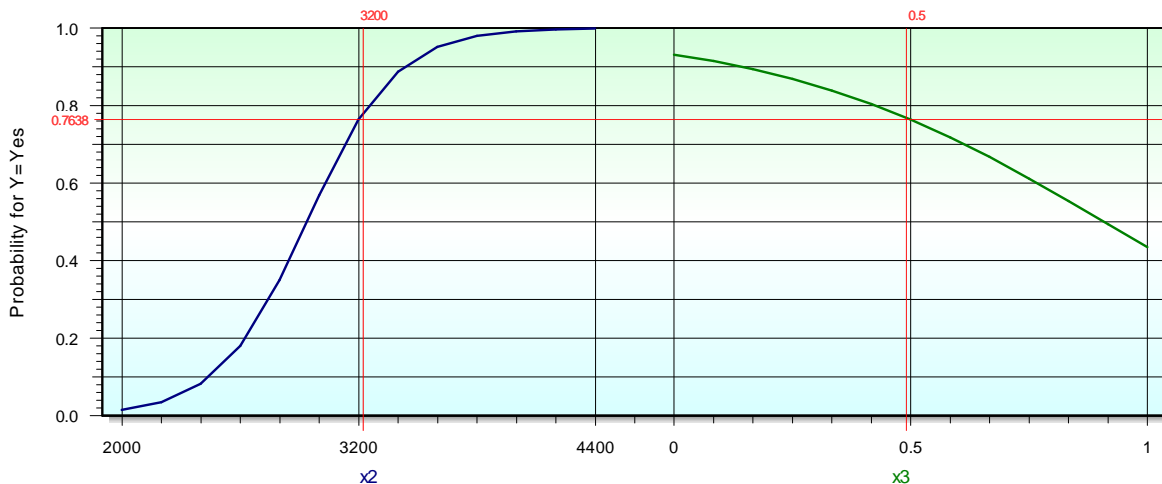
If certain response values have maybe a higher importance than other, this can be taken into account by a weighting factor δ

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 i.O). 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 residua 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.



An 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

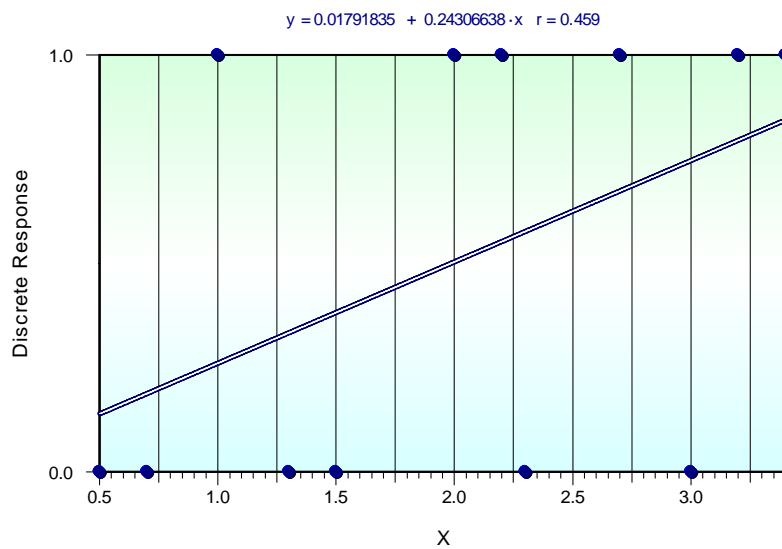
One understands an evaluation by a discrete regression with target values which do not have any steady measurement but qualitative character. The result of an examination could be judged only "well" or "badly", as rip available or not, for example. These statements represent the undermost level of the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

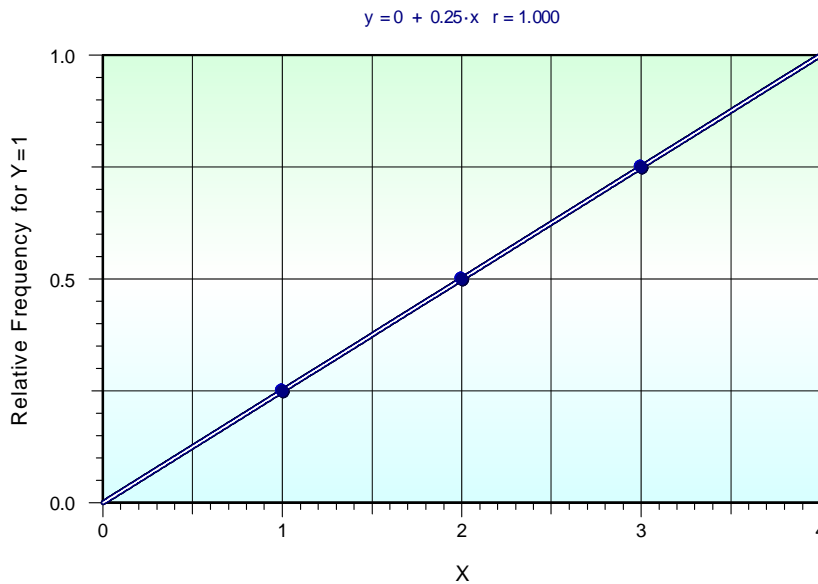
this one for the not satisfactory following regression leads (straight line approximation):



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
$n_x = \text{count } (y=1)$	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 this one makes 12 informations 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 $y=1$ (can not be 0, usually $n_{group} \geq 5$)

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

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

The expression $P/(1-P)$ 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 logarithmiert. 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 not weighted regression till now

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

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 minimised 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:

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

with

$$\delta = \text{diag}(s_1^2, s_2^2, \dots, s_n^2)$$

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:

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

Y_i 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 logarithmiert 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 be determined. The product of the LH function or the sum is charged to the LL with that for every data series. The same must be 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:

$$D = -2LL$$

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:

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

LL_0 : Log-Likelihood of the model, this is only the constant $y' = b_0$

LL_1 : Log-Likelihood of the concrete model $y' = b_0 + b_1x_1 +$

can not 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:

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

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 = $1 - \alpha$ 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 Differenzdevianz is:

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

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

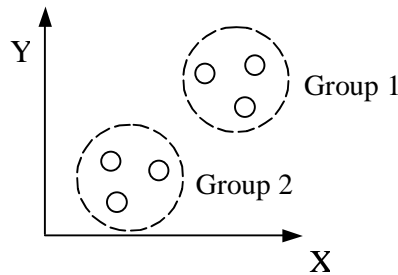
$$X^* = (X^T \delta X)^{-1}$$

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

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

Euklid's distance :

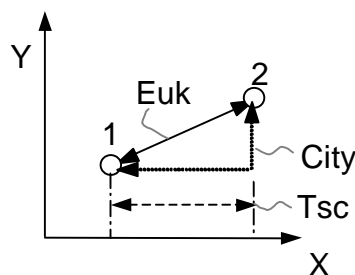
$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

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 $d=\text{ArcCos}(r)$ is used. Respectively higher distances caused through this equation. The similarities can not 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

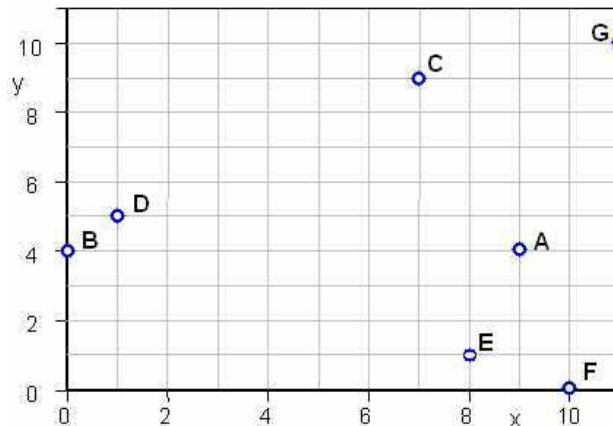
In Visual-XSel there is implemented the hierarchical agglomerativ method.

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

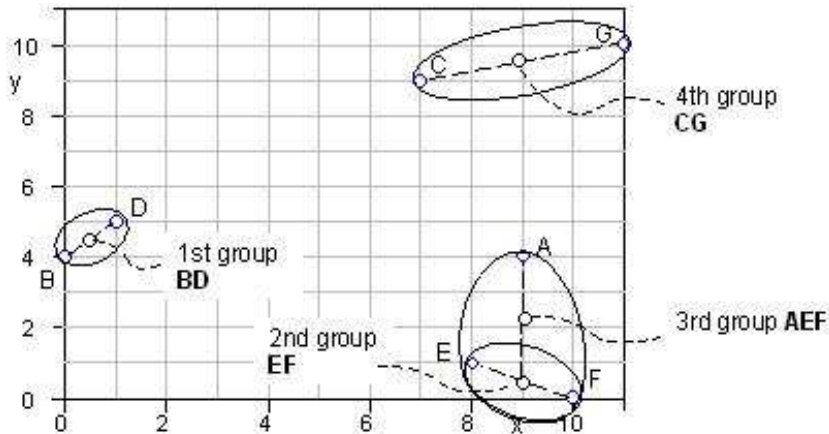


Only 2 coordinates are represented here. n dimensional coordinates (columns) are possible where 3 coordinates can be visualized in a 3D-diagram.

The distance matrix arises from the coordinates. (Values = Euklid'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 this 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 centre can be determined also about the following geometric relation:

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

The results of both variants, however, do not yield exactly the same because the geometric center is calculated is here only a approximation method.

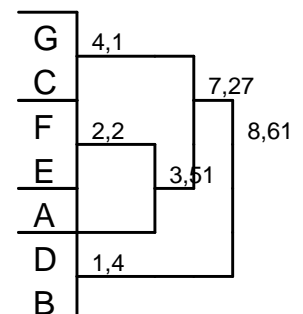
The distance of E and F amounts to 2.2 and therefore represents the 2nd group. The 3rd group is already a combination of 3 points AEF. After every run the complete table must be built up newly.

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	G	G
C	C	C
F	F	F
E	E	E
A	A	A
D	D	D
B	B	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.

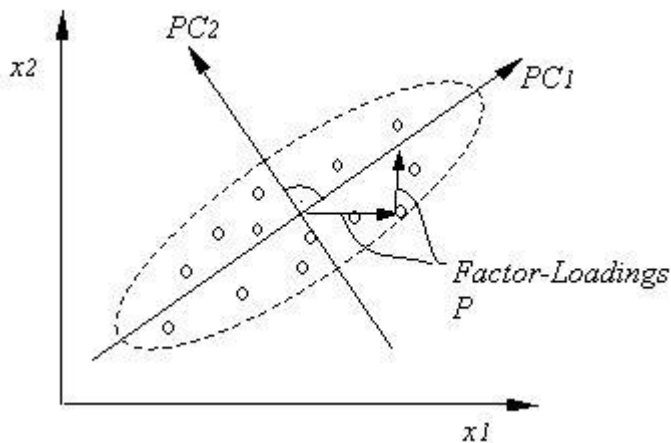
Categorical characteristics can not be defined directly. It is necessary to transform the basis data in a numerical format first. This can be done by producing columns with worth of 1 and 0 to describe the expressions. For example y can be transformed into the following numeric format:

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 Principle Component Analysis calculates new so-called latent variables. These are shortened called factors and represent the **Principle 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 x_1 and x_2 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 x_1 and x_2 . The second main axle stands vertically on the first one and explains the lower share of the variance. Therefore one also describes the principle components as eigenvectors.

For the determination of the principle 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 x_1 and x_2 . The dimension of the factor loadings is \rightarrow number of components \times number of variable x . The Score values T describe the projections on the main axes for every point. The dimension of T is \rightarrow number of component \times number of measurements. The connection is in matrix notation:

$$X = T P^T$$

The following condition applies to the factor loadings:

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

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

$$PC_i = \frac{t_i}{\sqrt{\lambda_i}}$$

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 can not 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 :

x_1	x_2	x_3	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 correlates 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.

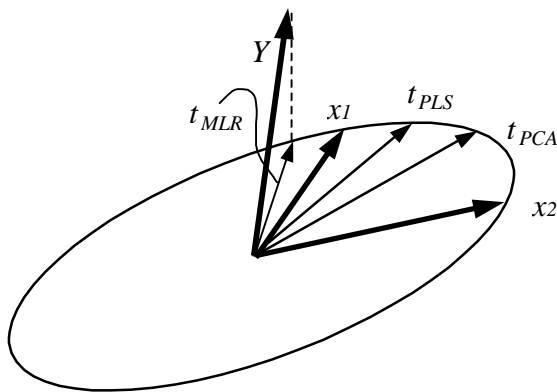
$x_3: F$	-0,958
$x_2: F$	-0,881
$x_1: F$	0,881
$x_1: x_3$	-0,800
$x_2: x_3$	0,800
$x_1: x_2$	0,600

It has to be taken into account here for the interpretation that the factor correlates with x_1 positively and with variable x_2 and x_3 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 exists in PCA. Also here the following condition applies to the weights:

$$w_1^2 + w_2^2 + \dots w_k^2 = 1$$

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

$$= \frac{\sum_{j=1}^z \text{cov}(y, x_j) x_j}{\sum_{j=1}^z \text{cov}(y, x_j)^2}$$

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. 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 than R^2_{MLR}
- Coefficients of PLS are less than MLR-> Errors have a less effect through this.
- PLS maximizes the covariance between the principle 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 $(X^T X)^{-1}$ 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:

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

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 has shown, that $VIP < 0.5$ are not unusual for important variables

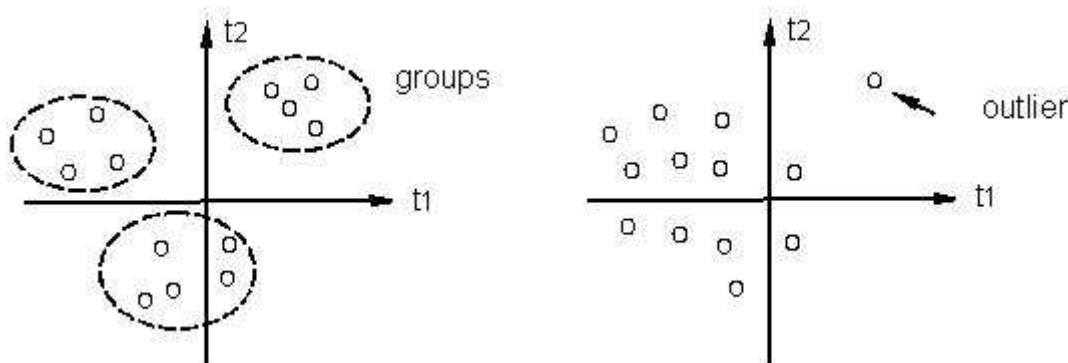
If there is the question whether a variable should be left out from the model, the coefficient size also has to be taken into account. Also the technical connections should be considered.

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 t_1 and t_2 . Possible samples and characteristics in common can be recognized. Also outliers can be recognized.

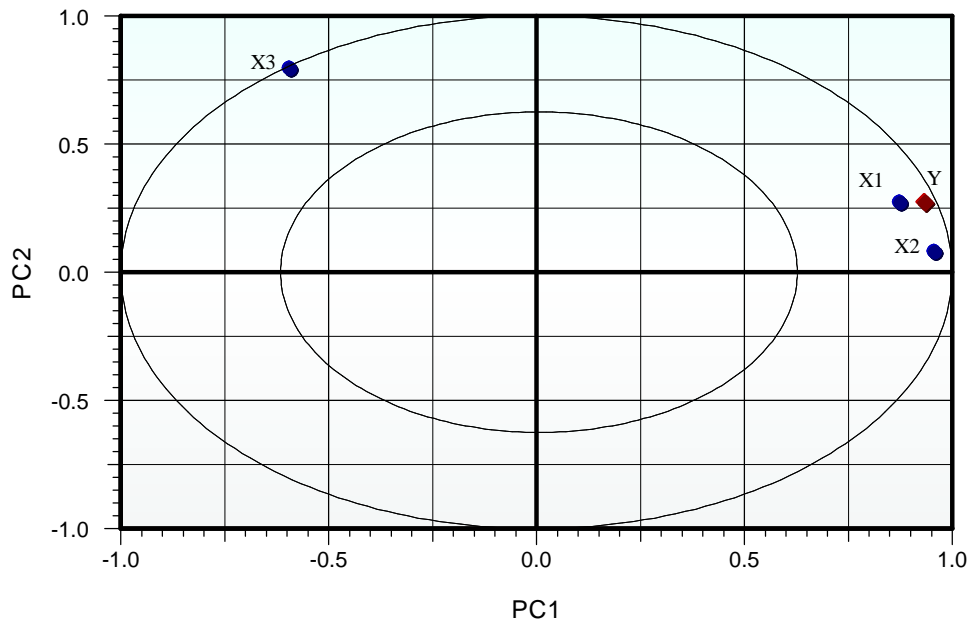


Correlation Loading Plot

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 are scaled as correlations, so it is: professed variance = correlation².

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.

Neural Networks

The fascination starting out from neural nets consists that they are able to solve problems of high complexity with simple means in some cases. The nerve cell represented in simplified terms the biological equivalent. (one guesses that about 100 billion neurons or nerve cells are responsible for the information processing and storage in the brain).

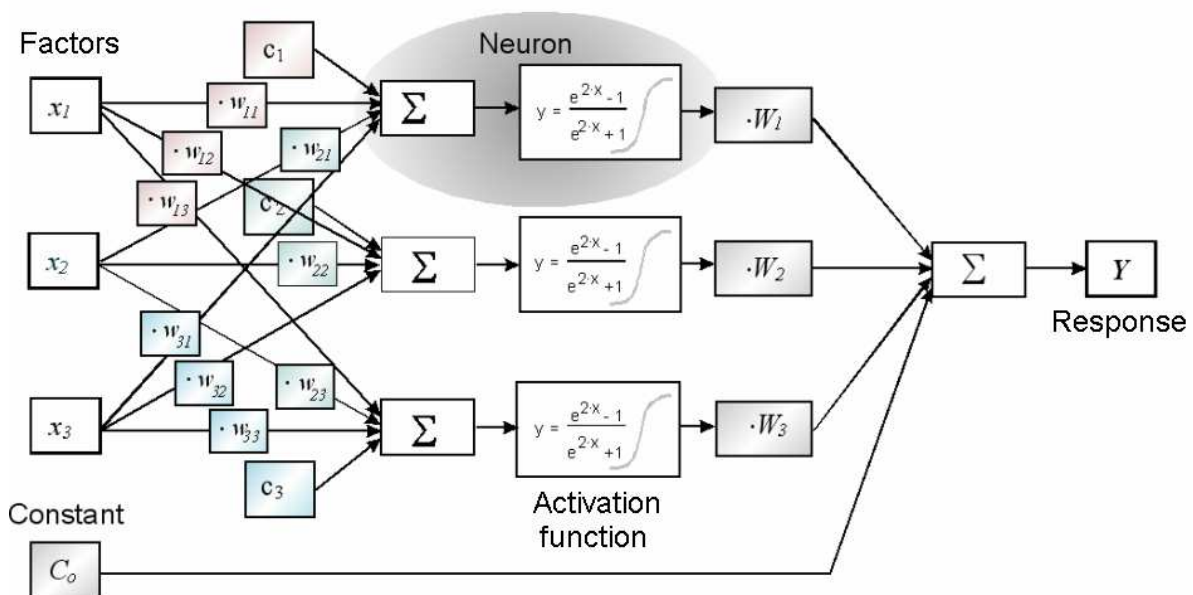
It is necessary in regard on a problem definition to be examined for neural nets neither to carry out more exact examinations nor to represent a formal model explicitly. No effortful algorithms are needed, there is merely the task to approximate data. Doesn't pass any guarantee for the training success either and the solutions can be different, though.

Depending on number of neurons the NN almost exclusively represents interactions. In principle nonlinear connections can be included. It is the advantage of the NN to be able to produce almost arbitrary curves of curve with several maxima and minima primarily while the relatively simple polynomials of the multiple regression can show at most cubic functions. NN is therefore also often used where e.g. characteristic maps shall be calculated by engines.

A found model only applies to the current attitude of the respectively other factors. Influence on the courses of curve of the others has a change of one or several factors in the high measure. One knows this behavior at the multiple regression also in connection with interactions, this is sometimes confusable. It is therefore more difficult to declare connections here.

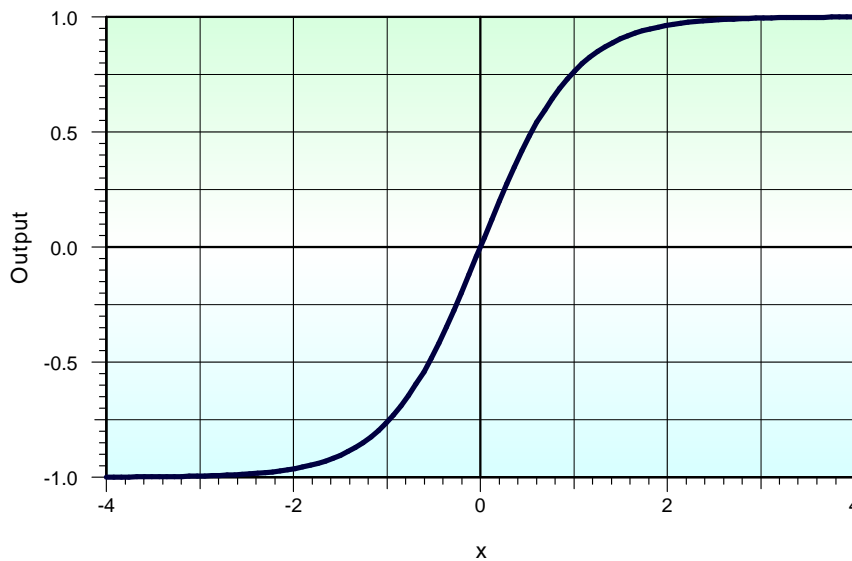
Topology

Example with 3 Factors



The so-called neuron has essentially an assignment or activation function. This is the math-function tanh in most cases. The output of the neuron is between -1 and 1.

$$y = \frac{e^{2 \cdot x} - 1}{e^{2 \cdot x} + 1}$$



With the weights W the quantitative effect of the neuron is fixed. On the initial side every neuron of every factor has also weighted entrances as well as in addition a constant value C over cross. The general model is:

$$Y = C_o + \sum_{j=1}^k \left(W_j \tanh \left(c_j + \sum_{i=1}^z x_i w_{i,j} \right) \right) \quad \tanh = \frac{e^{2x} - 1}{e^{2x} + 1}$$

k = number of neurons z = number of factors

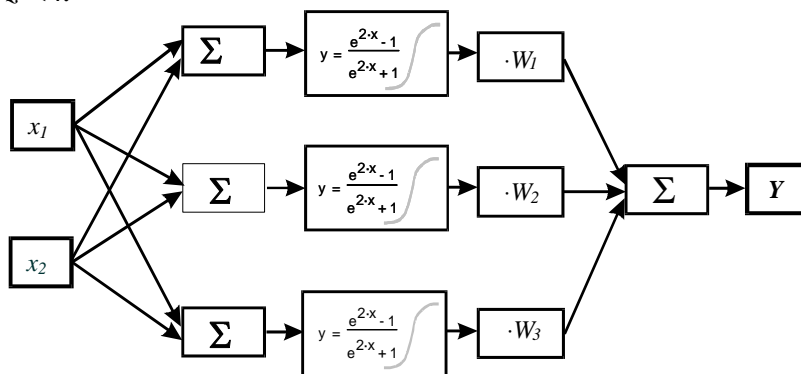
In principle, the factors x_i are standardized between -1 and +1

$$\text{Number of parameters} = z \cdot k + k \cdot 2 + 1$$

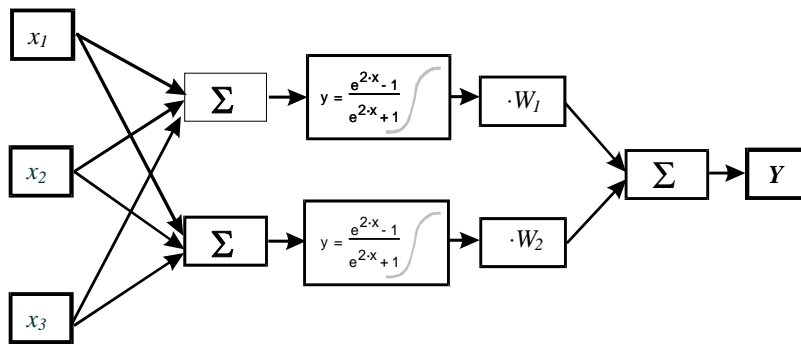
The output of the represented neurons can be inputs of wider neurons again. One calls this layered nets which shall not be given further treatment here.

The number of neurons doesn't have always to be so large like that one of the factors. Besides arbitrary further combinations the following representations are possible:

$z < k$



$z > k$



As a rule, a higher number of neurons as factors has, however, the disadvantage of the over-fit. The results can get relative discontinuous
 The most suitable model is dependent on the facts and can't be generalized.
 Besides the activation function tanh one also finds the following functions:

	Function	Name
	$Y' = x'$	Linear
	$Y' = \text{sign}(x')$	Signum
	$Y' = \text{tanh}(x')$	Sigmoid
	$Y' = e^{-(1/2 * x')^2}$	Normal-distribution

Exactly as in the case of the multiple regression there is to consider knowledge about the facts, which are to be examined. The corresponding activation function also has to be used for purely linear connections. E.g. the signum-function has to be chosen for switching processes or digital statements.

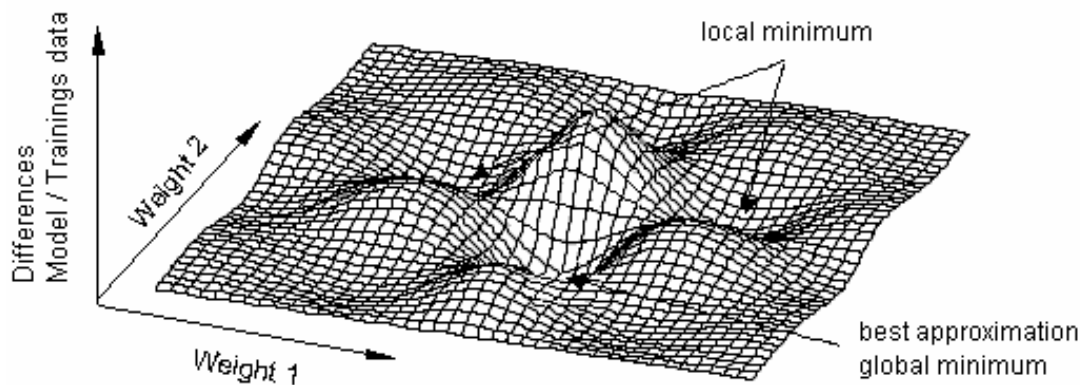
Training-Algorithm

Unlike the multiple regression a clear analytical solution isn't possible by means of matrices here. The coefficients or weights are determined by a training algorithm rather iteratively.

Meaningful start values for the training are:

- $w_{i,j}$: ± 5
- c_j : ± 2
- W_j : $\pm (Y_{max} - Y_{min})/2$
- C_o : $(Y_{max} + Y_{min})/2$ (is fix)

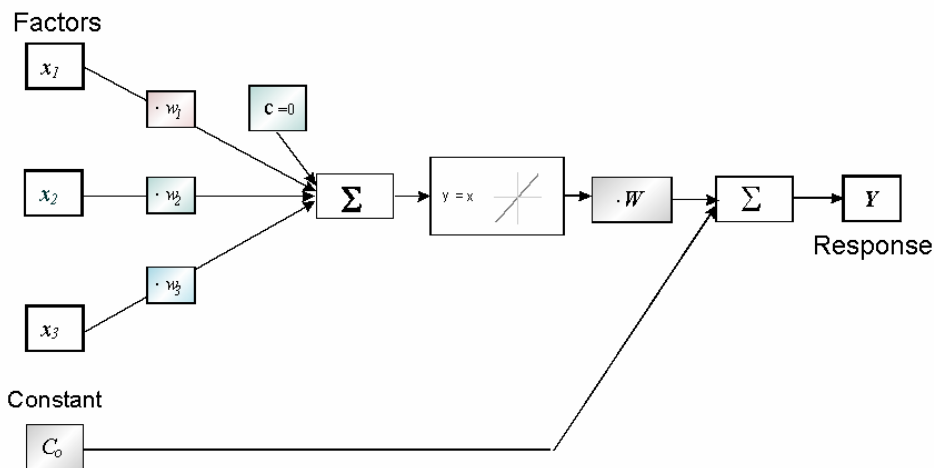
There often are several roughly equal solutions to the approximation of the available connections. However, there is a high risk to find only a local optimum. The model parameters, as in the case of the multiple regression, are sought. Instead of the polynomial coefficients the parameters the so-called model weights and the constants are here. The training process aims minimize the square sums between NN model and data (leaves square method) and is nothing else but an optimization method. These are described in detail in the literature. Methods which take into account the gradient of the object function for the search for the optimum are judged particularly well. It is a general problem that the training algorithm doesn't stop in a local minimum but searches the complete parameter room for the global minimum. The following picture shows the deviations for two parameters (weights) for the training data. This is necessary to minimize. Depending on start condition, perhaps a local minimum is found and this suggested as the solution. If there are no further investigations to search for other minima, not the best one will be found.



Depending on number of neurons the computation effort can increase considerably. A complete variation of all parameters is no longer feasible as of a certain number of factors and neurons. To prevent that one finds only a local minimum, several different "start points" are used for the weights at the training. To this several start points must be selected in sufficiently large number by randomizer. After an abort condition to be fixed the next optimization step is calculated with the weights of the best customization. These optimizes the best point adjusts iteratively for every weight. It happens, that in the practice there are much more or less equal solutions. I.e. the approximation to the data on hand many local minima in the range can be reached well by completely different weights equally (the global minimum).

Neural Network as an alternative for multiple regression

In principle, the neural network of the multiple regression is different by the networking and particularly by the activation function. One chooses the linear activation function for this then we will get the same as a polynomial of the regression:



The represented example of 3 factors corresponds to the model $C = 0$ and $W = 1$:

$$Y = x_1 \cdot w_1 + x_2 \cdot w_2 + x_3 \cdot w_3 + C_0$$

c and W having to be set in the definition of the start weights, so that they aren't further varied.

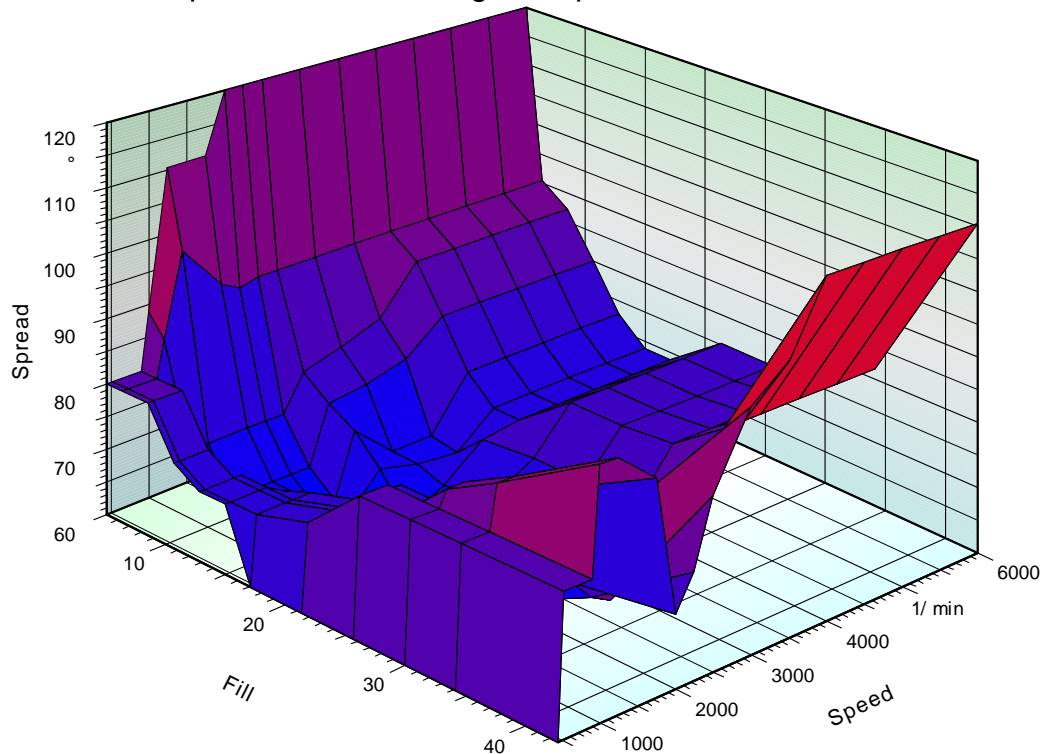
Interactions and square terms can be formed by pseudo-factors (e.g. $x^4 = x_1 \times x_2$ etc.). The training of this simple network must cause the equal clear result of the multiple regression. The weights then correspond to the coefficients. It would be the advantage of this procedure, for example, that certain coefficients are already known and become as "extremely" respected for technical or physical reasons. One then sets these values for the start weights as fix, and only looks for the remaining coefficients (weights). This isn't possible with the multiple regression.

Attributes of Neural Networks

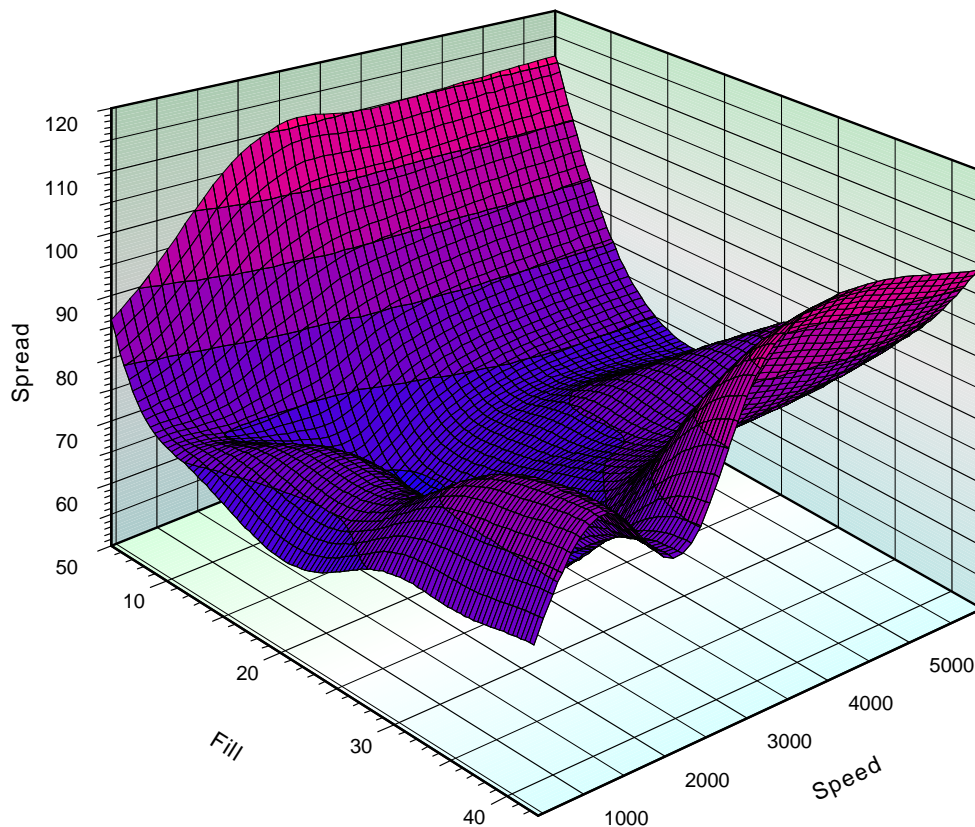
- A Neural Network is a dynamic system which is able $y = f(x)$ execute, an arbitrary function.
- The structure of a Neural Networks is defined by the number of elements, by the special topology and by the way of the activation function.
- One doesn't program a Neural Network but one trains it. Instead of a rule or a general algorithm the training demands a certain number of data.
- The "knowledge" of a Neural Network (longtime memory, program) sits in the totality and in the activation threshold of the elements. No clear symbolic form has this knowledge but a chaining is of factors, constants and weights whose connection cannot simply be recognized in comparison with the model of a multiple regression.
- There isn't any difference between hardware and software in the Neural Networks: It can be considered independent machines or has been simulated via software as a model.

Example

For example in an engine control the following characteristic map for an inlet spread of a camshaft. The parameters are filling and speed:



The normal regression cannot represent this even with a cubic model. The Neural Networks which offers more "degrees of freedom" to the approximation offers its services here. Due to the many discontinuities it will be necessarily to use a higher number of neurons as parameters. 10 neurons were used for the customization represented below:



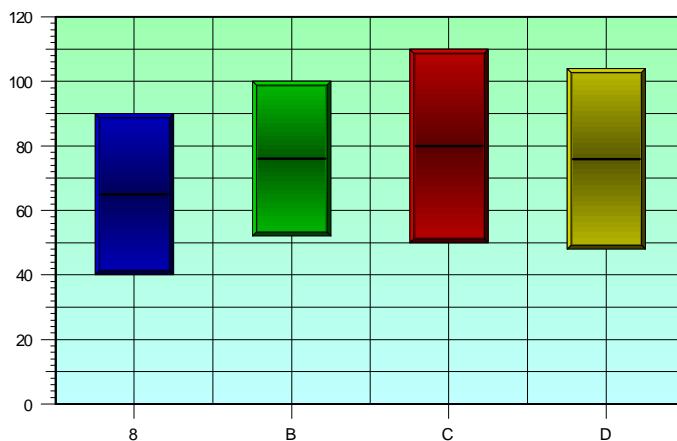
Further statistical charts

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

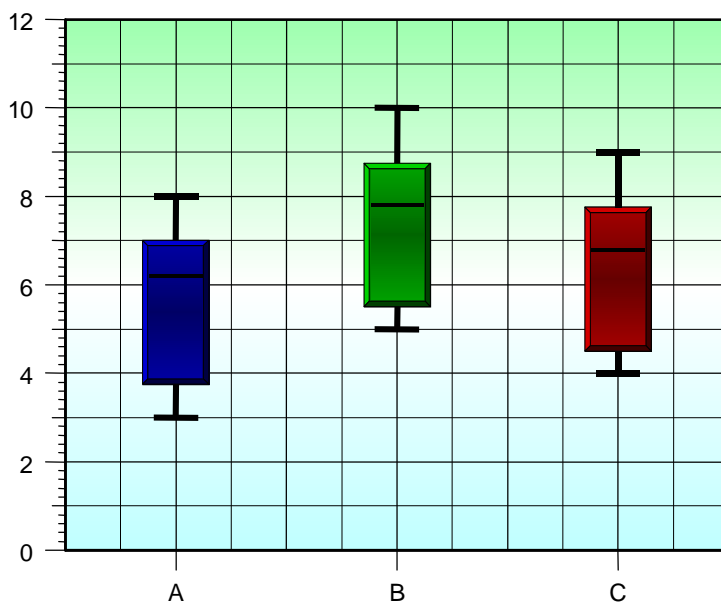
The boxplot is a special type of frequency scale. Here the values are depicted via the y-axis instead of the x-axis, whereas several boxplots in parallel are possible in one diagram. In the middle of the boxplot there is a line with the so called center-value resp. median. Optional also the median can 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

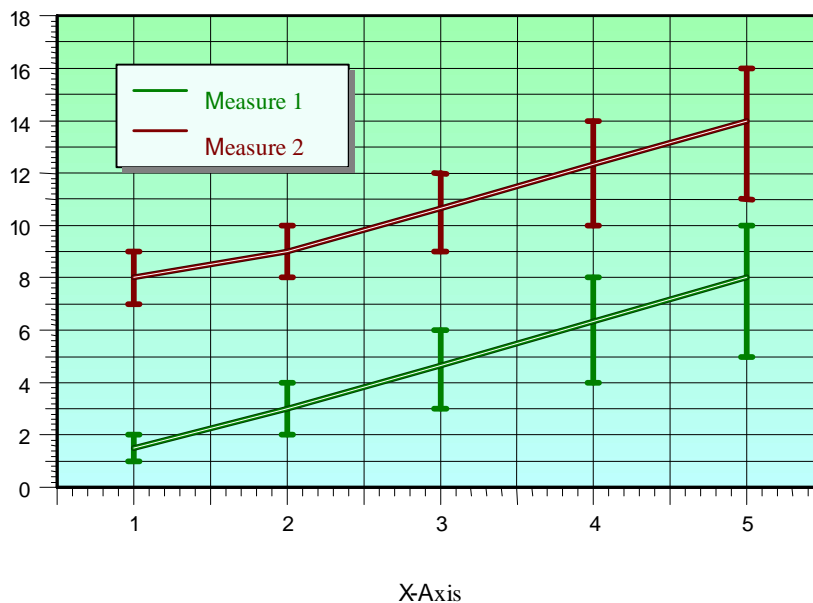
Median plot

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	1.5	2	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

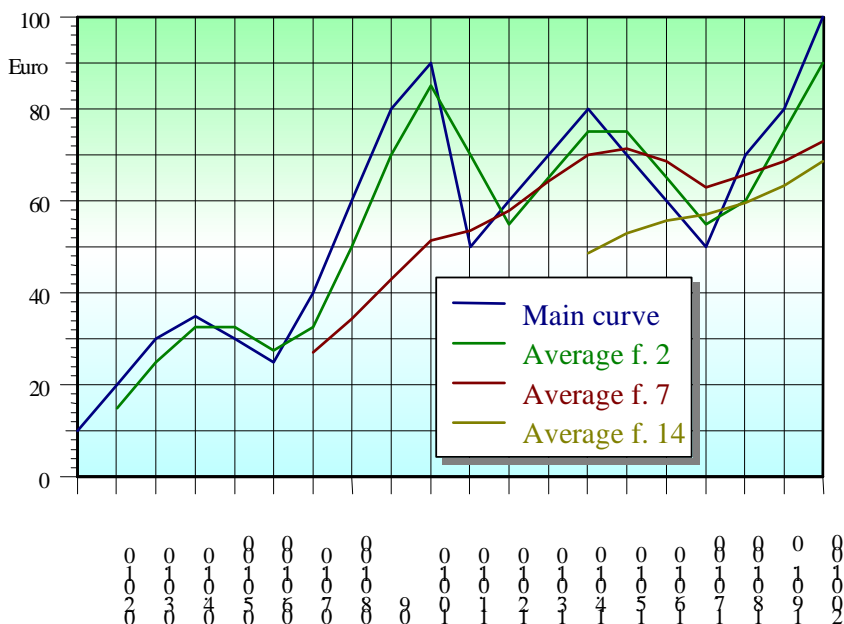
The illustration of gliding average corresponds to the line diagram with at first one „main curve“. The reference to the X-axis stands in the first column of the table. The data of the main curve (Y-values) are in the proximate column.

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 from 14
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:

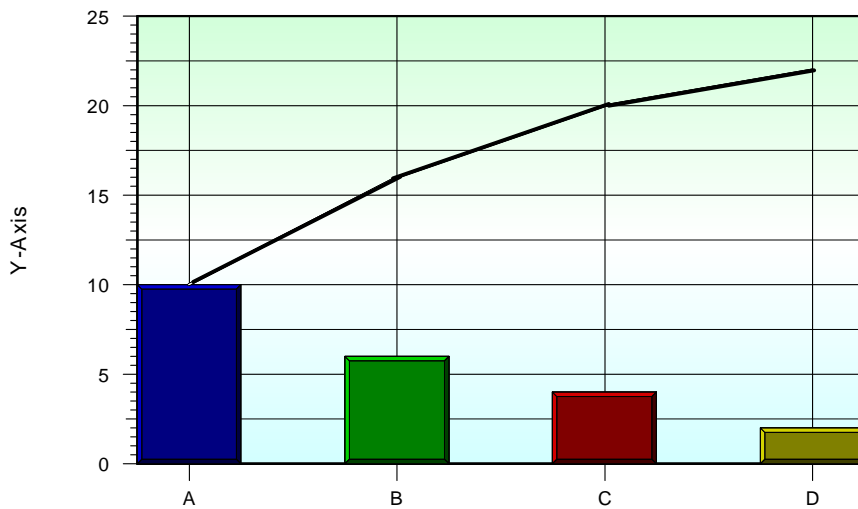


Pareto

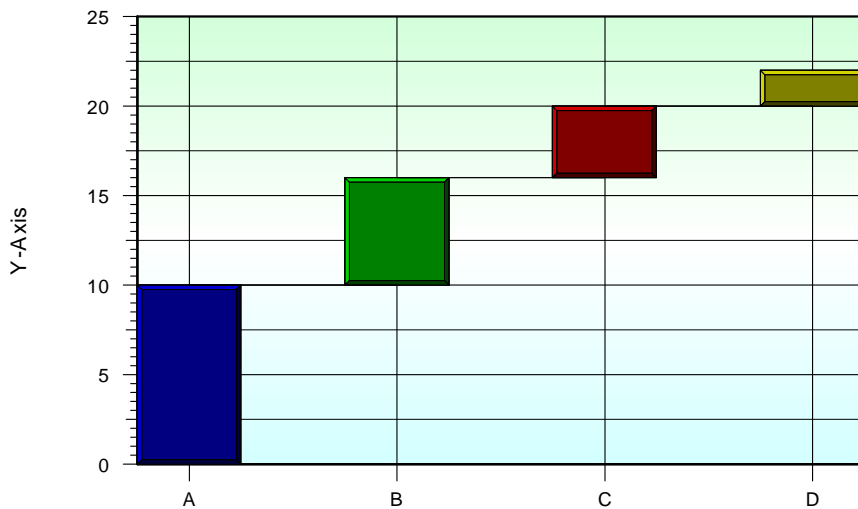
The Pareto diagram corresponds to a histogram, whereas the pursuant columns are depicted in turn, sorted according to the size. In addition the columns have different colours. The biggest value is at the beginning, the smallest at the end.

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:

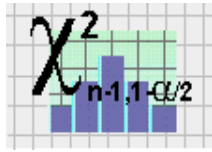


The other variation is the bars represent the sum of the prior values



(Also possible in horizontal representing)

Statistical Tests and Evaluations



For executing the most important statistical tests the following submission files with examples are available. Those are situated in the directory \Statistics:

ANOVA between two Series

Analysis of variance for comparison of two series

[ANOVA_2_Series.vxd](#)

ANOVA & ANOM with several Factors

Analysis of variance for several factors e.g. from an experimental design

[ANOVA_Multi.vxd](#) [ANOVA_Multi_Repetition.vxd](#)

χ^2 -Test of Goodness of Fit

Test of a sample versus a theoretical distribution

[StatTest_Chi2Goodness.vxd](#)

χ^2 -Homogeneity- or Independence Test

Verification on homogeneity or independence of two characteristics with contingency tables

[StatTest_Chi2Homogen.vxd](#)

χ^2 - Multi field Test

Check if the mean number of errors per unit is equal in all considered populations.

[StatTest_Chi2Multifield.vxd](#)

Kolmogorov-Smirnov-Assimilation

Check the assimilation of an observed to an optional expected distribution

[StatTest_KolmogSmirnow_Assim.vxd](#)

t-Test for two Samples

Check if the mean values of two samples are equal

[StatTest_t.vxd](#)

t-Test for Comparison of a Sample with a Default Value

Check if the mean value of a sample matches with a default value

U-Test for two Samples

Check with the rank order, if the mean from two samples are equal

[StatTest_U_Wilcoxon.vxd](#)

F-Test

It is checked if the samples are descended from the same population

[StatTest_F.vxd](#)

Rank Dispersion Test

It is checked if the samples are descended from the same population, if there is no normal distribution

[StatTest_Rank_Dispersion.vxg](#)

Outlier Test

Check a normal distributed sample on outlier and eliminate the values

[StatTest_Outlier.vxg](#)

Balanced simple Analysis of Variance

Comparison of expectations with samples of equal volume

[StatTest_BalVariance_Analysis.vxg](#)

Bartlett-Test

Comparison of more than two populations regarding their standard deviations

[StatTest_Bartlett.vxg](#)

Linearity Test

Check a series on linearity

[StatTest_Linearity.vxg](#)

Gradient Test of a Regression

Check the gradient of an equation straight line

[StatTest_Gradient_Regression.vxg](#)

Test of an Equation Straight Line

on linearity and gradient

[StatTest_StraightLine.vxg](#)

Test Regression Coefficient

Check two regression coefficients on equality

[StatTest_2_RegrCoeff.vxg](#)

Test on Independency of p Test Series

Check the correlation matrix of p test series on reciprocal independence

[StatTest_Independence_p_Series.vxg](#)

Weibull-Benchmark Test of 2 Distributions

Compares two constructions regarding their failure behaviour in the Weibull-net

[\ Weibull \ Weibull_Comparison.vxg](#)

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:

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

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 stands 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 , α , χ^2_{kr}) (with $\alpha = 1 - \alpha$). 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 categorised 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 $\chi^2 > \chi^2_{kr}$ the null hypothesis is refused 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 an 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}	$n_{4..}$	nc_1
M_{A2}	n_{12}	n_{22}	n_{32}	$n_{4..}$	nc_2
M_{A3}	n_{13}	n_{23}	n_{33}	$n_{4..}$	nc_3
$M_{A..}$	$n_{1..}$	$n_{2..}$	$n_{3..}$	$n_{4..}$	nc_4
M_{Ar}	n_{1r}	n_{2r}	n_{3r}	n_{4r}	N_{cr}

The expectation frequencies are calculated for each field by $H_E = n_i \cdot n_j / n$, 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 summarisation 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 **ChiContingency1** (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 **Critical-Worth_Chi2**(f , α , χ^2_{kr}) (with $\alpha = 1 - \alpha$). Here a degree of freedom f is needed, which is determined by: $f = (r-1) \cdot (c-1)$.

If $\chi^2 > \chi^2_{kr}$ the null hypothesis is refused 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 it's improvement measures is observed regarding it's failure behaviour:

	Starting constr.	Measure 1	Measure 2
Failure after 1 weeks	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 behaviour. 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 behaviour 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 in all observed populations.

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 **Critical-Worht_Chi2**(*f*, *alpha*, χ^2_{kr}) (with $\alpha = 1-\alpha$). 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.

If $\chi^2 > \chi^2_{kr}$, the null hypothesis has to be refused on the level of significance.

See also χ^2 -Homogeneity test

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 $T_{\text{prüf}}$.

$$T_{\text{prüf}} = \frac{\max |H_B - H_E|}{n}$$

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 **CriticalWorht_KS**(*n*, *alpha*, T_{kr}) (with $\alpha = \alpha$).

If $T_{\text{prüf}} > T_{\text{kr}}$ the null hypothesis is refused 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).

t-Test for two Samples

This test check the null hypothesis: The mean values of both samples are equal. From s and \bar{x} of both samples the test statistic t_{pr} 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*).

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

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 $t_{pr} > t_{kr}$ the null hypothesis is refused on the level of significance α .

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

The double sided confidence belt is determined by:

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

See also

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

Test for Comparison of a Sample with a Default Value

This test check the null hypothesis: the mean value of the sample corresponds to an alleged mean value μ_0 .

The test statistic is determined by

$$u_{pr} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}$$

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 $\alpha = 1 - \alpha/2$) by the function **InvNormal** (*alpha*, u_α).

The null hypothesis is refused if for the double sided test-case $\mu = \mu_0$ is:

$$|u_{pr}| > u_{kr}$$

The formula for u_σ can be found in the function library.

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

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

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

$$|t_{pr}| > t_{kr}$$

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

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 n₁ and n₂ 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
	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
	Σ = 89	Σ = 53

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

In the common order emerges with the ranked numbers for the spot check 1 and 2 the values represented right with the position-sums R₁ and R₂. 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 $U=U_1=11$, that is compared against a critical value of U_{crit} , is in the presentation-file

[StatTest_U_Wilcoxon.vxd](#)

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

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:

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

whereby the larger variance is always in the counter, so that $F_{pr} \geq 1$. 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_{kr}) (with $alpha = 1-\alpha/2$). 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 $F_{pr} > F_{kr}$ the null hypothesis is refused on the level of significance α .

The example file is called [StatTest_F.vxd](#), which can be used for own evaluations.

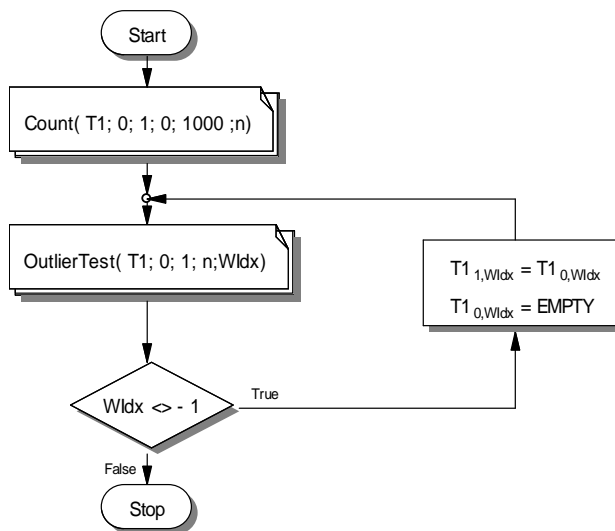
See also Rank Dispersion Test

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:

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

which will be compared to a critical value T_{kr} . 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:

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

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 $alpha = 1-\alpha$). The degree of freedom f_1 and f_2 results from $f_1 = k-1$ and $f_2 = k(n-1)$.

If $F_{pr} > F_{kr}$ the null hypothesis is refused on the level of significance α .

The double sided confidence belt is determined by the critical t-values (**CriticalWorht_t**(f_2 ; $alpha$; t) with $alpha = 1-\alpha/2$ to

$$x_i - t_{f_2, 1-\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq x_i + t_{f_2, 1-\alpha/2} \frac{s}{\sqrt{n}}$$

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 $n_i \geq 5$. The test statistic results to:

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

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_i^2 of the respective samples are taken for granted, resp. 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_{kr}) (with $alpha = 1 - \alpha$). Here a so called degree of freedom f is needed, which is determined by: $f = k - 1$.

If $\chi^2_{pr} > \chi^2_{kr}$, the null hypothesis has to be refused 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	38	52	69	87	109
n2=n1+1	11	29	42	57	74	93	115
n2=n1+2	12	32	45	61	79	99	121
n2=n1+3	13	35	49	65	84	105	127
n2=n1+4	14	38	53	70	89	110	134
n2=n1+5	14	42	57	74	94	116	140

H_0 is refused ($\alpha=0,05$ double sided resp. $\alpha=0,025$ one sided) if R_1 or R_2 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 $R_1=134$ and $R_2=76$ have been determined. As it can be seen in the above table that $R_1 < 78$ and $R_2 > 132$. So the null hypothesis has to be refused, there is no dispersion difference.

Spot 1	Spot 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 an Best Fit Straight Line

Any best fit straight line is tested on linearity and gradient. This test is a summarisation 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_{krit} on the level of significance $\alpha=5\%$. Both regression coefficients are equal, if $t < |t_{\text{krit}}|$.

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 categorised 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:

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

which will be compared on the level of significance $\alpha=5\%$. If $F < F_{\text{krit}(k-2, n-k)}$, 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 a equation straight line significantly differs from 0. A t-value is formed as quotient of a regression coefficient b to it's variation. From this the value α is determined from the student's t-distribution and compares this to the level of significance $\alpha=5\%$. If $\alpha < 5\%$, 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:

$$R' = |r_{\text{limit}}| \sqrt{\frac{n-2}{1-r_{\text{limit}}^2}}$$

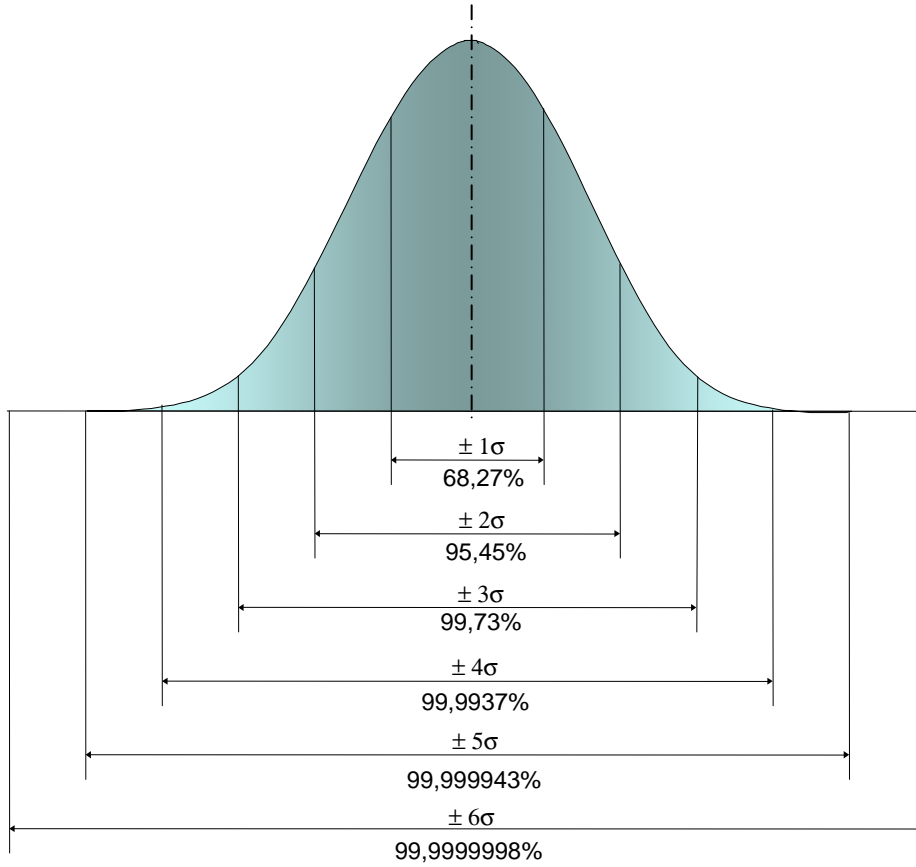
and iteratively r_{limit} is calculated, which is compared to the maximum founded correlation coefficient. If $r_{\text{max}} < r_{\text{limit}}$, then on the level $\alpha=5\%$ no pair of series of measurements significantly depends on each other.

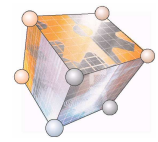
The file [StatTest_Independence_p_Series.vxg](#) is used as submission, which can be adjusted for own evaluations.

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	Characteristical service life at Weibull	For 63.2% failure frequency
w	Weighting	Number of alleged value
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	

Normal-distribution





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 ganzheitlichen 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 vielparametrischer Systeme in der Kfz-Antriebsentwicklung
Alexander Mitterer
Fortschritt-Bereich 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

Index

3 levels	30	coefficient of determination	45
agglomerativ	67	test of	48
Analysis Of Means	21	coefficients	72
Analyses of Variance		Comparison B versus. C	12
regression model	45	component	72
Analysis of Variance	17	components	70, 74
ANOM	19	Components Search	7
ANOVA	17, 19	Condition Number	50
Balanced simple ANOVA	97	confidence interval	
Bartlett-Test	98	linear regression	37
Bhote	104	Confidence interval	
Box	104	regression coefficient	50
Box-Behnken	30	response	50
Box-Cox	52	confidence intervals	31
Boxplot	83	confounding	25
Bravais - Pearson	35	constant	42
capability	9	contingency table	91
Categorical characteristic	69	contingency-table	89, 90
Categorical Factors	44	correlating data	72
CCC	30	correlation	35
CCD	31	correlation coefficient	37
CCF	30	correlation coefficient	35
Central Composite Circumscribed	30	Correlation Loading Plot	74
Central composite Design	23	correlation matrix	51, 66
Central Composite Design	29	Correlation matrix	36
Central Composite Face	30	Cubic	23
central point	29	cubic model	81
central points	31	Curve-diagram	54
chemical liquids	32	Data reduction	67
Chi ² Homogeneity Test	90	degrees of freedom	45
Chi ² Multi Field Test	91	dendrogram	68
Chi ² Test of Goodness of Fit	89	Design of Experiment	23
City-block Distance	66	determinant	31
cluster analysis	66	deviation	60
		Discrete regression	60

distance matrix.....	67	Kleppmann	104
D-Optimal	24	Kolmogorov-Smirnov-Assimilation Test	
Draper	104	92	
Effects	55	Lack of Fit	47
eigenvalue	70	latent variable	72
eigenvalues	50	LH	64
eigenvectors	70	Likelihoods.....	60
ellipse	70, 75	Linear	23
equidistantly.....	32	linear model	42
Erfüllungsgrad	59	linear regression	37
Euklid's distance	66	Linearity Test	100
experimental design.....	23	LL	60
factor loadings	70	loading	73
factors	70	loadings	70
fractional	27	logarithmic	29
Fractional.....	23, 25	Logits	60
fractional test	28	Log-Likelihood	64
F-Test	95	Mann	94
ANOVA	19	matrix form.....	42
Full factorial	23, 25	maximum	29
geometric center	68	Maximum Likelihood	60
Gliding average.....	85	Measurement-Comparison	9
Gradient Test of a Regression	101	measurements	57
Grubbs-Test.....	57	Median plot	84
hierarchical	67	minimum	29
Ho	99	Mixture	24
homogeneously	90	Mixture Plans	32
hypothesis	89, 90, 91, 92, 95, 98	MLR	72
Independence Test of p Series	101	model prediction	46
independent variables.....	41	Multiple Regression	42
Intensity-Relation-Matrix	14	Multivariate Analises.....	66
interaction	42, 57	Multi-Vari-Chart.....	10
Interaction.....	23	Neural Networks	76
Interaction model	42	neurons.....	76
interactions	31	NIPALS.....	72
Isoplot	9	NN	76
		nonlinear	29

nonlinearity	23	Regression	37
Norman.....	104	regression coefficients	100
null hypothesis	99	test of	48
number of tests	31	regression model	72
observation series.....	65	Regressionstypen	39
optimization	79	Reproducibility	48
Optimization.....	58	residual-matrix	73
orthogonal.....	23, 25, 28	Residues.....	56
orthogonality	50	Resolution.....	26
Outlier Test	96	RMS 50	
outliers 35, 46, 57, 74		Root mean squared	50
Paired Comparison	11	Sachs 104	
pairwise comparison	11	Scatter bars	82
Pareto 86		Scatter Plot	8
Pareto-diagram	55	Score 70	
partial correlation coefficient	36	Score plot	74
Partial Least Squares	72	Scores 72	
PCA 70, 72		Screening	26
Peterson	104	screening plans.....	46
Plackett-Burman	27	Screening-plans.....	27
PLS 72		Shainin 6	
Polynom.....	39	Spearman	35
Prediction Measure.....	46	spread 73	
Principle Component Analysis	70	square 29	
Priority Matrix.....	15	squared terms.....	42
Probst 104		Standard deviation.....	50
pseudo-R ²	60	standard plans	23
pure Error	47	Standardize.....	51
Q ² 46		star 29	
quadratic.....	29	Statistical Charts.....	54
Quadratic	23	Statistical Factors	102
qualitative factors.....	44	Statistical Tests.....	87
R ² 45		Sum of Squares	60
Rank Dispersion Test	99	Taguchi.....	23
ranking 99		Taguchi plans	28
reduction.....	70	Test for Comparison of a Sample with a Default Value	93
Red-X 6			

Test regression coefficients	100
theoretical distribution	89
Topology	76
Training-Algorithm	78
transformation	29
transformed	46
triangle	32
Tschebyscheff distance	66
t-Test for two Samples	92

U-Test	94
variance	74, 98
vector	42
VIP	74
weight matrix	72
Whitney	94
Wilcoxon	94
Wold	72, 74

