Robustness Investigation of a Numerical Simulation of the ECE-R14 with particular regard to correlation analysis

Rolf Henniger*, Klaus Hessenberger*, Heiner Müllerschön**

*DaimlerChrysler AG, Stuttgart, Germany
**DYNA*more* GmbH, Stuttgart, Germany

Abstract:

The stability of seats, seat-belts and seat-belt anchorages of vehicles are very important for the passengers' safety during an accident. Therefore, dedicated safety tests have to be passed in order to ensure a correct functionality of those parts. These tests are defined in the European regulation ECE-R14 in detail, whereas crashing loads are substituted by appropriate pulling forces on the seat-belts.
The complete configuration is designed, simulated and tested with certain (physical) parameters (geometry, materials, testing conditions, ...) which are assumed to be constant. In reality, most of these quantities may vary and take other values with appointed probabilities. Generally, different values for the design parameters will cause a variation in the simulation or testing result, which will also occur with a certain probability. Thus the results of arbitrary separate investigations are commonly not meaningful enough because their probability of occurrence is unknown.
The determination of this probability is part of a robustness investigation. In this paper the Monte Carlo Analysis as a very simple method is applied to a FEA model observing ECE-R14. Additionally, the identification of each parameter's importance for the simulation or testing result is a main topic. This is done by means of a linear correlation analysis, which is described in section 3 in detail.

Keywords:

ECE-R14, seat pull, LS-DYNA, LS-OPT, Monte Carlo Analysis, robustness investigation, reliability, Latin Hypercube Sampling, linear correlation coefficient, non-linear correlation coefficient, total correlation coefficient, regression coefficient, confidence interval

1 Introduction: Robustness Investigation and Monte Carlo Analysis

1.1 System Behavior due to Parameter Influence

Generally, system behavior depends on the influence of underlying *parameters*: For instance the elastic elongation of a pulled tensile bar is directly controlled by its cross-sectional area. This is a pure *deterministic* dependency. In the case of a pushing load the parameter's influence on the system is more complex because buckling modes my also occur. If the buckling direction of a ideally double-symmetric bar is investigated, the buckling direction can not be controlled by varying geometric parameters like sheet thickness or bar length. Those parameters may just control the occurrence itself in a deterministic way, but the buckling direction behaves completely *chaotic*. In most practical cases both deterministic and chaotic behavior occur.
Parameter values of a physical system are mostly not constant at all: As an example, the thickness of sheet metal varies statistically around a nominal value which is used for dimensioning of a structure. This is equal to an uncontrolled parameter variation, which influences the system behavior in a more or less deterministic and/ or chaotic way. Therefore, the behavior of a (physical) system as described above is usually not constant, so it may vary statistically, too.

1.2 Robustness Analysis

In the following, the system behavior is more precisely described as a set of system *responses* (e. g. bar elongation, velocity, acceleration, ...). System responses are often directly restricted with respect to a certain criterion, thus a statistical variation of a response may cause failure of the system. Any variation of a parameter or a response can be described by a statistical distribution. If the exact statistical distribution of each considered response is known, an exact probability of failure can be computed. In the case of a *reliable* system the probability of failure is smaller than a maximum given value.
Robust systems have ideally just a small range in their relevant responses, so the system behavior itself varies just in a permitted bandwidth statistically. Therefore, two different criterions characterize a robust system:

- System responses are mainly independent of statistically uncertain parameters.
- In analogy to the pushed tensile bar, the system has also just a minimum of buckling modes, so its responses are mainly unique.

Therefore, a *robustness analysis* includes primary the estimation and analysis of the statistical distributions of all relevant parameters and, upon that, of all relevant responses. "Relevant" indicates that just the most important quantities can be considered in most practical applications.
An analysis of a statistical distribution may be done basically by computing characteristic terms like reliability, mean or variance. Furthermore, the influence of each parameter on each system response may be investigated using correlation coefficients. In the case of linear correlation coefficients, this also allows the identification of the system responses' (physical) sensitivities with respect to the parameters.

1.3 Monte Carlo Analysis

Robustness itself is based obviously on statistical principles, so a detailed investigation has to utilize an appropriate method. The *Monte Carlo Analysis* is a very simple and universal statistic-based instrument, which is very useful for solving complex problems of any kind. It is not limited just to statistical problems, but its effectivity depends always more or less on the type of application.
Each probability distribution represents the distribution of a *population*. In the case of a discontinuous or "stepwise" distribution a population may have just a finite number of possible values for the observed quantity unlike a continuous distribution always bases on a infinite number. The latter is the usual case. The Monte Carlo Analysis tries to approximate arbitrary probability distributions by selecting just a subset of the entire population, which still represents the original distribution approximately. The elements of this subset may be interpreted as samples of the entire population, so they are called *samples* or *experiments*, generally.
Hence, each parameter of the investigated system, varying around a nominal value, may be substituted by a specific number of experimental values in order to describe its statistical distribution approximately. The more experiments are used the more accurate can the exact probability distribution be described. Generally, parameter values may be selected arbitrary and weighted with an additional factor in order to approximate the exact distribution in a feasible manner. More practical is a "direct" appropriate allocation of the parameter values, so the experiments have not to be weighted additionally (*Markov Chain*). The

most popular method for the latter case is the *Latin Hypercube Sampling* [6], [7], which distributes the values of each parameter in a half-structured way. The values of the different parameters are combined either randomly or via an optimization algorithm in order to generate a good multidimensional spreading (generally not necessary). Therefore, the values of every combination of two parameters are expected to be uncorrelated.
In a next step, the system is investigated by using the Monte Carlo experiments: If parameters have an influence on the system behavior or the system behavior itself is not completely unique, the responses will also follow stochastic distributions. These are consequently just approximations to exact distributions, which could be found by considering the entire population. In most practical cases, uncertain physical quantities own an infinite number of possible values so this will be quite impossible.
The approximations to the exact distributions can be used for computing several derivated quantities like reliability, mean or spreading (variance, standard deviation) of the responses. As mentioned above, the experiments are just a representative subset of the entire population. The choice of another subset would (probably) cause different distributions of the responses. Therefore, the value of any derivated quantity is an *estimation* to an (unknown) exact value. Due to the validity of the *law of large numbers*, the resulting quantity values out of any observed subset will converge with increasing number of experiments to the exact values. The uncertainty, caused by using a small subset, can be incorporated by computing appropriate *confidence intervals* for each derivated quantity. These intervals are based just on statistical fundamentals and depend not on the number of observed parameters. The law of large numbers is represented by decreasing confidence interval widths for an increasing number of experiments.
The definition of a confidence interval also needs the specification of a *statistical certainty* for expecting the exact value inside it. As an example, for a certainty of 90% and multiple Monte Carlo Analyses the exact value of a quantity will lie inside the confidence intervals in 90% of all analyses.

2 ECE-R14

2.1 Regulation

ECE-R14 is an European regulation which contains uniform provisions concerning the approval of vehicles with regard to safety-belt anchorages. In this paper ECE-R14 is applied to a commercial vehicle. There are several different restrictions defined in the regulation whereas just two of them are critical for the investigated model and the considered seat row. They are shown in Fig. 2.1: One end of each seat's upper safety-belt is mounted on top of each seat back (J). The motion of these anchorages as a result of predefined quasi-static pulling forces, applied to the safety-belts, is limited to a given permitted area (shaded). During the development stage, forces are usually increased by a given percentage in order to ensure the validity of testing or simulation results under ECE-R14's testing conditions. Those harder conditions will result in a larger probability of failure.

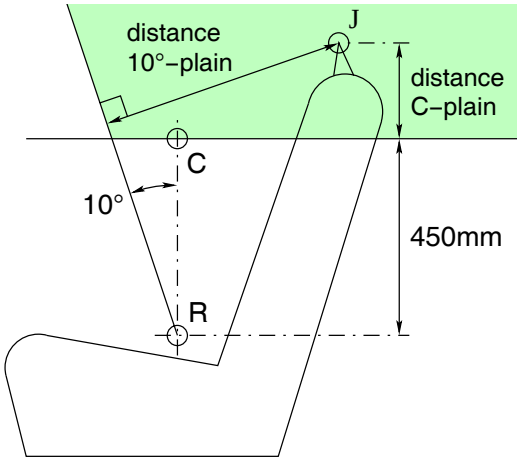

Fig. 2.1: Restrictions to the upper seat-belt anchorages J

2.2 Analysed FEA Model

Due to large nonlinearities, mainly caused by contacts, the simulation of the seat pull test is performed using LS-DYNA with explicit time integration scheme and limited time-step size [5]. In order to limit computing time, the (physical) simulation time is strongly reduced to a minimum, so it is no longer quasi-static rather highly dynamic.
The original design of the investigated vehicle and seat row, shown in Fig. 2.2, complies with the requirements in an exemplary single test and simulation. In order to achieve more detailed information about this conclusion, the Monte Carlo Analysis is applied to the FEA model using the Latin Hypercube Sampling, implemented in LS-OPT [10]. In order to achieve a maximum accuracy for the probability distributions and all derivated quantities, a large number of experiments has to be performed. In this application $N = 200$ experiments are aspired while the FEA model comprehends a comparatively large number of elements. In order to reduce computing time considerably, the original FEA model is reduced to a minimum of elements (about a third of the original model) by replacing selected areas of the structure by appropriate boundary conditions. These are time-dependent and can be found by simulating the complete model (Fig. 2.3): The data for the specified interface are written to a file that will be integrated in the stripped-down model as boundary conditions for all further simulations.

2.3 System Responses

Meaningful system responses are found directly from Fig. 2.1 by observing the distances from the belt anchorages to the restrictions. It turns out that just the restriction "distance 10°-plain" ($d_{10^\circ\text{-plain}}$) is violated during the analysis. On the other hand, the simulation time is too short in some cases because the anchorages are still in motion (dynamic simulation) as a result of certain coincident parameter variations: As a consequence it is not possible to rate wether one of the geometric restrictions may be violated at a later time. Restricting the velocity along the longitudinal axis at the end of the simulation ($v_{x,\text{end}}$) seems to be necessary, because a longer simulation time for each experiment should be prohibited. Experiments, which fail at this restriction, can now be rated as failed (simplest method) or evaluated again separately for a longer simulation time, afterwards ("correct" method). Before starting the Monte Carlo Analysis, the influence of the parameters is not known yet, so it is impossible to determine a sufficient long simulation time for all of the experiments. On the other hand, a short computation time is aspired generally, so restricting the velocity is a feasible way to handle this problem.

2.4 Parameters

Investigated parts and (out of that) parameters are shown in Fig. 2.4: As mentioned above, the influence of any parameter in the model is not known yet, so a preselection has to follow mostly subjective assumptions. Additionally, the number of significant parameters should not be too large, if their influence will be investigated in a correlation analysis afterwards. The more significant parameters are used at a given number of experiments, the less meaningful are the correlation coefficients as demonstrated in sec. 3. Due to this, it is aspired to limit this number, but at the same time the investigation of a broad bandwidth of parts and parameters is needed for computing "realistic" probability distributions of the responses. As a compromise, there are just two parameters selected for each part: The sheet thickness (Th.) and the yield stress (YS) seem to be a fairly good choice. Additionally, variation of the Young's Modulus and the coefficients of friction are investigated for all parts simultaneously by scaling their mean values for different materials. Variation of the pulling forces may also have a significant influence on the system responses.
A total of $n = 43$ parameters is varied in this analysis. All parameters are assumed to be normally distributed except the pulling forces (equally distributed). The mean values and standard deviations are obtained out of measuring data.

2.5 Evaluation

2.5.1 Probability Distributions

The scatter of the $N = 200$ Monte Carlo experiments (Latin Hypercube Sampling) is exemplarily shown in Fig. 2.5 for a material thickness vs. a yield stress. Parameters are independent among each other, so ideally there shouldn't be identifiable any directions or structures in the distribution of the experiments. After evaluating the experiments, system responses can be plotted vs. parameter values. As an example the response "distance 10°-plain" is plotted vs. the pulling forces in Fig. 2.6. Due to obviously existing

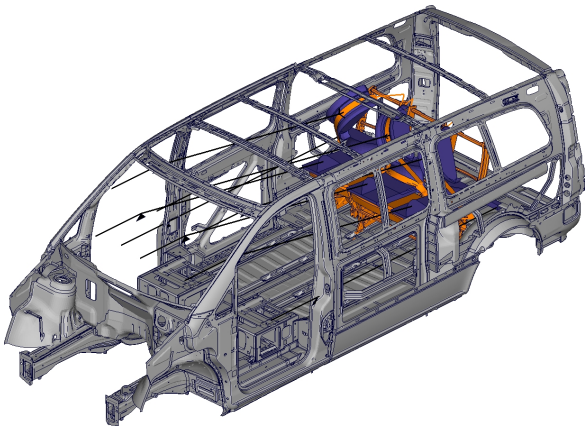

Fig. 2.2: Complete FEA Model

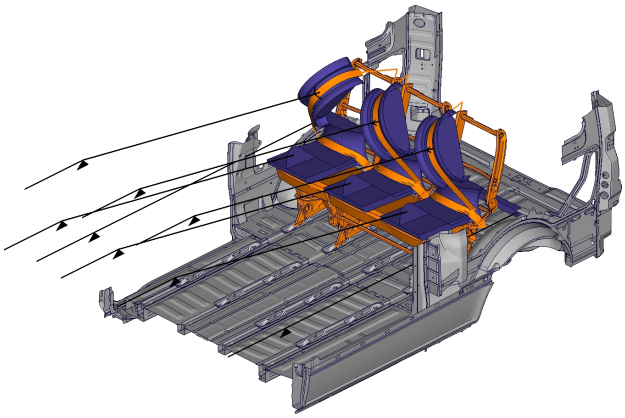

Fig. 2.3: Reduced FEA Model with boundary conditions

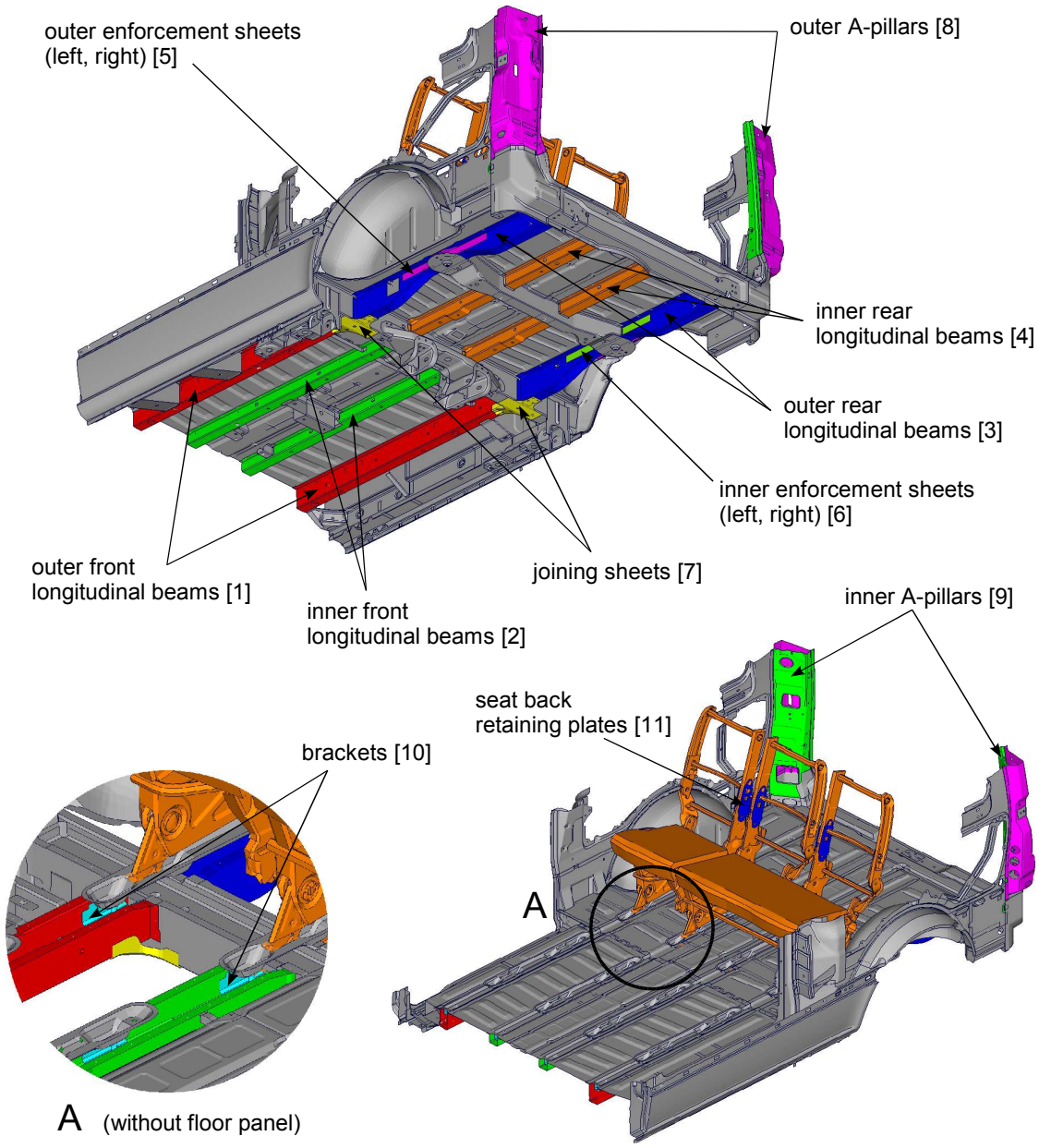

Fig. 2.4: Investigated parts

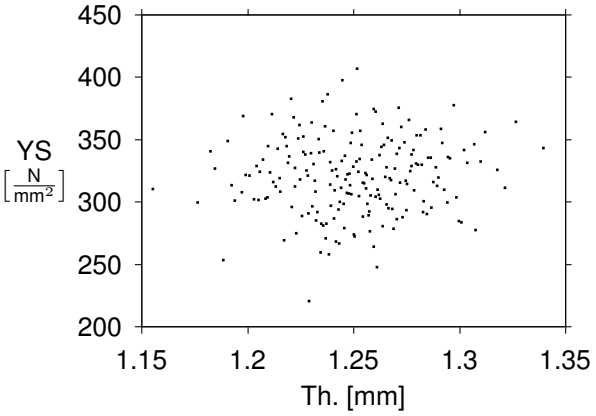

Fig. 2.5: Experiments displayed for two parameters (Th. = sheet thickness, YS = yield stress)

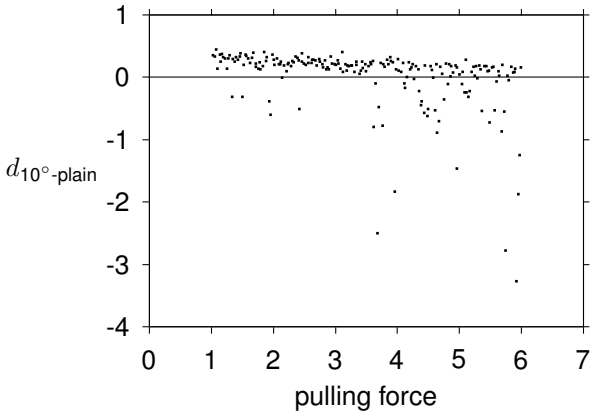

Fig. 2.6: Experiments displayed for a parameter vs. a response

parameter influence, an orientation of the experiments is now clearly visible. Because Monte Carlo experiments represent in this case an infinite large population, the distribution approximates the unknown exact probability distribution. This can be shown in histograms, which display the numbers of system responses belonging to certain intervals (Fig. 2.7 and 2.8). For comparison, a normal distribution based on the responses' mean and standard deviation is plotted, too. The more experiments are used the more "continuous" can the distributions be displayed (approximation quality also increases).

2.5.2 Quantities and Corresponding Confidence Intervals

The determination of several quantities and (additionally) appropriate confidence intervals allow a numerical characterization of the distributions [3], [4], [8]:

- Mean of N responses y_i:

$$\overline{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{2.1}$$

The determination of the dedicated confidence interval requires the knowledge of the *type* of the error's probability distribution. Generally, this is unknown so an appropriate approach has to be done, whereas the assumption of a normal distribution is a common choice. In this case half of the confidence interval is given by

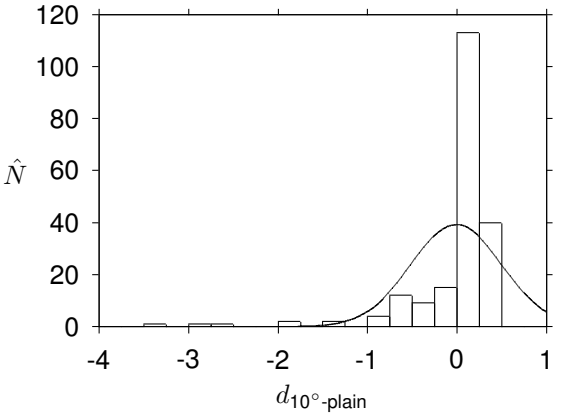

Fig. 2.7: Histogram for the distance to the 10°-plain

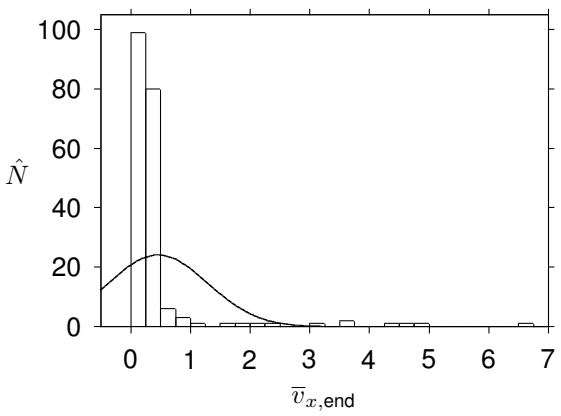

Fig. 2.8: Histogram for the final velocity

$$\Delta\overline{y}_{\mathsf{N}} = t_{(1-\frac{\alpha}{2},N)}\frac{\sigma_{\mathsf{N}}}{\sqrt{N}} \tag{2.2}$$

with σ_{N} as the estimated standard deviation of a standard normal distributed population (see below) and $t_{(1-\frac{\alpha}{2},N)} = -t_{(\frac{\alpha}{2},N)}$ as the *quantile* of the Student's Distribution as a function of the probability of falsity α. The (symmetric) confidence interval for the exact mean value μ_{N} is:

$$\overline{y}_{\mathsf{N}} - \Delta\overline{y}_{\mathsf{N}} \leq \mu_{\mathsf{N}} \leq \overline{y}_{\mathsf{N}} + \Delta\overline{y}_{\mathsf{N}} \tag{2.3}$$

Generally, the error caused by the assumption of a normal distributed error decreases very fast for a increasing number of experiments N.

- Variance σ^2 (standard deviation σ) of N responses y_i:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})^2 \tag{2.4}$$

The determination of the confidence interval suffers from the same problem as before (mean value): An assumption for the unknown exact probability distribution of the error is needed. For a normal distribution, the lower and upper interval bounds are

$$\sigma^2_{\mathsf{N},1} = (N-1)\frac{\sigma^2_{\mathsf{N}}}{\chi_{(\frac{\alpha}{2},N-1)}}, \tag{2.5}$$

$$\sigma^2_{\mathsf{N},2} = (N-1)\frac{\sigma^2_{\mathsf{N}}}{\chi_{(1-\frac{\alpha}{2},N-1)}}, \tag{2.6}$$

with $\chi_{(1-\frac{\alpha}{2},N-1)}$ as the quantile of the chi-square distribution. So the (non-symmetric) confidence interval for the exact variance $\tilde{\sigma}^2_{\mathsf{N}}$ is:

$$\sigma^2_{\mathsf{N},1} \leq \tilde{\sigma}^2_{\mathsf{N}} \leq \sigma^2_{\mathsf{N},2} \tag{2.7}$$

- Probability of failure for K responses which do not exceed any constraint at a total number of N responses (*Maximum Likelihood Estimation*):

$$P = \frac{K}{N} \tag{2.8}$$

The confidence interval for this quantity doesn't need further information about the distribution of the responses. It is based on the approximation of a binomial distribution, which describes the statistical relationship between P, K and N in eq. 2.8, by a normal distribution. This is allowed if the following heuristic constraint is fulfilled:

$$P\,N \geq 10 \quad \vee \quad (1-P)\,N \geq 10 \tag{2.9}$$

Then the upper and lower interval bounds are:

$$P_{1,2} = \frac{2K + n^2_{(1-\frac{\alpha}{2})} \pm \sqrt{\left(2K + n^2_{(1-\frac{\alpha}{2})}\right)^2 - 4\left(N + n^2_{(1-\frac{\alpha}{2})}\right)\frac{K^2}{N}}}{2\left(N + n^2_{(1-\frac{\alpha}{2})}\right)} \tag{2.10}$$

The confidence interval for the exact probability of failure $\tilde{P}$:

$$0 \leq P_1 \leq \tilde{P} \leq P_2 \leq 1, \tag{2.11}$$

- The linear correlation coefficient for N responses y_i belonging to the values $x_{i,k}$ of a parameter k:

$$\rho_{x_k y} = \frac{\sum_{i=1}^{N}(x_{i,k} - \overline{x}_k)(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_{i,k} - \overline{x}_k)^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}}, \qquad -1 \leq \rho_{x_k y} \leq 1 \tag{2.12}$$

  The confidence interval for the exact correlation coefficient $\tilde{\rho}_{x_k y}$, assuming a normally distributed estimating error:

$$\tanh\left[\frac{1}{2}\ln\left[\frac{1+\rho_{x_k y}}{1-\rho_{x_k y}}\right] - \frac{t_{(1-\frac{\alpha}{2},N)}}{\sqrt{N-3}}\right] \leq \tilde{\rho}_{x_k y} \leq \tanh\left[\frac{1}{2}\ln\left[\frac{1+\rho_{x_k y}}{1-\rho_{x_k y}}\right] + \frac{t_{(1-\frac{\alpha}{2},N)}}{\sqrt{N-3}}\right]. \tag{2.13}$$

Every confidence interval includes the so called *square root law* for a larger number of experiments N: The width of each confidence interval decreases with the inverse square root of the number of experiments N. Because the interval width represents the accuracy of the corresponding quantity, the accuracy of the results increases with the square root law, too.

2.5.3 Parameter Influence

The parameter influence is investigated by considering the linear correlation coefficients between system responses and parameters. As a special property, the confidence interval width of the estimated correlation coefficient decreases with an increasing absolute value of the correlation coefficient $|\rho_{x_k y}|$ (see eq. 2.13). Accordingly, the accuracy increases at the same time. The parameter influence on the distance to the 10°-plain ($d_{10°\text{-plain}}$), as the most important system response, is shown in Fig. 2.9 using a certainty of $1 - \alpha = 90\%$ for the confidence intervals (the bold line represents the estimated value).
A parameter is called "significant" if the associated interval width does not include the value $\rho_{x_k y} = 0$, so its influence can't be neglected. Apparently about half of the interval width, belonging to the estimated coefficients, is in almost all cases larger than the coefficient itself so just very few of the parameters can be identified as significant for the system response. This result is not satisfying, so it should be investigated how the quality can be improved in order to receive more detailed information and more precise conclusions in a further analysis.

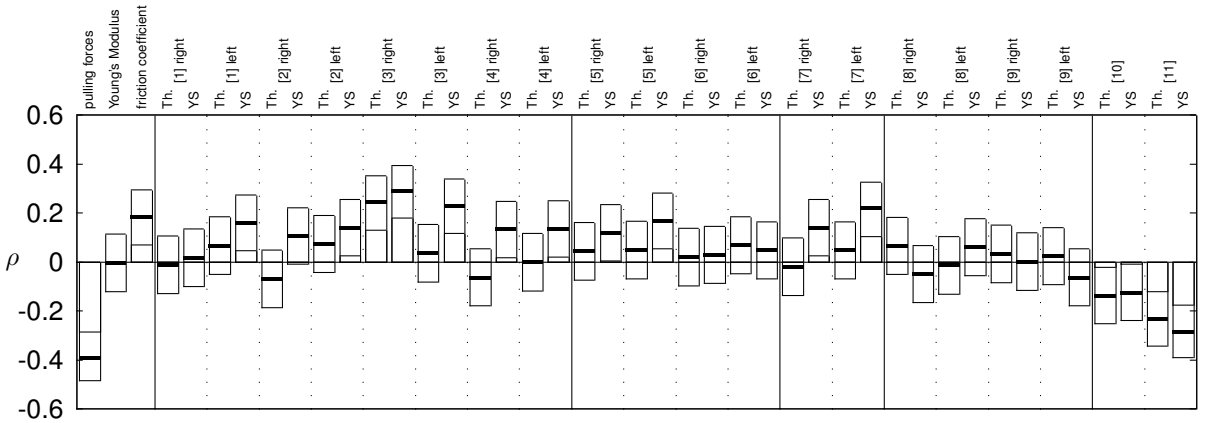

Fig. 2.9: Linear correlation coefficients of all parameters with respect to the response $d_{10°\text{-plain}}$. Each thick line indicates the estimated value of a correlation coefficient whereas the thinner upper and lower lines represent the corresponding confidence interval.
($\alpha = 10\%$; Th. = sheet thickness, YS = yield stress)

3 Properties of the Correlation Coefficient

3.1 Interpretation of the Correlation Coefficient

For a better understanding, a very simple system is investigated in detail: It depends on one parameter x_1 with a pure linear influence and a second parameter x_2 with a pure quadratic influence on the response y as shown in Fig. 3.1. This response surface is sampled with $N = 1000$ equal distributed experiments which can be projected to 2-dimensional plots (Fig. 3.2 and 3.3). In each plot of a parameter x_1 or x_2 vs. the response y, the scatter around a middle trend is obviously caused by the other remaining (not directly displayed) parameter. The correlation coefficient considers just 2-dimensional information between the values of two quantities (e. g. one parameter and one response). This observation is not limited to 2-dimensional problems, thus scatter in a 2-dimensional plot of an arbitrary multi-dimensional system is always caused by the remaining parameters. Even noise may be interpreted as the influence of other (unknown and therefore not controllable) parameters, so each arbitrary system response may always be assumed to be dependent on a certain number of parameters in a *purely deterministic* way. Therefore each system response could be described exactly by a unique surface.
If more than one parameter influences the system response, scatter is always present in any plot of a parameter vs. a system response.

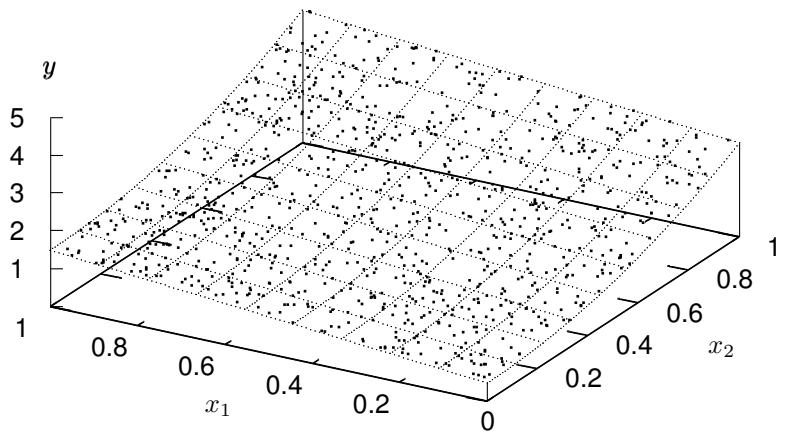

Fig. 3.1: Distribution of $N = 1000$ experiments sampling the response surface $y = y(x_1, x_2)$ (linear relation with respect to x_1, quadratic relation with respect to x_2)

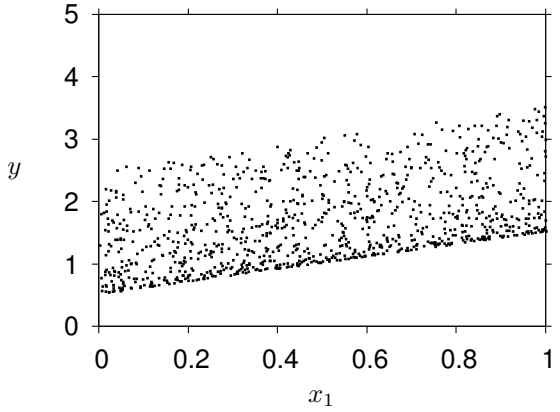

Fig. 3.2: Distribution of parameter values x_1 and responses y ($\rho_{x_1 y} = 0.454$)

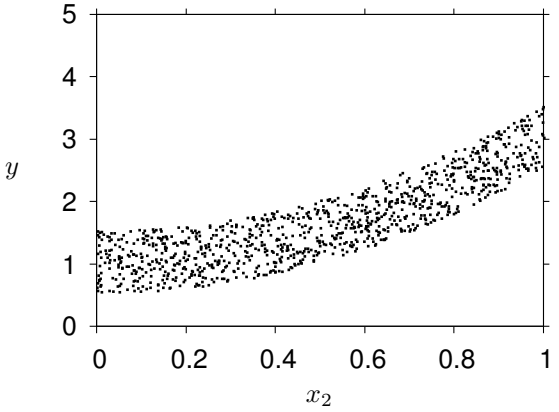

Fig. 3.3: Distribution of parameter values x_2 and responses y ($\rho_{x_2 y} = 0.867$)

3.2 Values of the Correlation Coefficient

The correlation coefficient is non-dimensional and limited to the interval $\rho_{x_k y} \in [-1, 1]$ (eq. 2.12). In the following, different cases of 2-dimensional distributions are investigated with respect to the (linear) correlation coefficient.

3.2.1 Linear System Response in Dependency of a Single Parameter

In this first case it is assumed that just one (or no) parameter has an exactly linear influence on the system response. Therefore, no scatter can occur in the plot of the response vs. this parameter (Fig. 3.4). The relation between each system response y_i and the observed parameter values $x_{i,k}$ is simply given by

$$y_i = a_1 x_{i,k} + a_0 \quad \text{and} \quad \overline{y} = a_1 \overline{x}_k + a_0 \tag{3.1}$$

with arbitrary values for a_0 and a_1. Inserting into eq. 2.12 leads to the extrem values

$$\rho_{x_k y} = = \begin{cases} +1 & \text{for} \quad a_1 \geq 0 \\ -1 & \text{for} \quad a_1 \leq 0 \end{cases} \tag{3.2}$$

because the standard deviations in the denominator (see eq. 2.12) are selected as positive, while the numerator may have both signs. For $a_1 = 0$ the linear correlation coefficient is not unique, because it has two possible solutions $\rho_{x_k y} = \pm 1$.

3.2.2 Arbitrary System Response and Observation of a Non-Influencing Parameter

For the opposite case it is assumed that in a multidimensional problem at least one parameter has a large influence on the system response. Investigating another parameter with no influence on the response, as shown in Fig. 3.5, leads to the correlation coefficient

$$\rho_{x_k y} = 0. \tag{3.3}$$

The scatter caused by the other parameters is larger than the variability caused by the observed parameter.

3.2.3 Arbitrary System Response and Observation of an Arbitrary Influencing Parameter

- Between the extreme cases in eq. 3.2 and 3.3, the (linear) correlation coefficient may have any value $\rho_{x_k y} \in [-1, 1]$.
- In the case of a linear correlation coefficient, its sign conforms to the sign of the mean slope of the response with respect to the investigated parameter.
- A parameter with a large influence on the system response causes a wide scatter in the plot of any other parameter vs. the response (see Fig. 3.2 and 3.3). Vice versa, a parameter with a small influence causes a small scatter in the plot of any other parameter vs. the response.

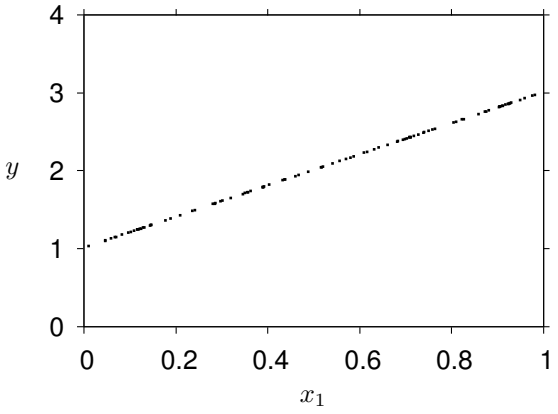

Fig. 3.4: Distribution of parameter values x_1 and responses y ($\rho_{x_1 y} = 1$)

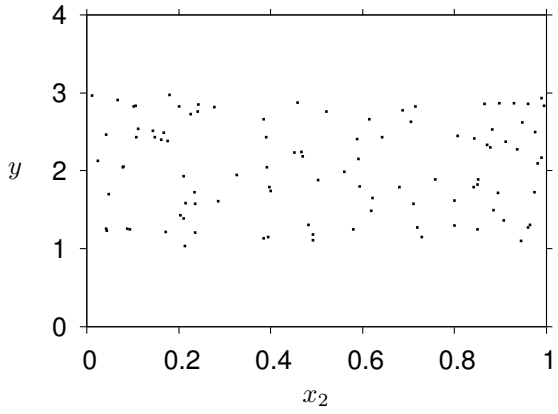

Fig. 3.5: Distribution of parameter values x_2 and responses y ($\rho_{x_2 y} = 0$)

- Therefore, parameters with large influence *compared to all other* parameters cause large absolute values of the correlation coefficient. The correlation coefficient *ranks* the parameters' influence (absolute value). This is demonstrated in the following sections more precisely.

3.3 Relation Between the Linear Correlation Coefficient and the Regression Coefficient Using Linear Polynomials

Generally, regression coefficients are found by means of the Least Square Criterion: The system response at a single experiment $\mathbf{x}_i = [x_{i,k}]$ is defined as $y(\mathbf{x}_i) = y_i$ which can be approximated globally by a sum of L shape functions $\phi_s(\mathbf{x}_i)$, weighted with the regression coefficient a_s:

$$f(\mathbf{x}_i) = \sum_{s=0}^{L} a_s \phi_s(\mathbf{x}_i) = f_i \tag{3.4}$$

The difference between y_i and f_i is the *regression* or *modelling* error $d(\mathbf{x}_i) = d_i$:

$$y(\mathbf{x}_i) = f(\mathbf{x}_i) + d(\mathbf{x}_i) \tag{3.5}$$

The matrices $\mathbf{y} = [y_i]$, $\mathbf{f} = [f_i]$ and $\mathbf{d} = [d_i]$ contain the corresponding values of all experiments $\mathbf{x}_i$. With the shape functions $\mathbf{X} = [\phi_k(\mathbf{x}_i)] = [X_{i,k}]$ and the regression coefficients $\mathbf{a} = [a_k]$ the approximated system responses may be written as

$$\mathbf{f} = \mathbf{X}\mathbf{a}. \tag{3.6}$$

It is easy to demonstrate that

$$\mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.7}$$

returns the demanded regression coefficients in terms of the least square method. In the special case of a linear approach, eq. 3.4 may directly be written as

$$f(\mathbf{x}_i) = a_0 + \sum_{k=1}^{n} a_k x_{i,k}. \tag{3.8}$$

The number of shape functions is now $L = n + 1$ with the problem's dimension n. Therefore, the first column of the matrix $\mathbf{X}$ contains just the constant value $X_{i,0} = 1$ while the other columns are identical to the columns of $\mathbf{x}$. Eq. 3.6 integrated in eq. 3.7:

$$\mathbf{X}^T\mathbf{f} = \mathbf{X}^T\mathbf{y} = \begin{bmatrix} \sum_{i=1}^{N} f_i \\ \sum_{i=1}^{N} x_{i,1} f_i \\ \sum_{i=1}^{N} x_{i,2} f_i \\ \vdots \\ \sum_{i=1}^{N} x_{i,k} f_i \\ \vdots \\ \sum_{i=1}^{N} x_{i,L} f_i \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_{i,1} y_i \\ \sum_{i=1}^{N} x_{i,2} y_i \\ \vdots \\ \sum_{i=1}^{N} x_{i,k} y_i \\ \vdots \\ \sum_{i=1}^{N} x_{i,L} y_i \end{bmatrix} \tag{3.9}$$

The first position contains a rule for the mean values of the actual and the approximated system responses:

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} f_i = \overline{f} \tag{3.10}$$

The mean of the values $x_{i,k}$ of a single parameter is simply

$$\overline{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{i,k}. \tag{3.11}$$

The *expected value* (operator "E") of every approximated system response f_i is

$$\mathsf{E}[f_i] = \overline{f} + a_k(x_{i,k} - \overline{x}_k) = \overline{y} + a_k(x_{i,k} - \overline{x}_k). \tag{3.12}$$

A multiplication with any value of a parameter and following summation is allowed:

$$\begin{aligned}\sum_{i=1}^{N} \mathsf{E}[x_{i,k} f_i] &= \sum_{i=1}^{N} \left[x_{i,k}\overline{y} + a_k(x_{i,k}^2 - \overline{x}_k x_{i,k})\right] = \sum_{i=1}^{N} x_{i,k}\overline{y} + a_k \sum_{i=1}^{N}(x_{i,k}^2 - \overline{x}_k x_{i,k}) \\ &= \mathsf{E}\left[\sum_{i=1}^{N} x_{i,k} f_i\right] = \mathsf{E}\left[\sum_{i=1}^{N} x_{i,k} y_i\right]\end{aligned} \tag{3.13}$$

The second row is a consequence of eq. 3.9. Together with

$$N\overline{x}_k\overline{y} = \sum_{i=1}^{N} \overline{x}_k y_i = \sum_{i=1}^{N} x_{i,k}\overline{y} = \sum_{i=1}^{N} \overline{x}_k\overline{y} \tag{3.14}$$

and

$$N\overline{x}_k^2 = \sum_{i=1}^{N} \overline{x}_k x_{i,k} = \sum_{i=1}^{N} \overline{x}_k^2, \tag{3.15}$$

eq. 3.13 may be expanded and converted:

$$\mathsf{E}\left[\sum_{i=1}^{N} x_{i,k} y_i\right] = \sum_{i=1}^{N}(x_{i,k}\overline{y} + \underbrace{[\overline{x}_k y_i - \overline{x}_k\overline{y}]}_{=0}) + a_k \sum_{i=1}^{N}(x_{i,k}^2 - \overline{x}_k x_{i,k} + \underbrace{[\overline{x}_k^2 - \overline{x}_k x_{i,k}]}_{=0}) \tag{3.16}$$

Changing the expectation operator to the regression coefficient a_k

$$\sum_{i=1}^{N} x_{i,k} y_i = \sum_{i=1}^{N}(x_{i,k}\overline{y} + \overline{x}_k y_i - \overline{x}_k\overline{y}) + \mathsf{E}[a_k] \sum_{i=1}^{N}(x_{i,k}^2 - 2\overline{x}_k x_{i,k} + \overline{x}_k^2), \tag{3.17}$$

solving to $\mathsf{E}[a_k]$ and gathering the sums provide the expected value for the regression coefficient in a very simple form:

$$\mathsf{E}[a_k] = \frac{\sum_{i=1}^{N}(x_{i,k} - \overline{x}_k)(y_i - \overline{y})}{\sum_{i=1}^{N}(x_{i,k} - \overline{x}_k)^2} \tag{3.18}$$

In comparison to eq. 2.12, the relation between the linear correlation coefficient and the regression coefficient using linear polynomials is evidently

$$\mathsf{E}[a_k] = \rho_{x_k y}\frac{\sigma_y}{\sigma_{x_k}}, \tag{3.19}$$

with σ_{x_k} as the standard deviation of the parameter values $\mathbf{x}_k = [x_i]_k$ and σ_y as the standard deviation of the system responses $\mathbf{y} = [y_i]$.
It is important to note that this equation contains just an expected value. The value of a_k is equal to the expected value $\mathsf{E}[a_k]$ if the entire population (normally consisting of an infinite number of possible parameter combinations $\mathbf{x}_i$) is used for computing $\rho_{x_k y}$, σ_{x_k} and σ_y.

3.4 Linear and Non-Linear Correlation Coefficient

In the case of a linear influence of a parameter x_k on the system response y, the standard deviation of the experiments σ_{x_k} is projected by means of the sensitivity a_k on the system response with the standard deviation σ_{y_k}:

$$\sigma_{y_k} = a_k\sigma_{x_k} \tag{3.20}$$

The variance $\sigma_{y_k}^2$ is obviously caused by the parameter x_k. It is always possible to associate a part of the system response's variance $\sigma_{y_k}^2$ with the influence of an arbitrary parameter x_k, even if the influence

is not linear.
Inserting 3.20 into 3.19 leads to the general definition of the correlation coefficient:

$$\mathsf{E}[\rho_{x_k y}] = \frac{\sigma_{y_k}}{\sigma_y} \tag{3.21}$$

The correlation coefficient relates the variance $\sigma^2_{y_k}$, caused by the influence of a parameter x_k (which doesn't have to be linear), to the total variance σ^2_y, caused by all parameters together.
Therefore, an arbitrary non-linear correlation coefficient for a parameter x_k and the system response y can be computed by performing a 1-dimensional regression

$$f^k(\mathbf{x}_i) = \sum_{r=0}^{M} b_r \phi_r^k(\mathbf{x}_i) = f_i^k \tag{3.22}$$

using M (1-dimensional) shape functions $\phi_r^k = \phi_r(x_k)$. In analogy to eq. 3.7, the regression coefficients $\mathbf{b} = [b_r]$ may be found in

$$\mathbf{b} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}, \tag{3.23}$$

using

$$\mathbf{f}^k = \mathbf{Z}\,\mathbf{b}. \tag{3.24}$$

Thus, the general correlation coefficient $\rho_{x_k y}$ is

$$\mathsf{E}[\rho^2_{x_k y}] = \mathsf{E}\left[\frac{\mathsf{Var}[f_i^k]}{\mathsf{Var}[y_i]}\right] = \mathsf{E}\left[\frac{\mathsf{Var}[f_i^k]}{\sigma_y^2}\right] \leq 1. \tag{3.25}$$

Var[ ] is the operator of the variance. Eq. 3.25 may be used for the analysis of "real" systems.

3.5 Total Linear and Non-Linear Correlation Coefficient

As mentioned above, each arbitrary system response may always be assumed to be dependent on a certain number of parameters in a purely deterministic way, so it could be described exactly by a unique surface (see sec. 3.1).
Therefore, established system parameters $\hat{x}_l$ (which are assumed to be directly controllable, "*design parameters*" [13]) have to be distinguished from not established parameters $\tilde{x}_m$ (which are assumed to be unknown, "*noise parameters*" [13]). As an example, buckling of a rotational symmetric tensile bar may be controlled directly by varying its cross-sectional area or the pushing force, but the buckling direction is only influenced by disturbances in a chaotic way (e. g. inhomogen material properties, geometrical inaccuracies, ...). Disturbances are usually not controllable and not established. Not established parameters may be defined for modelling of noise in real system responses. They cause a *process variation* of the system response [13].

If an arbitrary system depends on more than one parameter (established or not), the total variance σ^2_y of the system response can be separated in the variance caused by a single parameter x_k and a variance caused by all other parameters $x_{l\neq k}$. Therefore, the system response for any experiment may be displayed as

$$y_i = \overline{y} + y_{i,k} + y_{i,l\neq k} \tag{3.26}$$

with the (predefined) expected values

$$\mathsf{E}[y_i - \overline{y}] = \mathsf{E}[y_{i,k}] = \mathsf{E}[y_{i,l\neq k}] = 0. \tag{3.27}$$

Squaring of eq. 3.26 leads to

$$\mathsf{E}\left[(y_i - \overline{y})^2\right] = y^2_{i,k} + y^2_{i,l\neq k} \tag{3.28}$$

because $y_{i,k}$ and $y_{i,l\neq k}$ are independent from each other ($\mathsf{E}[y_{i,k}\ y_{i,l\neq k}] = 0$). Thus, the total expected variance is

$$\mathsf{E}[\sigma_y^2] = \sigma^2_{y_k} + \sigma^2_{y_{l\neq k}}. \tag{3.29}$$

This operation can be done for the variance $\sigma^2_{y_{l\neq k}}$ recursively so the overall variance σ^2_y of the system response is simply:

$$\mathsf{E}[\sigma_y^2] = \sum_{k=1}^{n} \sigma^2_{y_k} \quad \text{or} \quad \mathsf{E}\left[\sum_{k=1}^{n} \frac{\sigma^2_{y_k}}{\sigma^2_y}\right] = 1 \tag{3.30}$$

Inserting eq. 3.21 into eq. 3.30 results in

$$\mathsf{E}\left[\sum_{k=1}^{n}\rho_{x_k y}^2\right]=\begin{cases}1 & \text{for} \quad \forall\, y_i \neq y_0 = const.\\ n & \text{for} \quad \forall\, y_i = y_0 = const.\end{cases} \tag{3.31}$$

The first case is valid as long as at least one parameter influences the response y, respectively at least two different y_i have different values. In the other case every single correlation coefficient is exactly $\rho_{x_k y} = \pm 1$ (threshold value).

In an analysis of an arbitrary system, only the established parameters are recognized, so in eq. 3.30 the variances $\hat{\sigma}_{y_l}^2$ and $\tilde{\sigma}_{y_m}^2$, caused by the parameters $\hat{x}_l$ and $\tilde{x}_m$, are summed up to

$$\mathsf{E}[\sigma_y^2] = \sum_{l=1}^{n_{\text{est.}}}\hat{\sigma}_{y_l}^2 + \sum_{m=n_{\text{est.}}+1}^{n}\tilde{\sigma}_{y_m}^2 = \hat{\sigma}_y^2 + \tilde{\sigma}_y^2 \quad \text{or} \quad \mathsf{E}\left[\frac{\hat{\sigma}_y^2}{\sigma_y^2}\right] \leq 1 \tag{3.32}$$

because of

$$0 \leq \tilde{\sigma}_y^2 \leq \sigma_y^2. \tag{3.33}$$

This is nothing but the extension of eq. 3.21 to a multi-dimensional correlation analysis:

$$\mathsf{E}[\rho_{\mathbf{x}\,y}^2] = \mathsf{E}\left[\sum_{l=1}^{n_{\text{est.}}}\rho_{x_l y}^2\right] = \mathsf{E}\left[\frac{\hat{\sigma}_y^2}{\sigma_y^2}\right] \leq 1 \quad \text{for} \quad \forall\, y_i \neq y_0 = const. \tag{3.34}$$

For a "real" system (cp. eq. 3.25):

$$\mathsf{E}[\rho_{\mathbf{x}\,y}^2] = \mathsf{E}\left[\sum_{l=1}^{n_{\text{est.}}}\rho_{x_l y}^2\right] = \mathsf{E}\left[\frac{\mathsf{Var}[f_i]}{\mathsf{Var}[y_i]}\right] = \mathsf{E}\left[\frac{\mathsf{Var}[f_i]}{\sigma_y^2}\right] \leq 1 \tag{3.35}$$

$\rho_{\mathbf{x}\,y} = \rho_{\text{tot}}$ is the *total correlation coefficient* [8] that is better known as $R^2 = \rho_{\text{tot}}^2$. f_i are the system responses associated to the experiments $\mathbf{x}_i$, approximated by arbitrary (linear or non-linear) shape functions (eq. 3.4).
Ineq. 3.34 or 3.35 implicate that the absolute value of each correlation coefficient $\rho_{x_k y}^2$ decreases with an increasing number of significant parameters.

4 Conclusions

- Estimated values of arbitrary statistical quantities must be regarded always together with their corresponding confidence intervals.
- Parameters with a large influence on the system response cause large absolute values of the corresponding correlation coefficients. Therefore, the correlation coefficient may be used for the **ranking of the parameter's importance with respect to the system responses**. In the case of linear correlation coefficients, their signs conform to the signs of the corresponding physical sensitivities or regression coefficients.
- "Linear" regression coefficients (= physical sensitivities) may be approximated using the correlation coefficients, the standard deviations of the parameter values and the standard deviations of the system response values (eq. 3.19). This is always possible. All of the three items are estimated, so the error of this operation will be gathered. Therefore, it is **recommended to compute the linear regression coefficients directly (eq. 3.7) in order to minimize their error of estimation**. However, this is only possible if the problem's number of degrees of freedom is positive ($N > n$) and the matrix $\mathbf{X}^T\mathbf{X}$ not singular.
- The absolute values of the correlation coefficients are limited due to the validity of ineq. 3.34 or 3.35. As mentioned in sec. 2.5.3, the accuracy of the correlation coefficients decreases with decreasing absolute values, too. Therefore, it is **recommended not to use a large number of significant parameters** for a given number of experiments, as far as this can be estimated before the analysis. This is contrary to the purpose of the robustness investigation and a lack of the correlation analysis. On the other hand, **the accuracy of each statistical quantity may always be increased by using a larger numbers of experiments**. High accuracies are represented by small confidence intervals.

- It should always be considered, that a pure regression analysis is an appropriate alternative to the pure correlation analysis, if the number of degrees of freedom is sufficient large. A parameter screening might be applied using the **Analysis of Variance (ANOVA)**, usually performed with a D-optimal design of experiments. ANOVA also provides confidence intervals for the regression coefficients [1], [9], [11], [12], [14].

5 References

[1] Craig K. J.; Stander N.; Dodge, D. A.; Varadappa, S.: Automotive crashworthiness design using response surface-based variable screening and optimization, Int. J. Comp.-Aided Engng and Sofw, 2005

[2] Gärtner, T.; Reuter, R.: Stochastische Crashsimulation mit LS-DYNA am Beispiel des Kopfaufpralls nach FMVSS 201. 17. CAD-FEM Users' Meeting, http://www.easi.de/company/publications/dyn_ralf/dyn_ralf.htm, 1999.

[3] Kersting, G.: Zufallsvariable und Wahrscheinlichkeiten. Eine Einführung in die Stochastik. http://ismi.math.uni-frankfurt.de/kersting/lecturenotes/Stochastik.pdf, 2005.

[4] Kokoska, S.; Zwillinger, D.: Standard Probability and Statistics Tabels and Formulae. Chapman & HallCRC, 2000.

[5] LS-DYNA Theoretical Manual. Livermore Software Technology Corporation, 1998.

[6] McKay, M. D.; Conover, W. J.; Beckman, R. J.: A Comparison of Three Methods for Selection Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics, 1979.

[7] McKay, M. D.: Latin Hypercube Sampling as a Tool in Uncertainty Analysis of Computer Models. Proceedings of the 24th Conference on Winter Simulation. ACM Press, 1992.

[8] Mühlbach, G.: Repetitorium der Wahrscheinlichkeitsrechnung und Statistik. Binomi, 2000.

[9] Myers, R. H.; Montgomery, D. C.: Response Surface Methodology: Process and Product Optimization Using Designed Experiments. John Wiley & Sons Inc, New York, 1995.

[10] Stander, N.; Eggleston, T.; Craig, K.; Roux, W.: LS-OPT Users's Manual. A Design and Probabilistic Analysis Tool for the Engineering Analyst. Livermore Software Technology Corporation, 2004.

[11] Roux, W. J.; Stander, N.; Haftka, R. T.: Response Surface Approximation for Structural Optimization. Int. J. Numer. Meth. Engng, 1998.

[12] Stander, N.; Reichert, R.; Frank, T.: Optimization of Nonlinear Dynamical Problems Using Successive Linear Approximations in LS-OPT. 6^{th} International LS-DYNA Users Conference, Detroit, 2000.

[13] Roux, W. J.; Stander, N. Günther, F.; Müllerschön, H.: Stochastic Analysis of Highly Nonlinear Structures. Int. J. Numer. Meth. Engng., accepted July 2005

[14] Stander, N.: The Successive Response Surface Method Applied to Sheet-Metal Forming. First MIT Conference on Computational Fluid and Solid Mechanics, Cambridge, 2001.

