

# Statistische Versuchsplanung mit JMP – von der Klassik zur Moderne

David Meintrup  
Hochschule Ingolstadt  
Fakultät Maschinenbau  
Esplanade 10  
85049 Ingolstadt  
David.Meintrup@haw-ingolstadt.de

## Zusammenfassung

In der Statistischen Versuchsplanung (DoE – Design of Experiments) wird mit statistischen Methoden eine Serie von Einzelexperimenten aufgestellt, die den Versuchsraum optimal erfassen und eine effiziente Analyse der Versuchsergebnisse ermöglichen. So kann man für beliebige Anwendungsgebiete, in denen experimentell gearbeitet wird, Produkte und Prozesse verbessern, die Robustheit erhöhen und Kosten sparen.

In diesem Beitrag erläutern wir die Grundprinzipien statistischer Versuchspläne und stellen klassische Designs, wie Screening-Designs zur Auswahl wichtiger Faktoren und Optimierungsdesigns vor. Die moderne Versuchsplanung geht mit Hilfe von optimalen Plänen über diese klassischen Ansätze hinaus. Optimale Designs werden exakt auf die Bedürfnisse und Nebenbedingungen des jeweiligen Experiments zugeschnitten. Die Statistik-Software JMP, eine SAS-Lösung, deren Stärken in der Interaktivität und Visualisierung liegen, stellt alle benötigten Hilfsmittel für die Erstellung und die Analyse statistisch geplanter Versuche zur Verfügung.

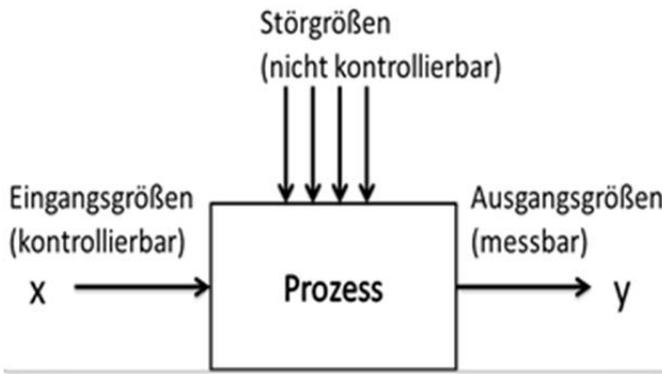
**Schlüsselwörter:** Statistische Versuchsplanung, Design of Experiments, optimale Versuchspläne, JMP

## 1 Einführung in die Versuchsplanung

### 1.1 Der Ansatz der statistischen Versuchsplanung

Experimente sind in fast jeder Wissenschaft das Mittel der Wahl, um neue Erkenntnisse über ein Produkt oder einen Prozess zu gewinnen. Das Ziel der Statistischen Versuchsplanung besteht darin, die einzelnen Versuche so zu planen und durchzuführen, dass man möglichst viel Information aus möglichst wenigen Versuchen erhält.

Wir gehen dabei von einem Prozessmodell aus, bei dem gewisse Eingangsgrößen durch ein System oder einen Prozess in einen Output verwandelt werden. Der Output kann mit Hilfe von messbaren Ausgangsgrößen  $y$  beschrieben werden. Diese Zielgrößen werden von kontrollierbaren Eingangsgrößen ( $x$ ) beeinflusst. Andere Größen sind nicht kontrollierbar oder auch unbekannt. Sie werden als Störgrößen bezeichnet und tragen zum Rauschterm bei.



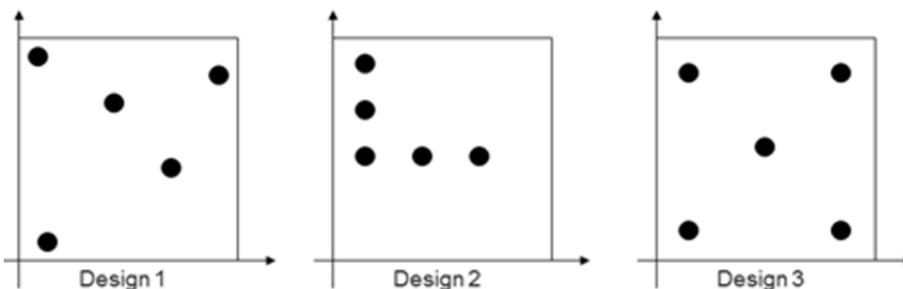
**Abbildung 1:** Prozessmodell mit Eingangs-, Stör- und Ausgangsgrößen

Wir gehen davon aus, dass es einen funktionalen Zusammenhang  $f(x)$  zwischen den Zielgrößen  $y$  und den kontrollierbaren Faktoren  $x$  gibt, so dass wir mathematisch die Situation durch eine Funktion

$$y = f(x) + \varepsilon$$

beschreiben. Dabei bezeichnet  $\varepsilon$  den Störterm, der unter anderem durch die nicht kontrollierbaren Faktoren entsteht. Unsere Aufgabe besteht darin, den funktionalen Zusammenhang  $f$  zwischen den Eingangs- und Ausgangsgrößen möglichst gut zu bestimmen.

Dabei verfolgen wir folgende Strategie: Wir legen bestimmte Experimente mit festgesetzten  $x$ -Werten fest, führen diese durch und messen die zugehörigen  $y$ -Werte. Anschließend passen wir das Modell  $f(x)$  möglichst gut an die Daten-Paare  $(x,y)$  an. Die entscheidende Frage der statistischen Versuchsplanung lautet: Welche Experimente, d.h. welche Festlegung von  $x$ -Werten wählt man aus? Gehen wir davon aus, dass wir zwei Eingangsgrößen  $x_1$  und  $x_2$  haben, die wir jeweils in einem Intervall variieren können, so ist der mögliche Versuchsraum ein Rechteck in der Ebene. Jeder Punkt dieses Rechtecks stellt ein mögliches Experiment dar. Es gibt unendlich viele Möglichkeiten, zum Beispiel 5 Einzelexperimente durchzuführen. In Abbildung 2 sind drei exemplarisch dargestellt. Welcher dieser Versuchspläne ist der geeignetste? Die Statistische Versuchsplanung verwendet statistische Methoden, um in den unterschiedlichsten experimentellen Situationen genau diese Frage zu beantworten.



**Abbildung 2:** Drei mögliche Versuchspläne für zwei Faktoren

## 1.2 Terminologie

Wie bereits beschrieben, besteht ein Gesamtexperiment aus der Durchführung mehrerer Einzelexperimente. Die Unterscheidung fällt im Englischen leichter, da dort die Einzelexperimente als „Runs“ bezeichnet werden. Jedes Einzelexperiment entspricht der Festlegung eines Satzes von Eingangsvariablen. Die Eingangsgrößen  $x$  eines Prozesses, die unabhängig voneinander variiert werden können, heißen Faktoren. Die Werte, auf die ein Faktor während eines Experiments gesetzt wird, bezeichnet man als Stufen. Die Faktoren variiert man im Laufe eines Experiments in einem bestimmten zulässigen Bereich. Alle Punkte, in denen jeder Faktor sich im zulässigen Bereich befindet, bilden zusammen den Versuchsraum. Im einfachsten Fall, wenn jeder der  $N$  Faktoren unabhängig von allen anderen in einem Intervall verändert werden darf, ist der Versuchsraum ein ( $N$ -dimensionaler) Quader. Die Variablen, die am Ausgang des Prozesses gemessen werden, heißen Zielgrößen. Der Einfluss, den ein Faktor auf eine Zielgröße besitzt, wird als Effekt bezeichnet. Unter einem Modell versteht man die Funktion  $f$ , die den Zusammenhang zwischen den Faktoren und den Zielgrößen beschreibt. Alle Einflüsse, die nicht durch das Modell erfasst werden, bilden den Rausch- oder Störterm. Dieser kann sowohl zufällig streuende Variablen enthalten als auch z. B. technische Größen, deren Einfluss nicht bekannt ist. Für einen realisierten Einzelversuch heißt die Differenz zwischen dem vom Modell vorausgesagten Wert und dem tatsächlichen Messwert das Residuum.

## 1.3 Modelle

Allgemein beschreiben wir den Zusammenhang zwischen den Zielgrößen und den Faktoren durch eine Funktion

$$y = f(x) + \varepsilon$$

Die Funktion  $f$  bezeichnet man als Modell. Ziel ist es, das Modell  $f$  möglichst gut zu bestimmen. Es ist dabei nicht sinnvoll, alle Funktionen  $f$  zur Konkurrenz zuzulassen. Zum einen würde dies die Gefahr des Overfitting bergen: Das berechnete, eventuell sehr komplizierte Modell würde den gegebenen Datensatz gut beschreiben, aber eben nicht die Natur hinter dem experimentell untersuchten Phänomen. Zum anderen würde die Interpretierbarkeit des Modells verloren gehen.

Es ist vernünftig, davon auszugehen, dass die Funktion  $f$  in eine Taylor-Reihe entwickelbar ist. Über einem beschränkten Versuchsraum kann man daher die Funktion  $f$  quadratisch approximieren. Im Falle von zwei Faktoren führt dies zu

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon$$

Dieses Modell ist aufgrund der quadratischen Terme in der Lage, ein Minimum oder Maximum der Zielgröße zu modellieren. Darüber hinaus enthält es Haupteffekte und Wechselwirkungsterme. Damit ist es sehr vielseitig einsetzbar, es ist das am meisten verwendete Modell für Optimierungszwecke.

Man kann das Modell weiter vereinfachen, indem die quadratischen Effekte gestrichen werden. So erhält man ein lineares Modell mit Wechselwirkungen

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

Streicht man aus diesem Modell die Wechselwirkungsterme, so erhält man das einfachst mögliche Modell, ein lineares Modell

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

das nur noch die Haupteffekte enthält. Beide Modelle werden verwendet, wenn das Ziel der Experimente darin besteht, zwischen wichtigen und unwichtigen Einflussgrößen zu unterscheiden. Sie werden im Allgemeinen die wahre Natur des Prozesses nur unzureichend wiedergeben, aber sie genügen, um die Faktoren nach Relevanz zu trennen. Durch ihre Einfachheit erlauben sie, mit vielen Faktoren gleichzeitig zu arbeiten. In obigen Formeln sind wir der Übersichtlichkeit halber von zwei Faktoren ausgegangen, aber natürlich lassen sich alle drei Modelle auf beliebig viele Faktoren verallgemeinern.

## 2 Klassische Versuchsplanung

### 2.1 Versuchspläne für Screening-Experimente

Es gibt in der Versuchsplanung stets einen Zusammenhang zwischen dem Ziel des Experiments, der Modellklasse und dem Versuchsplan. Alle drei müssen aufeinander abgestimmt sein. In diesem Kapitel beschreiben wir Versuchspläne, die für Screening-Experimente eingesetzt werden. Ein Screening-Experiment wird in einer Situation eingesetzt, bei der viele Faktoren einen Einfluss auf eine oder mehrere Zielgrößen haben könnten. Man möchte die (wenigen) Faktoren herausfiltern, die tatsächlich einen signifikanten Einfluss auf die Zielgrößen haben. Die Tatsache, dass nur wenige Faktoren einen relevanten Einfluss auf die Zielgröße ausüben, ist eine Voraussetzung für den sinnvollen Einsatz eines Screening-Experiments.

Bevor wir die Versuchspläne vorstellen, beschreiben wir kurz die Schritte, die typischer Weise für eine erfolgreiche Durchführung eines statistisch geplanten Versuchs erforderlich sind (vgl. [4,9]):

1. **Problembeschreibung und Versuchsziel:** So offensichtlich dieser Punkt zu sein scheint, so schwierig erweist er sich in der Praxis. Gerade für die statistische Versuchsplanung ist es sehr wichtig, dass alle Beteiligten sich auf eine klare Problemdefinition einigen und ein spezifisches, messbares und realistisches Versuchsziel formulieren.
2. **Auswahl der Zielgrößen:** Die Zielgrößen sollen das Versuchsziel messbar repräsentieren. In der Regel sind mehrere Zielgrößen erforderlich. Darüber hinaus benötigt man für jede Zielgröße eine Aussage über ihre (Mess-)Genauigkeit. Ist diese unzureichend, werden ggf. nur große Effekte erkannt oder es werden mehr Einzelexperimente erforderlich sein.
3. **Auswahl der Faktoren:** Geht man zunächst von einer Gesamtliste aller möglichen Einflussgrößen aus, so kann man diese in drei Gruppen aufteilen: Störgrößen, die man nicht beeinflussen kann; Variablen, die nicht ins Experiment aufge-

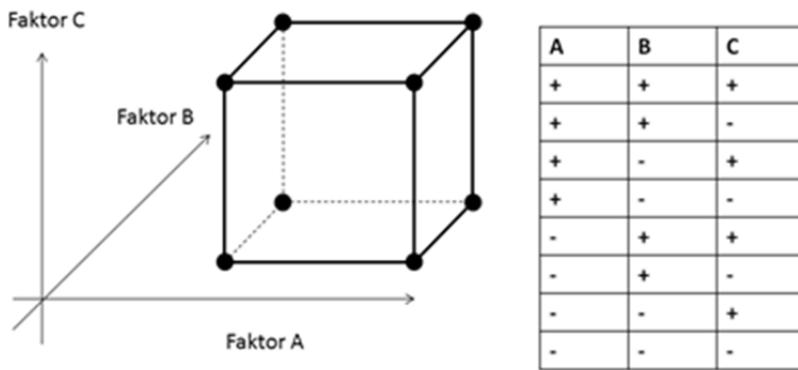
nommen werden (Diese sollten möglichst konstant gehalten werden, damit sie keinen Einfluss ausüben.); Faktoren, die im Experiment variiert werden sollen.

4. **Festlegung der Faktorstufen:** Steht die Liste der Faktoren fest, so muss entschieden werden, in welchem Bereich diese getestet werden sollen. Dabei muss das gesamte Expertenwissen der Wissenschaftler zum Einsatz kommen, um z.B. nicht realisierbare Faktorstufenkombinationen auszuschließen. Am Ende dieser Phase sollten das Versuchsziel, die Zielgrößen, die Faktoren mit ihren zulässigen Bereichen und ihre Stufen feststehen.
5. **Erstellung des Versuchsplans:** In Abhängigkeit des Versuchsziels, der Anzahl Faktoren und der Fallzahlplanung wird ein statistischer Versuchsplan erstellt. Die Einzelversuche werden randomisiert, um systematische Fehler zu vermeiden.
6. **Durchführung des Versuchs:** Die Einzelexperimente werden in der randomisierten Reihenfolge durchgeführt. Dabei ist aufgrund der möglichst niedrigen Zahl der Experimente besondere Sorgfalt erforderlich.
7. **Analyse der Daten:** die Ergebnisse werden mit Hilfe statistischer Software analysiert und validiert. Je nach Versuchsziel werden wichtige Faktoren identifiziert, Vorhersagen getroffen o.ä..

In einem Screening-Experiment werden lineare Modelle mit und ohne Wechselwirkungen verwendet. Diese können mit vielen Faktoren umgehen und genügen, um die Größe des Effekts eines Faktors zu bestimmen. Die zugehörigen voll- und teilfaktoriellen Versuchspläne stellen wir nachfolgend vor.

### 2.1.1 Vollfaktorielle Versuchspläne

Für die Erstellung eines vollfaktoriellen Plans werden für jeden Faktor zwei oder mehr Stufen festgelegt. Anschließend werden alle möglichen Kombinationen der Faktorstufen als Einzelexperimente durchgeführt. Aufgrund dieser Kombinatorik wächst die Anzahl der Versuche sehr schnell mit der Zahl der Faktoren und der Faktorstufen. Da man diese Pläne aber gerade für viele Faktoren benötigt, beschränkt man sich oft auf zwei Faktorstufen. Man wählt also für jeden Faktor eine niedrige Stufe (-) und eine hohe Stufe (+) aus. Für  $n$  Faktoren erhält man daher  $2^n$  Einzelexperimente. Diese entsprechen geometrisch den Ecken eines  $n$ -dimensionalen Würfels, der zugehörige Versuchsplan wird als  $2^n$ -Plan bezeichnet. Abbildung 3 zeigt einen  $2^3$ -Plan für drei Faktoren sowohl graphisch repräsentiert als auch in Form einer Tabelle.



**Abbildung 3:** Vollfaktorieller Versuchsplan mit drei Faktoren

Die Tabelle in Abbildung 3 zeigt die Experimente in systematischer Reihenfolge, d.h. von +++ zu ---. Für die Durchführung des Experiments werden die Versuche stets randomisiert, also in eine zufällige Reihenfolge gebracht, um systematische Fehler zu vermeiden. Vollfaktorielle  $2^n$ -Pläne haben sehr gute statistische Eigenschaften, so sind sie u.a. balanciert (jede Faktorstufenkombination tritt gleich oft auf) und orthogonal (Parameterschätzer sind unkorreliert).

Setzt man jeden Faktor auf den Mittelwert der hohen und niedrigen Faktorstufe, so erhält man einen so genannten Mittelpunktversuch. Drei oder vier Mittelpunktversuche gehören in der Regel in jeden Versuchsplan, da sie erlauben, die Linearitätsannahme der Modelle zu überprüfen und eine Fehlerabschätzung vorzunehmen.

### 2.1.2 Teilfaktorielle Versuchspläne

Vollfaktorielle Versuchspläne spielen in der Praxis eine untergeordnete Rolle, da mit steigender Zahl Faktoren die Anzahl der Experimente auch bei nur zwei Faktorstufen zu groß wird. So muss man für 7 Faktoren bereits  $2^7 = 128$  Experimente durchführen, was praktisch oft schon nicht mehr möglich ist.

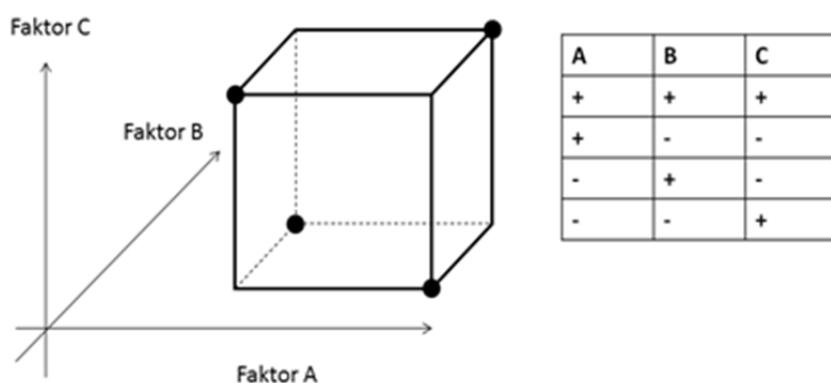
Vollfaktorielle Pläne sind in der Lage, sämtliche Wechselwirkungen zwischen den Faktoren zu schätzen, also bei vier Faktoren z.B. auch die Vierfachwechselwirkung  $x_1x_2x_3x_4$ . In fast allen experimentellen Situationen wird ein solcher Effekte aber vernachlässigbar sein. Daher ist die Idee nahe liegend, auf die Schätzbarkeit der Wechselwirkungen höherer Ordnung zu verzichten und dafür Experimente einzusparen. Genau so arbeiten teilfaktorielle Versuchspläne.

Teilfaktorielle Pläne werden als  $2^{n-k}$ -Pläne bezeichnet. So besteht z.B. ein  $2^{3-1}$ -Plan aus  $2^2 = 4$  Einzelexperimenten, die dadurch entstehen, dass man aus einem vollfaktoriellen  $2^3$ -Plan die Hälfte der Experimente streicht, wie in Abbildung 4 dargestellt. Ein  $2^{5-1}$ -Plan entsteht dadurch, dass man einen vollfaktoriellen  $2^5$ -Plan systematisch um die Hälfte reduziert. Die Halbierung der Einzelexperimente geschieht so, dass die erwünschten statistischen Eigenschaften, Balanciertheit und Orthogonalität, erhalten bleiben. Der  $2^{5-1}$ -Plan ist in der Lage, alle Haupteffekte und Zweifachwechselwirkungen zu schätzen, aber keine Wechselwirkungen höherer Ordnung. Diese sind in der Regel aber nicht aktiv. Für ein typisches Screening-Modell

$$y = \beta_0 + \beta_1x_1 + \dots + \beta_4x_4 + \beta_{12}x_1x_2 + \dots + \beta_{34}x_3x_4 + \varepsilon$$

kann man daher den teilfaktoriellen Versuchsplan mit nur  $2^{5-1} = 16$  Versuchen verwenden.

Natürlich zahlt man für die Reduktion der Versuchszahl einen Preis, dieser wird als Vermengung bezeichnet. Mit Vermengung bezeichnet man das Phänomen, dass man Effekte einzeln gar nicht schätzen kann, sondern tatsächlich stets die Summe mehrerer Effekte schätzt. Die Einzelheiten der Vermengung würden hier den Rahmen sprengen, den interessierten Leser verweisen wir auf die Literatur [4,9]. Wir beschränken uns auf den Hinweis, dass der Schweregrad der Vermengung durch die so genannte Auflösung gemessen wird. Pläne mit Auflösung III sind zu vermeiden, Auflösung IV kann gelegentlich zum Einsatz kommen, Auflösungen von V oder größer, wie sie z.B. der  $2^{5-1}$ -Plan besitzt, sind im Normalfall unproblematisch.



**Abbildung 4:** teilfaktorieller Versuchsplan für 3 Faktoren (Auflösung III)

## 2.2 Versuchspläne für Optimierungen

Wie bereits einleitend erwähnt, muss ein polynomiales Modell, das zur Optimierung verwendet wird, mindestens quadratischer Ordnung sein. Das nachfolgende Modell gilt für zwei Faktoren, ist aber leicht auf mehr Faktoren verallgemeinerbar:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_{12}x_1x_2 + \beta_{11}x_1^2 + \beta_{22}x_2^2 + \varepsilon$$

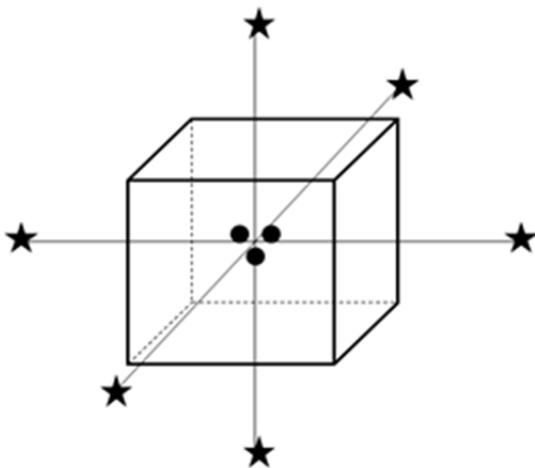
Den Graphen dieser Funktion bezeichnet man als Wirkungsfläche (Response Surface Methodology, siehe [10] für eine ausführliche Darstellung). Ziel des Experiments ist es, aus den Versuchsergebnissen die Parameter zu schätzen und damit die Wirkungsfläche zu beschreiben. Diese lässt sich sowohl graphisch als auch analytisch untersuchen. Die Kenntnis der Wirkungsfläche in der Nähe eines Prozesspunktes kann für verschiedene Fragestellung von Interesse sein:

- **Verständnis:** Durch graphische und analytische Beschreibungen der Wirkungsfläche kann man verstehen, wie sich die Zielgrößen in Abhängigkeit der Faktoren qualitativ und quantitativ verhalten.
- **Optimierung:** Mit Hilfe des Modells können eine oder mehrere Zielgrößen optimiert werden. Dabei kann es sich um Minimierungen, Maximierungen oder Ein-

stellungen auf einen Zielwert handeln, die gleichzeitig durchgeführt und untereinander gewichtet werden können.

- **Spezifikationen:** Die Wirkungsfläche kann dazu verwendet werden, ein Prozessfenster zu definieren, also einen Bereich, in dem die Faktoren schwanken dürfen, ohne dass z.B. kundenspezifische Spezifikationen für die Zielgrößen verletzt werden.
- **Robustheit:** Ziel einer Robustheitsstudie ist es, die Faktoren so einzustellen, dass kleinere Schwankungen aufgrund von Störgrößen oder nicht kontrollierbaren Einflüssen sich möglichst wenig auf die Zielgrößen auswirken. Gesucht sind also flache Bereiche auf der Wirkungsfläche.

Es gibt verschiedene Designs, die zur Anpassung von quadratischen Optimierungsmodellen zum Einsatz kommen. Allen gemeinsam ist, dass in jeder Dimension mindestens drei Faktorstufen einbezogen werden, da man diese für die Anpassung einer quadratischen Funktion benötigt. Das am weitesten verbreitete Design ist der zentral zusammengesetzte Versuchsplan (Central Composite Design, CCD), wie in Abbildung 5 für den Fall von drei Faktoren dargestellt.



**Abbildung 5:** Zentral zusammengesetzter Plan (CCD) für drei Faktoren

Ein CCD besteht aus drei Teilen. Sind  $n$  Faktoren zu berücksichtigen, so ist ein vollfaktorieller  $2^n$ -Plan ein Teil eines CCD. Hinzu kommen so genannte Axialpunkte mit den Koordinaten

$$(\pm a, 0, \dots, 0), (0, \pm a, 0, \dots, 0) \dots (0, \dots, 0, \pm a)$$

Diese sind in Abbildung 5 durch einen Stern gekennzeichnet. Verschiedene Abstände  $a$ , den die Axialpunkte vom Zentrum haben, liefern unterschiedliche Varianten des Versuchsplans, die wir im Anschluss diskutieren. Den dritten Teil eines CCD bilden Mittelpunktsversuche. Ihre Zahl kann je nach gewünschten Eigenschaften des Versuchsplans variieren, in Abbildung 5 sind drei Mittelpunktsversuche dargestellt.

Setzt man für die Axialpunkte  $a = 1$ , so bedeutet dies, dass die Axialpunkte auf den Seitenflächen liegen. Diese Variante, die auch als CCF (Central Composite on Face)

bezeichnet wird, verwendet man z.B. dann, wenn es aus technischen Gründen nicht möglich ist, größere Werte zu realisieren.

Bezeichnen wir mit  $n_0$  die Zahl der Mittelpunktversuche, so ergibt sich mit  $n$  Faktoren die Gesamtzahl der Versuche  $N$  durch

$$N = 2^n + 2n + n_0$$

Wählt man den Abstand  $a$  der Axialpunkte durch

$$a^2 = \frac{1}{2} \sqrt{2^n \cdot N} - 2^{n-1}$$

so ist der CCD orthogonal. Orthogonalität bedeutet, dass die Varianz-Kovarianz-Matrix der Modellparameter diagonal ist. Damit lassen sich die Modellparameter unabhängig voneinander schätzen und man erhält die kleinst möglichen Konfidenzintervalle für die Schätzer. Daher ist Orthogonalität eine wichtige Eigenschaft für einen Versuchsplan.

Alternativ kann man die Axialpunkte so setzen, dass die Größe der Konfidenzintervalle für die Zielgröße, die man vorhersagen möchte, nicht von der Richtung, sondern nur vom Abstand zum Mittelpunkt abhängt. Pläne, die eine solche Rotationssymmetrie aufweisen, nennt man rotierbar. Dies ist für ein CCD erfüllt, wenn

$$a^2 = \sqrt{2^n}$$

Kann man sich ein wenig Flexibilität in der Gesamtzahl der Einzelversuche erlauben, so kann man oft erreichen, dass der zentral zusammengesetzte Versuchsplan rotierbar und nahezu orthogonal ist. Für 3 Faktoren ist dies z.B. mit 9 Mittelpunktversuchen und  $a^2 = \sqrt{8}$  der Fall. Die Gesamtzahl der Einzelexperimente liegt dann bei 23.

### 2.3 Fallbeispiel: Papier-Helikopter

Zur Illustration der klassischen Versuchsplanung stellen wir in diesem Abschnitt ein Beispiel vor, das von einem der bedeutendsten Statistiker des 20. Jahrhunderts eingeführt wurde, G.E.P. Box [3]. Ziel des Experimentes ist es, einen Papier-Helikopter so zu konstruieren, dass er aus einer vorgegebenen Höhe möglichst langsam zu Boden fällt, also seine Flugzeit maximal wird. Für die Auswertung dieses Experiments verwenden wir die Statistik-Software JMP, eine SAS-Lösung, deren Stärken in der Interaktivität und Visualisierung liegen. Mit JMP kann man schnell, einfach und interaktiv alle gängigen Verfahren der statistischen Datenanalyse durchführen. Daher ist es insbesondere für Nicht-Statistiker als Statistik-Werkzeug gut geeignet. Im Bereich der statistischen Versuchsplanung liegt ein besonderer Schwerpunkt von JMP. Zum einen bietet die Plattform Design of Experiments (DoE) die gesamte Bandbreite der klassischen Verfahren der Versuchsplanung, von Screening-Designs über Mischungsdesigns bis zu Taguchi-Feldern. Zum anderen sorgt der Chef-Entwickler für statistische Versuchsplanung bei JMP Bradley Jones, einer der weltweit führenden Wissenschaftler auf dem Gebiet der modernen Versuchsplanung [5,6,7], dafür, dass alle Methoden für optimale Designs in JMP implementiert werden.

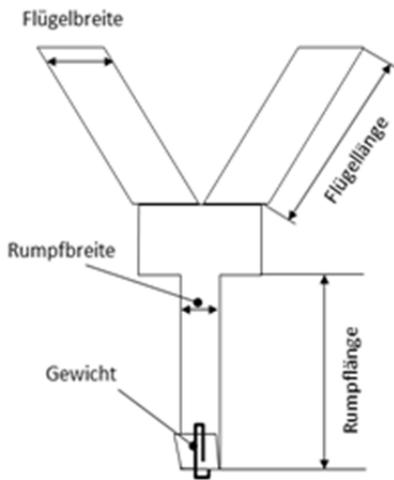


Abbildung 6: Papier-Helikopter mit 5 Faktoren

Tabelle 1: teilfaktorieller  $2^{5-1}$  Versuchsplan für Papier-Helikopter

Muster	Flügel- länge	Flügel- breite	Rumpf- Länge	Gewicht	Rumpf- breite	Flugzeit [sek]
+++++	10	4	9	2	3	3,12
+++--	10	4	9	1	2	2,65
+++-	10	4	6	2	2	3,54
++-+	10	4	6	1	3	2,99
+---	10	2	9	2	2	2,7
+--+	10	2	9	1	3	2,11
+-++	10	2	6	2	3	2,92
+---	10	2	6	1	2	2,16
--++	6	4	9	2	2	2,43
--+-	6	4	9	1	3	1,93
-+--	6	4	6	2	3	2,29
-+---	6	4	6	1	2	1,89
---++	6	2	9	2	3	1,98
---+-	6	2	9	1	2	1,6
----+	6	2	6	2	2	2,02
-----	6	2	6	1	3	1,89
00010	8	3	7,5	1	2,5	2,31
00010	8	3	7,5	1	2,5	2,36
00020	8	3	7,5	2	2,5	2,58
00020	8	3	7,5	2	2,5	2,54

Der Papier-Helikopter ist in Abbildung 6 dargestellt. Wir haben uns für 5 Faktoren entschieden, die Flügelgröße und -breite, die Rumpflänge und -breite, sowie das zusätzliche Gewicht, das durch eine oder zwei Büroklammern angebracht wird. Der zugehörige teilfaktorielle  $2^{5-1}$ -Plan ist zusammen mit der gemessenen Flugzeit in Tabelle 1 dargestellt. Alle Längenangaben sind in cm, das Gewicht hat die Stufen eine oder zwei Büro-

klammern. Zusätzlich zu den 16 Versuchen, die der  $2^{5-1}$ -Plan liefert, wurden vier Mittelpunktversuche, je zwei mit einer und mit zwei Büroklammern durchgeführt.

Das volle Modell besteht aus insgesamt 16 Effekten: dem Achsenabschnitt  $\beta_0$ , den 5 Haupteffekten und 10 Zweifachwechselwirkungen. Eine Analyse dieser Effekte mit graphischen und analytischen Mitteln zeigt, dass von den 16 Effekten vier einen signifikanten Einfluss auf die Flugzeit haben. Es handelt sich um die drei Haupteffekte Flügellänge, Flügelbreite und Gewicht sowie die Wechselwirkung von Flügellänge und Flügelbreite. Ihre Parameterschätzer und p-Werte sind in Tabelle 2 dargestellt.

**Tabelle 2:** Signifikante Effekte im Screening-Modell

Term	Schätzer	t-Wert	p-Wert
Achsenabschnitt	2,4	67,70	<,0001*
Flügellänge ( $x_1$ )	0,385	9,71	<,0001*
Flügelbreite ( $x_2$ )	0,216	5,45	<,0001*
Gewicht ( $x_3$ )	0,212	-5,96	<,0001*
Flügellänge*Flügelbreite ( $x_1x_2$ )	0,085	2,14	0,0488*

Die Wechselwirkung zwischen Flügellänge und Flügelbreite entspricht der Flügelfläche, so dass es auch aus physikalischer Sicht Sinn macht, dass sich diese im Modell für die Flugzeit wiederfindet. Das Screening-Modell, in dem sich nur noch die signifikanten Terme befinden, lautet jetzt:

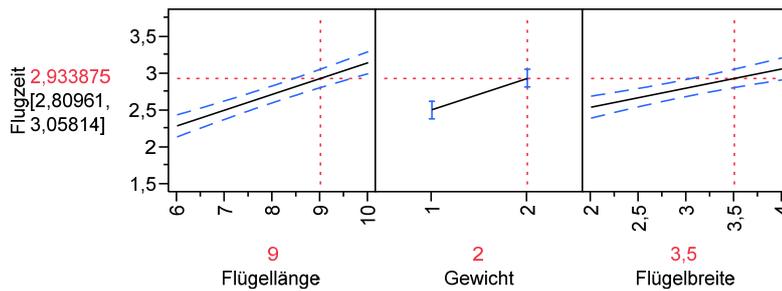
$$\hat{Y} = 2,4 + 0,385x_1 + 0,216x_2 + 0,085x_1x_2 + 0,212x_3$$

Dabei ist zu beachten, dass es sich hierbei um eine Gleichung für skalierte Faktoren handelt. Die hohe Faktorstufe (+), z.B. 10 cm Flügellänge, entspricht dem Wert  $x_1 = 1$ , die niedrige Faktorstufe (-), z.B. 6 cm Flügellänge, entspricht dem Wert  $x_1 = -1$ . Mit anderen Worten, die Variable  $x_1$  entsteht durch folgende lineare Transformation

$$x_1 = \frac{(\text{Flügellänge} - 8)}{2}$$

Genauso verfährt man mit allen anderen Faktoren. Ein Vorteil dieser Konvention besteht darin, dass sie die Effektgrößen untereinander vergleichbar macht. Die Schätzer für die Flügelbreite und das Gewicht sind in etwa gleich (0,21). Das bedeutet, dass der Wechsel von einer zu zwei Büroklammern als Gewicht etwa den gleichen Effekt auf die Flugzeit hat wie die Veränderung der Flügelbreite von der niedrigen (2cm) auf die hohe Stufe (4cm).

## Analysediagramm - Vorhersageanalyse



**Abbildung 7:** Analysediagramm der Flugzeit

Das Analysediagramm in JMP, siehe Abbildung 7, stellt das Screening-Modell graphisch dar. Zum einen erlaubt diese Darstellung, die Effektstärken graphisch zu erfassen und miteinander zu vergleichen. Zum anderen kann man die Faktorwerte interaktiv einstellen, und dabei beobachten, wie sich die anderen Effekte und die Zielgröße verhalten. So kann man in der Abbildung 7 ablesen, dass die vorhergesagte Flugzeit für einen Helikopter mit Flügellänge 9 cm, Flügelbreite 3,5 cm und zwei Büroklammern bei 2,93 Sekunden (95%-Konfidenzintervall [2,81; 3,06]) liegt.

Eine interessante Weiterentwicklung des Versuchs zum Papier-Helikopter hat Annis [1] vorgenommen.

## 3 Moderne Versuchsplanung

### 3.1 Optimale Versuchspläne

Die von uns bisher vorgestellten Versuchspläne haben sehr gute statistische Eigenschaften wie Orthogonalität, Minimierung von Konfidenzbereichen u.ä., solange keine ungewöhnlichen Randbedingungen vorliegen. So könnte z.B. der Versuchsraum irregulär sein, das Modell aus wissenschaftlichen Gründen andere, z.B. nicht-polynomiale Terme enthalten, es können kategoriale Faktoren vorkommen oder eine bestimmte Anzahl Einzelexperimente vorgeschrieben sein. Man könnte in diesen Fällen versuchen, die Randbedingungen so zu verändern, dass man mit einem klassischen Design arbeiten kann. Die moderne Versuchsplanung geht den umgekehrten Weg: ausgehend von allen Randbedingungen wird ein einzigartiges Design erzeugt, das speziell auf genau diese experimentelle Situation zugeschnitten ist. Der Slogan der modernen Versuchsplanung lautet daher:

Fit the design to the problem, not the problem to the design!

Optimale Versuchspläne sind zu unterscheiden von Versuchsplänen für Optimierung, wie wir sie im vorherigen Kapitel besprochen haben. Der Begriff „optimal“ bezieht sich auf den Auswahlprozess des Designs, nicht auf das Ziel des Experiments. So werden optimale Versuchspläne häufig für Screening-Experimente eingesetzt.

Vor der Erzeugung eines optimalen Versuchsplans muss man definieren, was man unter Optimalität versteht. Genauer bedeutet dies, dass man ein Kriterium angeben muss, das

in der Lage ist, für zwei gegebene Versuchspläne zu entscheiden, welcher der bessere ist. Wir gehen im nächsten Abschnitt näher auf die Optimalitätskriterien ein. Dann sind alle Randbedingungen festzulegen, die für das zu betrachtende Experiment gelten. Schließlich benötigt man einen Computer, der die algorithmische Arbeit erledigt. Der Computer erzeugt immer wieder neue Designs, die allen Randbedingungen genügen. Diese werden mit Hilfe des festgelegten Optimalitätskriteriums miteinander verglichen und das beste („optimale“) Design wird ausgewählt.

Die theoretische Entwicklung optimaler Designs entstand etwa 1960. Da die Auswahl und der Vergleich von Designs auch für eine relativ geringe Anzahl von Faktoren schon rechenintensiv ist, konnte sich diese Methode erst durchsetzen, seitdem leistungsfähige Computer für jedermann zur Verfügung stehen. Heute gehören optimale Design zum Standard und sind in jeder guten Software zur statistischen Versuchsplanung enthalten.

### 3.2 Optimalitätskriterien

In diesem abschließenden Abschnitt wollen wir etwas näher auf Optimalitätskriterien und ihren Einsatz eingehen. Wir setzen dabei ein wenig Vertrautheit mit der Theorie allgemeiner Regressionsmodelle voraus. Wir beginnen mit einem linearen Regressionsmodell in Matrixnotation:

$$Y = X\beta + \varepsilon$$

$Y$  ist ein  $N$ -dimensionaler Vektor, der die Zielgrößenwerte aus den  $N$  Einzelversuchen enthält.  $\beta$  enthält alle Modellparameter in einem Vektor. In der Modellmatrix  $X$  stehen in jeder Zeile alle Effekte, die in dem Modell betrachtet werden. Der gewöhnliche Kleinste-Quadrate-Schätzer für  $\beta$  ist

$$\hat{\beta} = (X'X)^{-1}X'Y$$

wobei  $X'$  die zu  $X$  transponierte Matrix bezeichnet. Die Varianz-Kovarianz-Matrix von  $\hat{\beta}$  ist gegeben durch

$$\text{var}(\hat{\beta}) = \sigma_{\varepsilon}^2(X'X)^{-1}$$

Die Fehlervarianz  $\sigma_{\varepsilon}^2$  können wir nicht beeinflussen, daher konzentrieren wir uns auf die Matrix  $(X'X)^{-1}$ . Wir wünschen uns, dass die Einträge in dieser Matrix betragsmäßig so klein wie möglich sind. Denn die Elemente neben der Diagonalen geben die Kovarianz zwischen zwei Modellparametern an, sie verhindern also die unabhängige Schätzung der Parameter. Die Diagonalelemente stellen die Varianzen der Modellparameter dar, sie steuern also die Größe der Konfidenzbereiche, und damit die Signifikanz von Hypothesentests usw. Nun ist der kleinst mögliche und realisierbare Wert für die Kovarianzen 0 und für die Varianzen  $1/N$ , wobei  $N$  die Anzahl der Einzelversuche ist. In diesem Fall ist die Varianz-Kovarianz-Matrix diagonal, und alle Diagonalelemente haben den Wert  $1/N$ . Dies ist äquivalent dazu, dass die Inverse  $(X'X)$  der Varianz-Matrix diagonal ist mit Diagonaleinträgen  $N$ . Ein Design, das diesem Kriterium genügt, heißt orthogonal. In einem orthogonalen Design hat die Determinante von  $(X'X)$ , die Informationsmatrix genannt wird, den Wert  $N^p$ , wobei  $p$  die Anzahl der Terme im Modell ist. Dies ist gleichzeitig der größte Wert, den die Determinante einer Informations-

matrix annehmen kann. In diesem Sinne kann man die Determinante von  $(X'X)$  als globales Maß für die Varianz des Parameterschätzers  $\hat{\beta}$  betrachten. Genau hier setzt das erste Optimalitätskriterium an. Unter allen möglichen Designs, die in einer gegebenen Situation allen Randbedingungen genügen, wählen wir dasjenige mit der größten Determinante der Informationsmatrix  $(X'X)$  aus. Ein solches Design heißt D-optimal, das Kriterium D-Optimalität (D wie Determinante).

D-Optimalität ist das geläufigste Kriterium für optimale Versuchspläne. Betrachtet man den  $p$ -dimensionalen gemeinsamen Konfidenzbereich des Schätzers  $\hat{\beta}$ , so minimiert ein D-optimaler Plan dessen Volumen. Daher werden D-optimale Versuchspläne in Screening-Experimenten mit linearen Modellen eingesetzt, bei denen das primäre Ziel in der Trennung von wichtigen und unwichtigen Faktoren und einer guten Schätzung der Modellparameter besteht.

Möchte man Versuchspläne miteinander vergleichen, so kann man z.B. ihre relative D-Effizienz bestimmen, in dem man ihre Determinanten ins Verhältnis setzt:

$$\text{relative D-Effizienz} = \frac{\det(X_1'X_1)^{\frac{1}{p}}}{\det(X_2'X_2)^{\frac{1}{p}}}$$

Beim Vergleich eines CCD mit einem D-optimalen Plan in einer Situation, die z.B. durch Nebenbedingungen zu einem irregulären Versuchsraum führen, kann man relative D-Effizienzen von 50% bekommen. Das bedeutet, dass man in dem CCD etwa doppelt so viele Versuche durchführen muss, um die gleiche Genauigkeit zu erreichen.

Ein weiteres Optimalitätskriterium ist die I-Optimalität (I steht hier für ein Integral). Minimiert wird bei I-optimalen Versuchsplänen die mittlere relative Varianz der Vorhersage über dem Versuchsraum. Die Herleitung der I-Optimalität ist etwas komplizierter als die D-Optimalität, daher verweisen wir den interessierten Leser auf die Literatur [2,8]. Da eine Vorhersagevarianz minimiert wird, werden I-optimale Pläne besonders dann verwendet, wenn eine präzise Vorhersage der Zielgrößenwerte wichtiger ist als die genaue Bestimmung der einzelnen Effektgrößen, also z.B. in Wirkungsflächenplänen zweiter Ordnung.

**Literatur**

- [1] Annis, D.: Rethinking the Paper Helicopter: Combining Statistical and Engineering Knowledge. *The American Statistician*, Vol. 59, No. 4, 2005.
- [2] A. Atkinson, A. Donev, R. Tobias: *Optimum Experimental Designs, with SAS*. Oxford University Press, New York, 2007.
- [3] G. Box: Teaching Engineers Experimental Design with a Paper Helicopter. *Quality Engineering*, 4, 1992.
- [4] G. Box, W. Hunter, J. Hunter: *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley, New York, 2005.
- [5] P. Goos, B. Jones: *Optimal Design of Experiments*. Wiley, New York, 2011.
- [6] B. Jones, P. Goos: D-optimal designs of split-split-plot experiments. *Biometrika*, 96, 2009.
- [7] B. Jones, C. Nachtsheim: A class of three-level designs for definitive screening in the presence of second-order effects. *Journal of Quality and Technology*, 43, 2011.
- [8] R. Meyer, C. Nachtsheim: The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, 37, 1995.
- [9] D. Montgomery: *Design and Analysis of Experiments*. Wiley, New York, 2008.
- [10] R. Myers, D. Montgomery, C. Anderson-Cook: *Response Surface Methodology*. Wiley, New York, 2009.

