4.1 Datenvorverarbeitung

Kapitel

4 Datenvorverarbeitung

- Übersicht, Sinn von Vorverarbeitung
  - Datenerhebung gut zu planen und durchzuführen bringt mehr als (zu) spät(er) lange an den Methoden zu feilen!
- Versuchsplanung
  - Welche Wertekonstellationen sollen untersucht werden?
  - Wieviele Versuche sind notwendig?
  - Welche Modelle sind günstig für die Auswertung?

4.1 Datenvorverarbeitung

- elementare Stichproben (inkl. Schichtung)
  - Grundgesamtheit repräsentativ erfassen
  - Schätzung mit möglichst geringer Varianz
  - Stichproben auch aus der Datenbank selbst, falls Datensätze für die Verarbeitung zu groß (bzgl. Speicher / Rechenzeit) sind!
- Behandlung fehlender Werte: später
- Datentransformation mit Beispiel: später

4.2 Versuchsplanung

Kapitel

4 Datenvorverarbeitung

4.2 Versuchsplanung: Einleitung und Orthogonalität

(aus: Weihs, Jessenberger, 1999)

Im Folgenden soll durch **systematisches Sammeln von Daten** mit **maximaler Information** mit möglichst geringem Aufwand ein möglichst vollständiges Verständnis bestimmter Zusammenhänge zwischen Einflussfaktoren und Zielgrößen erreicht werden.

4.2 Versuchsplanung

- **Statistische Versuchsplanung** hat das Ziel, mit möglichst wenigen Versuchen **diejenigen** Faktoren zu identifizieren, die einen signifikanten Einfluss auf eine Zielgröße haben.
- Bei der **Faktorreduktion** werden typischerweise **lineare Modelle** unterstellt, um diejenigen Faktoreffekte zu identifizieren, die deutlich größer als das Rauschen sind. Nur diese Faktoren werden dann weiter untersucht.

4.2 Versuchsplanung

Definition

Versuchsplan

Ein (**statistischer**) **Versuchsplan** ist eine bzgl. einer bestimmten Fragestellung optimale Zusammenstellung sogenannter **Versuche**, in denen jedem Faktor, von dem vermutet wird, dass er einen Einfluss auf ein Qualitätsmerkmal hat, ein Wert (ein sog. **Niveau**) zugewiesen wird.

4.2 Orthogonale Versuchspläne

Definition

Faktoreffekt

Ein Faktor habe in einem Versuchsplan nur zwei verschiedene Niveaus, die mit tief/hoch $\pm$ oder $-1/+1$ bezeichnet werden. Dann gilt:
Der (**Faktor-**)**Effekt** ist die Differenz zwischen dem Mittelwert der Zielgrößenwerte bei den Versuchen mit
"Faktor hoch" und dem Mittelwert der Zielgrößenwerte bei den Versuchen mit "Faktor tief".

4.2 Orthogonale Versuchspläne

- Der Effekt eines Faktors ist also der durchschnittliche Unterschied zwischen den Werten einer Zielgröße auf den beiden Niveaus dieses Faktors. Durch die Durchschnittsbildung werden die unterschiedlichen Niveaus der anderen Faktoren vernachlässigt, d.h. von Interesse sind nur die Einstellungen des betrachteten Faktors unabhängig von den Einstellungen der anderen Faktoren. So werden mögliche Einflüsse anderer Faktoren auf die Zielgröße quasi "ausgemittelt" und damit eliminiert.

4.2 Orthogonale Versuchspläne

Orthogonale Versuchspläne

- Bevor der Versuchsplan nach der Durchführung der Versuche ausgewertet werden kann, muss zunächst ein Maß für den Einfluss eines Faktors bestimmt werden. Der Einfluss eines Faktors auf die Zielgröße ist intuitiv gegeben durch die Veränderung der Zielgröße bei unterschiedlichen Einstellungen des Faktors.
- In diesem Abschnitt nehmen wir zunächst an, dass der Einflussfaktor **nur zwei Einstellungen** (Niveaus) hat und sein Einfluss durch den sog. Faktoreffekt quantifiziert werden kann.

4.2 Orthogonale Versuchspläne

- Bei zwei Niveaus kann man in einem Versuchsplan die echten Niveaus durch $\pm, -1/+1$ oder hoch/tief ersetzen. Eine solche **Kodierung** erhöht nicht nur die Übersichtlichkeit des Versuchsplans, sondern erleichtert auch das Rechnen. Außerdem ermöglicht erst sie es, mit den Niveaus von qualitativen Faktoren zu rechnen. **Kodierte Versuchspläne** sind Versuchspläne für sog. kodierte Faktoren.
- Wir beschränken uns nun auf Faktoren mit 2 unterschiedlichen Niveaus, dann kann man in einem Versuchsplan die echten Niveaus durch $\pm, -1/+1$ oder hoch/tief ersetzen.

4.2 Orthogonale Versuchspläne

Definition

Kodierung

Sei X ein quantitativer Faktor.
Der **kodierte Faktor** X_c ist definiert durch:

$$X_c := \frac{X - m}{d},$$

wobei m die Mitte und d die Halbspanne des Wertebereichs von X ist.
Also gilt: $X \in [m - d, m + d]$, und das ist äquivalent zu $X_c \in [-1, 1]$.

4.2 Orthogonale Versuchspläne

Falls Faktor X die Grenzen x_1 und x_2 seines Wertebereichs hat, dann sind **Zentrum** und **Halbspanne** gegeben durch

$$m := \frac{x_1 + x_2}{2} \text{ und } d := \frac{x_2 - x_1}{2}$$

und der kodierte Faktor ist

$$X_c := \frac{X - \frac{x_1+x_2}{2}}{\frac{x_2-x_1}{2}} = \frac{2X - (x_1 + x_2)}{x_2 - x_1}.$$

Modelle für den Zusammenhang von Einflussfaktoren und Zielgrößen werden oft in kodierten Faktoren formuliert und geschätzt.
Danach wird das Modell wieder in die Originalmerkmale umgeformt.
Dabei bleibt die Linearität in den Faktoren erhalten.

4.2 Orthogonale Versuchspläne

Ein **lineares Modell in kodierten Faktoren**

$$Y = \beta_1 + \beta_2 X_{c1} + \beta_3 X_{c2} + \epsilon$$

hat, ausgedrückt in Originalfaktoren, die Form:

$$\begin{aligned} Y &= \beta_1 + \beta_2 \frac{X_1 - m_1}{d_1} + \beta_3 \frac{X_2 - m_2}{d_2} + \epsilon \\ &= \left(\beta_1 - \beta_2 \frac{m_1}{d_1} - \beta_3 \frac{m_2}{d_2}\right) + \frac{\beta_2}{d_1} X_1 + \frac{\beta_3}{d_2} X_2 + \epsilon \\ &:= \tilde{\beta}_1 + \tilde{\beta}_2 X_1 + \tilde{\beta}_3 X_2 + \epsilon \end{aligned}$$

Es ist also auch ein lineares Modell in den Originalfaktoren.

4.2 Orthogonale Versuchspläne

- Bei ordinalen Faktoren mit **mehr als zwei Niveaus** werden die "extremen" Niveaus mit -1 und $+1$ kodiert,
- bei nominalen Faktoren beliebig vom Anwender ausgewählte Niveaus.
- **Qualitative Faktoren** werden in Modellen grundsätzlich kodiert verwendet, um damit rechnen zu können.
- Dabei ist ein Modellterm $5 \cdot S$ als -5 zu interpretieren, wenn $S = -1$ ist, also S das "kleinere" Niveau annimmt, und als $+5$, wenn $S = +1$ ist, also S das "größere" Niveau annimmt.
- Wir betrachten jetzt zunächst sogenannte **Screening-Pläne**, d.h. **lineare Modelle** in den kodierten Einflussfaktoren (**Haupteffekte**), um einfache Modelle zu bekommen, und damit kleine Versuchspläne.

4.3 Screening Pläne

Kapitel

4 Datenvorverarbeitung
4.3 Versuchsplanung: Screening

Definition: Ein **Screening-Modell** hat die Form:

$$y_i = \beta_1 + \sum_{j=1}^{L} x_{cij} \beta_{j+1} + \epsilon_i, \quad \epsilon_i \sim \text{u.i. } N(0, \sigma^2),$$

wobei

- y_i das Ergebnis der Zielgröße beim i-ten Versuch ist,
- x_{cij} das kodierte Niveau des j-ten Faktors im i-ten Versuch,
- β_1 der Achsenabschnitt (intercept, overall mean),
- β_{j+1} der Halbeffekt des j-ten Faktors (s.u.),
- ϵ_i der Fehler beim i-ten Experiment und
- σ^2 die Fehlervarianz.

4.3 Screening Pläne

- Die Fehler werden unabhängig identisch verteilt (u.i.v.) vorausgesetzt.
- In Matrixschreibweise gilt also:

$$y = X\beta + \epsilon, \quad X = \begin{pmatrix} 1 & \\ \vdots & A \\ 1 & \end{pmatrix}$$

wobei X die **Designmatrix** ist mit der **Planmatrix** A mit Spaltendarstellung

$$A = (x_{c1} \dots x_{cK}), \text{ wobei } x_{cj} = (x_{c1j} \dots x_{cnj})^T.$$

4.3 Screening Pläne

- X ist demnach eine Matrix mit lauter Einsen in der ersten Spalte und den Spalten der Matrix A. Außerdem ist

$$\beta := (\beta_1 \beta_2 \cdots \beta_{K+1})^T$$

der Vektor der unbekannten **Modellkoeffizienten** und analog $\epsilon := (\epsilon_1 \cdots \epsilon_n)^T$ der Fehlervektor.
- Die n **Zeilen der Planmatrix** A entsprechen den **Versuchen**, die K Spalten den beteiligten **Faktoren**.
- Jeder Faktor nimmt nur zwei Niveaus an, die mit -1 und $+1$ kodiert werden.
- Eine solche Planmatrix definiert einen **Screening-Plan**, wenn die kodierten Faktoren sämtlich **Mittelwert** 0 haben und paarweise **empirisch unkorreliert** sind!

4.3 Screening Pläne

- Man beachte, dass die Fehler ϵ_i neben den **Versuchsfehlern** auch noch die sog. **Modellfehler** umfassen, die dadurch entstehen, dass das lineare Modell die Wirklichkeit nur näherungsweise wiedergibt.
- Einen **Screening-Plan**, also einen Versuchsplan, der die Anforderungen an Mittelwert, empirische Varianz und Kovarianzen der kodierten Faktoren erfüllt, kann man wie folgt **konstruieren**:
  - Jede Spalte von A besteht aus **genau so vielen** -1 **wie** $+1$. Damit hat jeder Faktor den Mittelwert 0.
  - Jede Spalte von A steht orthogonal auf jeder anderen, d.h.

$$x_{cj}^T x_{ck} = 0, \quad j \neq k.$$

- Daraus folgt, dass $X^T X = n \cdot I$, wobei I die Einheitsmatrix ist.
- Daraus folgt offenbar die empirische **Unkorreliertheit** der Spalten, also der Faktoren, da jeder Faktor Mittelwert 0 hat.

4.3 Screening Pläne

- Die Orthogonalität kann man allerdings nur dann erreichen, wenn man mindestens $n > K$ Versuche hat.
- Die **Unkorreliertheit der kodierten Faktoren** ist eine wichtige Eigenschaft, denn nur sie garantiert, dass die **Kleinste-Quadrate-Schätzung** der unbekannten Koeffizienten eine einfache Form hat und dass die **Effekte unabhängig voneinander bestimmbar** sind.

4.3 Screening Pläne

Berechnung der Kleinste-Quadrate-Schätzung

Da $X^T X = n \cdot I$ ist, gilt:

$$\hat{\beta} = (X^T X)^{-1} X^T y = \frac{1}{n} X^T y.$$

An dieser Berechnungsformel sieht man, dass die Koeffizienten für die verschiedenen Einflussfaktoren unabhängig voneinander bestimmt werden können, weil die beteiligte Inverse eine Diagonalmatrix ist.

4.3 Screening Pläne

Tatsächlich hängt der Koeffizient des j-ten Faktors nicht von den Beobachtungen der anderen Faktoren ab, denn für die Schätzungen der einzelnen Koeffizienten erhält man durch Einsetzen der kodierten Niveaus:

$$\hat{\beta}_{j+1} = \frac{1}{n} \left(\sum_{i \text{ mit } x_{cij}=+1} y_i - \sum_{i \text{ mit } x_{cij}=-1} y_i \right), j = 1, \ldots, K.$$

Weil wegen der Voraussetzungen an Screening-Pläne die Anzahl der Einstellungen von -1 und $+1$ für jeden Faktor $\frac{n}{2}$ ist, ist der Effekt des j-ten Faktors gegeben durch:

$2\hat{\beta}_{j+1} =$ Effekt von Faktor j
$=$ Zielgrößenmittelwert bei den Versuchen mit "Faktor j hoch"
$-$ Zielgrößenmittelwert bei den Versuchen mit "Faktor j tief".

4.3 Screening Pläne

- Der geschätzte Koeffizient für den j-ten Faktor ist also nicht gleich dem Effekt des Faktors j, sondern nur gleich der Hälfte des Effekts, dem sog. **Halbeffekt**.
- Weiter gilt:

$$\hat{\beta}_1 = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y},$$

d.h. der Achsenabschnitt ist gleich dem Mittelwert der Versuchsergebnisse für die Zielgröße.

- Im Screening ist das Ziel die Identifizierung der wirklich einflussreichen Faktoren (**Faktorreduktion**). Es erhebt sich deshalb die Frage, wie die relevanten Einflussfaktoren von den unwichtigen getrennt werden können.

4.3 Screening Pläne

- Dazu können z.B. t-**Tests der Effekte** (d.h. der geschätzten Koeffizienten) auf Erwartungswert Null, etwa auf dem Niveau 20% (!), verwendet werden. Man beachte, dass im Screening in keinem Fall zu viele Faktoren eliminiert werden sollten. Deshalb wird ein Signifikanzniveau $> 5\%$ gewählt.
- Es reicht, die Tests für die Faktoren einzeln durchzuführen, da die Einflussfaktoren unkorreliert sind und die Effekte deshalb unabhängig voneinander bestimmt werden können.
- Außerdem ist die Varianz aller Faktoren wegen der Kodierung identisch $= \frac{\hat{\sigma}^2}{n}$, da allgemein die Kovarianzmatrix der Koeffizienten $= \hat{\sigma}^2 (X^T X)^{-1}$ ist.
- Damit wird bei einer schrittweisen Auswahl der relevanten Faktoren zuerst derjenige Faktor mit dem größten Effekt ausgewählt, dann derjenige mit dem zweitgrößen, usw. bis die Effekte nicht mehr signifikant sind.

4.3 Screening Pläne

- Diese **Faktorreduktion** ist bei Screening-Modellen erwünscht und aus mathematischer Sicht sogar notwendig, da in Screening-Plänen die Anzahl der Versuche n häufig nur 1 größer ist als die Anzahl Faktoren K, also $n = K + 1$.
- Dann sind zwar die Effekte berechenbar, allerdings ist bei $n = K + 1$ die Kleinste-Quadrate-Schätzung für die Fehlervarianz nur dann definiert, wenn nicht alle Faktoren in das Modell aufgenommen werden; denn

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - K - 1} = \frac{RSS}{n - K - 1},$$

wobei $e_i =$ geschätztes Residuum für Beobachtung i.
- Das prominenteste Beispiel für Screening Pläne sind die **Plackett-Burman-Pläne** (Plackett, Burman, 1946). Leider liegen Plackett-Burman-Pläne nur mit **Versuchsanzahlen** n vor, die ein Vielfaches von 4 sind.

4.3 Plackett Burman Pläne

- Die folgenden beiden Tabellen zeigen die Plackett-Burman-Pläne mit 4 bzw. 8 Versuchen:

	X_1	X_2	X_3
1	−	−	−
2	**−**	**+**	**+**
3	+	−	+
4	+	+	−

Tab. 1 : Plackett-Burman-Plan mit 4 Versuchen

4.3 Plackett Burman Pläne

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	−	−	−	−	−	−	−
2	**−**	**−**	**+**	**−**	**+**	**+**	**+**
3	+	−	−	+	−	+	+
4	+	+	−	−	+	−	+
5	+	+	+	−	−	+	−
6	−	+	+	+	−	−	+
7	+	−	+	+	+	−	−
8	−	+	−	+	+	+	−

Tab. 2 : Plackett-Burman-Plan mit 8 Versuchen

4.3 Plackett Burman Pläne

- Man versucht, immer den kleinstmöglichen Plackett-Burman-Plan zu verwenden, bei K Faktoren also einen Plan mit $K + 1$ Versuchen, da auch $K + 1$ unbekannte Koeffizienten vorliegen (K für die Faktoren und 1 Konstante).
- Falls $K + 1$ kein Vielfaches von 4 ist, werden einige Spalten des nächst größeren Plackett-Burman-Plans weggelassen, im Allg. diejenigen mit den meisten **Niveauänderungen**, da diese Spalten den größten Versuchsaufwand bedeuten.

4.3 Plackett Burman Pläne

- **Plackett-Burman-Pläne** lassen sich aus einer einzigen sog. "**Ausgangszeile**" mit Hilfe von zyklischen Permutationen konstruieren. Diese Zeile ist in den vorhergehenden Tabellen markiert.
- Die nächste Zeile des Plans ergibt sich durch Schieben der Zeile um eine Position nach rechts und durch Hinzufügen des "verloren gegangenen" Werts an den Anfang der Zeile.
- Der Plan wird komplettiert durch eine Zeile mit ausschließlich "-".

4.3 Plackett Burman Pläne

- Die nächster Tabelle zeigt die Ausgangszeilen für Plackett-Burman-Pläne mit 12, 16 und 20 Versuchen und höchstens 11, 15 bzw. 19 Faktoren.

n = 12:	+ − − − + − + + − + +
n = 16:	− − − + − − + + − + − + + + +
n = 20:	− − − − + − + − + + + + − − + + − + +

Tab. 3 : Ausgangszeilen für Plackett-Burman-Pläne mit n Versuchen

- Man beachte, dass in Plackett-Burman-Plänen in jeder Spalte genauso viele "+" wie "-" vorkommen und dass die Spalten untereinander und zu der Konstante orthogonal sind.
- Damit erfüllen Plackett-Burman-Pläne die Voraussetzungen für **Screening-Pläne**.

4.4 (D-)Optimale Versuchsplanung

Kapitel

4 Datenvorverarbeitung
4.4 Optimale Versuchsplanung

Optimale Versuchsplanung

- Die Motivation für die Screening-Pläne war, dass die Orthogonalität der Faktoren eine Identifizierung der wichtigen Faktoren besonders einfach macht.
- Neben solchen pragmatischen Gründen für die Wahl eines Versuchsplans gibt es vielfältige mathematische Optimalitätskriterien. Es gibt allerdings so viele unterschiedliche "Optimalitäten", dass man zur Bezeichnung fast das ganze Alphabet benötigt. Man spricht deshalb auch von alphabetischer Optimalität. Wir konzentrieren uns hier auf sogenannte **D-optimale Pläne**, die sogar generell **unkodierte Faktoren** verwenden.

4.4 (D-)Optimale Versuchsplanung

- Das verwendete (lineare) Modell hat die Form $y = X\beta + \epsilon$ mit beliebigen Faktoren und einer beliebigen Anzahl B von unbekannten Koeffizienten, wobei für die Komponenten ϵ_i des Modellfehlers gilt: $\epsilon_i \sim$ u.i.$N(0, \sigma^2)$.
- **D-Optimalität bedeutet, dass die Unsicherheit über die unbekannten Modellkoeffizienten β möglichst klein ist.**
- Im Fall von obigen Modellen sind die Unsicherheitsbereiche ellipsoidförmig (ellipsenförmig bei 2 unbekannten Koeffizienten, intervallförmig im 1-dimensionalen Fall). Man spricht von einem **Konfidenzellipsoid** (bzw. Konfidenzintervall).

4.4 (D-)Optimale Versuchsplanung

- **D-optimale Pläne minimieren das Volumen dieses Ellipsoids.**
- Genauer wird das Volumen des sogenannten $\alpha \cdot 100\%$-Konfidenzellipsoids minimiert. Dieses Konfidenzellipsoid ist eine mehrdimensionale Verallgemeinerung des $\alpha \cdot 100\%$-Konfidenzintervalls für einen einzelnen Modellkoeffizienten.
- Während sich Konfidenzintervalle auf einen einzelnen unbekannten Koeffizienten beziehen, charakterisieren Konfidenzellipsoide mehrere unbekannte Koeffizienten gleichzeitig.

4.4 (D-)Optimale Versuchsplanung

Definition

Konfidenzellipsoid

Das $\alpha \cdot 100\%$-**Konfidenzellipsoid** (Konfidenzellipsoid zum **Konfidenzniveau** α) gibt den Bereich um die B geschätzten Modellkoeffizienten $\hat{\beta}$ an, in dem die wahren Modellkoeffizienten β "mit mindestens $\alpha \cdot 100\%$ Wahrscheinlichkeit" liegen, wenn das Modell richtig ist.

4.4 (D-)Optimale Versuchsplanung

Bei dem Modell $y = X\beta + \epsilon, \quad \epsilon \sim$ u.i. $N(0, \sigma^2)$, hat das $\alpha \cdot 100\%$-**Konfidenzellipsoid** der B wahren Modellkoeffizienten β bzgl. der nach der Methode der kleinsten Quadrate geschätzten Koeffizienten $\hat{\beta}$ die Form:

$$(\beta - \hat{\beta})' X' X (\beta - \hat{\beta}) \leq B\hat{\sigma}^2 \cdot F_{B,n-B;\alpha},$$

wobei

$$\hat{\sigma}^2 := \frac{1}{n-B} \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

die geschätzte Varianz des Modellfehlers ist und $F_{B,n-B;\alpha}$ das α-Quantil der F-Verteilung mit B und $n - B$ Freiheitsgraden ist.

4.4 (D-)Optimale Versuchsplanung

Wenn zusätzlich $X^T X = n \cdot I$ ist, dann gilt für das Ellipsoid:

$$\sum_{j=1}^{B} (\beta_j - \hat{\beta}_j)^2 \leq \frac{B}{n} \hat{\sigma}^2 \cdot F_{B,n-B,\alpha}.$$

Wir haben es also mit einer **Kugel** um den geschätzten Koeffizientenvektor $\hat{\beta}$ zu tun!

Man kann zeigen, dass man zur Minimierung des Konfidenzellipsoids die Determinante der Matrix $X^T X$ maximieren muss (vgl. Weihs, Jessenberger (1999), S. 252 ff.).

4.4 (D-)Optimale Versuchsplanung

Definition

D-optimale Versuchspläne

Ein Versuchsplan mit der Designmatrix X heißt **D-optimal**, wenn das Volumen des dazugehörigen Konfidenzellipsoids für die unbekannten Modellkoeffizienten β minimal wird.

Sei D die Determinante der Matrix $X^T X$. Wenn $X^T X$ invertierbar ist, dann ist $D^{-0.5}$ proportional zum Volumen des Konfidenzellipsoids, und ein Plan ist D-optimal, wenn D für die dazugehörige Designmatrix X maximal ist für alle Versuchspläne im zulässigen Bereich.

4.4 (D-)Optimale Versuchsplanung

- Allerdings ist der absolut beste D-optimale Plan extrem schwierig zu konstruieren, da im Prinzip unendlich viele Pläne im Wertebereich verglichen werden müssten.
- Das D-Kriterium lässt sich aber als **relatives Kriterium** dazu verwenden, Versuchspläne bzgl. ihrer Eignung als Optimierungspläne zu bewerten.
- Üblicherweise gibt man sich eine Reihe von möglichen Versuchsplänen vor und bestimmt aus diesen den relativ D-besten Versuchsplan.
- Aber was sind sinnvolle Kandidaten für näherungsweise D-optimale Versuchspläne?
- Zunächst heißt es, die Vielfalt der möglichen Pläne einzuschränken.

4.4 (D-)Optimale Versuchsplanung

- Tatsächlich ist bekannt, dass D-optimale Pläne in gewisser Weise immer den vollständigen Planbereich ausnutzen, denn eine **Vergrößerung des Planbereichs** führt immer zu einer Vergrößerung des D-Kriteriums (vgl. Weihs, Jessenberger (1999), S. 255).
- Die Maximierung von D entspricht also der Maximierung des **Planbereichs**, d.h. desjenigen Bereichs, in den Planpunkte gelegt werden.
- In diesem Sinne verwenden D-optimale Pläne immer den gesamten zulässigen Bereich.
- Daher ist es sinnvoll, als **Kandidaten für Planpunkte** nur Punkte auf dem Rand des zulässigen Bereichs (und gewisse Mittelpunkte) zuzulassen.

4.4 (D-)Optimale Versuchsplanung

Definition

Ecken-Zentroid-Pläne

Ecken-Zentroid-Pläne verwenden als Kandidaten für Planpunkte ausschließlich

- Eckpunkte des zulässigen Bereichs,
- Kantenmittelpunkte des zulässigen Bereichs,
- Mittelpunkte von Begrenzungsflächen des zulässigen Bereichs und den
- Mittelpunkt des zulässigen Bereichs (overall centroid) oder den Referenzpunkt.

Ein Kandidat kann mehrmals gewählt werden.

4.4 (D-)Optimale Versuchsplanung

- In der Praxis werden häufig **D-optimale Ecken-Zentroid-Pläne mit einer vorgegebenen Anzahl von Versuchen** verwendet.
- Dazu wird natürlich eine Berechnungsvorschrift benötigt, die Planpunkte ihrer Wichtigkeit nach aus den Kandidaten auswählt, so lange, bis die gewünschte Anzahl Planpunkte erreicht ist.
- Hier wird die Methode von Mitchell (1974) vorgestellt, die zwar u. U. den optimalen Plan nicht findet, aber besonders leicht verständlich ist.

4.4 (D-)Optimale Versuchsplanung

Konstruktion

D-optimale Pläne bei vorgegebener Anzahl von Versuchen

Sei die Anzahl n der Versuche vorgegeben, dann gehe wie folgt vor:

1. Zufällige Auswahl eines Plans richtiger Größe aus den Kandidaten, so dass $X^T X$ invertierbar ist.
2. Hinzufügen des Kandidaten, der den D-Wert maximal vergrößert.
3. Weglassen des Kandidaten, der den D-Wert minimal verkleinert.
4. Wiederholung der Schritte 2 und 3, bis keine Verbesserung mehr möglich ist.
5. Austausch des ursprünglichen Plans durch den so konstruierten, falls der neue D-Wert größer ist.

4.4 (D-)Optimale Versuchsplanung

Dass D-optimale Pläne die Determinante von $X^T X$ maximieren, bedeutet, dass die lineare Unabhängigkeit der Spalten der Matrix X maximiert wird, und also die zu den Faktoren gehörenden Spalten von X so unkorreliert wie möglich sind!
Das erinnert uns an die Screening-Pläne, deren Spalten ja orthogonal sein sollten.

Tatsächlich kann man zeigen, dass die **Screening-Pläne D-optimal** sind!

4.5 Literatur – Versuchsplanung

- Mitchell, T.J. (1974): An algorithm for the construction of D-optimal experimental designs. Technometrics, 16, 203–210.
- Murphy, P.M., Aha, D.W. (1994): UCI repository of machine learning databases.
- Plackett, R. L., Burman, J. P. (1946): The design of optimum multifactorial experiments. Biometrica, 33, 305–325.
- Weihs, C., Jessenberger, J. (1999): Statistische Methoden zur Qualitätssicherung und -optimierung. Wiley-VCH, Weinheim.

4.5 Literatur – Versuchsplanung

Kapitel

4 Datenvorverarbeitung
4.5 Literatur – Versuchsplanung

- Atkinson, A.C., Donev, A.N. (1992): Optimum Experimental Design. Clarendon Press Oxford.
- Hedayat, A., Sloane, N., Stufken, J. (1999): Orthogonal Arrays. Springer Verlag, New York.
- Kohavi, R., John, G.H. (1998): The wrapper approach. In H. Liu and H. Motoda, (eds.). Feature Extraction, Construction and Selection: A data mining perspective, 33–50, Kluwer.

4.6 Elementare Stichprobentheorie

Kapitel

4 Datenvorverarbeitung
4.6 Elementare Stichprobentheorie

- Grundgesamtheit repräsentativ erfassen
- Schätzung mit möglichst geringer Varianz
- Stichproben auch aus der Datenbank selbst, falls Datensätze für die Verarbeitung zu groß (bzgl. Speicher / Rechenzeit) sind!

Idee:
Um eine Wahrscheinlichkeitsaussage über die Genauigkeit einer Vorhersage durch eine Stichprobe auf eine Grundgesamtheit zu bekommen, benötigt man eine unabhängige Zufallsauswahl bei der Wahl der Elemente der Stichprobe.

4.6 Zufallsstichproben

Definitionen:

- Die Gesamtheit der Elemente, über die Information gewünscht wird, heißt **Zielpopulation**.
- Eine Menge von identisch verteilten Zufallsvariablen $X_1, \ldots, X_N$ mit Dichte $f_X = f_{Xi}, i = 1, \ldots, N$, heißt eine (**Zufalls-**) **Stichprobe** der Größe N aus einer Grundgesamtheit, wenn für die gemeinsame Dichte gilt:

$$f_{X_1, X_2, \ldots, X_N}(x_1, x_2, \ldots, x_N) = f(x_1) f(x_2) \ldots f(x_N),$$

d.h. wenn die Zufallsvariablen unabhängig sind.

4.6 Zufallsstichproben

- Die Zufallsvariable X_i repräsentiert das i-te Element der Stichprobe in der **Reihenfolge des Ziehens**. Häufig werden nicht die Zufallsvariablen $X_1, \ldots, X_N$, sondern ein Satz ihrer **Realisierungen** $x_1, \ldots, x_N$ Zufallsstichprobe genannt. Da die X_i identisch verteilt sind, spricht man auch von **Werten** $x_1, \ldots, x_N$ **einer einzigen Zufallsvariable** X.
- Die Gesamtheit der Elemente, die in einer Stichprobe vorkommen können, heißt **Stichprobenpopulation**.
- Die gemeinsame Verteilung der Stichprobe $X_1, \ldots, X_N$ der Größe N heißt **Stichprobenverteilung**.

4.6 Geschichtete Stichproben

Idee:
Schätzer mit größerer Güte als "$\bar{X}$ bei Zufallsstichprobe" lassen sich z.B. konstruieren, wenn Informationen über die Struktur der Stichprobenpopulation bekannt sind, die in Zusammenhang mit derjenigen Zufallsvariablen stehen, deren Erwartungswert geschätzt werden soll.

Wenn z.B. die Körpergröße X in einer Population untersucht werden soll, dann erhält man durch die Zusatzinformation "Frau oder Mann" einen "informativen" Hinweis über die zu erwartende Körpergröße, da Männer im Durchschnitt größer sind als Frauen.

4.6 Geschichtete Stichproben

Diese Zusatzinformationen kann man ausnutzen, um die ("zu erwartende") mittlere Körpergröße in der Population zu schätzen. Dabei betrachtet man die Zusatzinformation als Ergebnis einer weiteren Zufallsvariablen Y auf der Stichprobenpopulation, zieht zunächst Zufallsstichproben aus den **Schichten**, d.h. denjenigen Teilen der Population, in denen Y einen festen Wert hat (also z.B. nur aus den Männern oder Frauen), bildet die dazugehörigen (bedingten) Erwartungswerte und erhält den (unbedingten) Erwartungswert von X durch Bildung des Erwartungswerts über die bedingten Erwartungswerte.

Falls man die Stichprobengröße in den Schichten "geschickt" wählt, ist die Güte dieses Schätzers größer als diejenige von "$\bar{X}$ bei Zufallsstichprobe".

4.6 Geschichtete Stichproben

Definition:

- Die Elemente B_l einer Partition $(B_1, \ldots, B_L)$ einer Grundgesamtheit Ω heißen **Schichten** (**Strata**) der Grundgesamtheit. Dabei habe die Schicht B_l ein bekanntes **Schichtgewicht** (Anteil!) $w_l, \sum_{l=1}^{L} w_l = 1$, an der Grundgesamtheit.

4.6 Geschichtete Stichproben

Beispiele: Schichten

- Geschlecht zur Schichtung von Studierenden bei der Bestimmung der mittleren Körpergröße
- Bundesländer zur Schichtung der Bevölkerung bei einer Wahlprognose,
- produzierende Maschinen zur Schichtung der Produkte bei der Qualitätskontrolle,
- Artikelgruppen zur Schichtung eines Lagerbestandes bei einer Inventur,

4.6 Geschichtete Stichproben

Beispiel: Effekt von Schichtung
Betrachte eine Grundgesamtheit vom Umfang $M = 7$, aus der eine Zufallsstichprobe der Größe $N = 3$ gezogen werden soll, um den Erwartungswert einer Zufallsvariablen X mit den folgenden Werten zu schätzen:

Objekt i	1	2	3	4	5	6	7
x_i	1	3	4	8	16	22	30

Beim Ziehen ohne Zurücklegen (**ACHTUNG : abhängige Auswahl**), erhält man bei Zufallsstichproben als minimalen bzw. maximalen Erwartungswertschätzer:

$$\bar{x}_{min} = \frac{1+3+4}{3} = \frac{8}{3}$$

und

$$\bar{x}_{max} = \frac{16+22+30}{3} = \frac{68}{3}.$$

4.6 Geschichtete Stichproben

Zerlegt man die Grundgesamtheit dagegen in 2 Schichten, so dass Schicht 1 := $\{x_1, x_2, x_3, x_4\}$ und Schicht 2 := $\{x_5, x_6, x_7\}$, und wählt zufällig 2 Elemente aus Schicht 1 und 1 Element aus Schicht 2, dann gilt für den minimalen bzw. maximalen Erwartungswertschätzer:

$$\bar{x}_{Smin} = \frac{1+3+16}{3} = \frac{20}{3}$$

und

$$\bar{x}_{smax} = \frac{4+8+30}{3} = \frac{42}{3}$$

Die **Spannweite der Schätzungen**, die ja ein Streuungsmaß ist, ist also **bei den geschichteten Stichproben deutlich kleiner**!

4.6 Geschichtete Stichproben

Definition

geschichtete Zufallsstichprobe

- Eine **geschichtete Zufallsstichprobe** der Größe N ist die Vereinigung von unabhängigen Zufallsstichproben der Größe N_l, $l = 1, \ldots, L$, $\sum_{l=1}^{L} N_l = N$, in den L Schichten einer Grundgesamtheit.

4.6 Geschichtete Stichproben

Satz:
Sei X eine Zufallsvariable auf der Grundgesamtheit Ω mit einer Schichtung mit Schichtgewichten w_l, $\sum_{l=1}^{L} w_l = 1$.
Für eine geschichtete Zufallsstichprobe gilt:

1. $$\bar{X}_S := \sum_{l=1}^{L} w_l \bar{X}_l$$
   ist ein erwartungstreuer Schätzer für $E[X]$
2. $$\hat{\text{var}}(\bar{X}_S) := \sum_{l=1}^{L} w_l^2 \hat{\text{var}}(\bar{X}_l)$$
   ist ein erwartungstreuer Schätzer für
   $$\text{var}(\bar{X}_S) = \sum_{l=1}^{L} w_l^2 \,\text{var}(\bar{X}_l),$$

4.6 Proportional Geschichtete Stichproben

Definition

proportional geschichtete Zufallsstichprobe

Eine geschichtete Zufallsstichprobe heißt **proportional geschichtete Zufallsstichprobe**, wenn gilt:

$$\frac{N_l}{N} = w_l, \quad l = 1, \ldots, L$$

4.6 Proportional Geschichtete Stichproben

Satz
Für eine proportional geschichtete Zufallsstichprobe gilt:

1. $$\bar{X}_S := \sum_{l=1}^{L} w_l \bar{X}_l = \bar{X},$$
2. $$\text{var}(\bar{X}_S) = \frac{1}{N} \sum_{l=1}^{L} w_l \sigma_l^2,$$

wobei

$$\sigma_l^2 := \text{var}(X|Y = l)$$

und

$$Y(\omega) := \text{ Schichtindex } (\omega), \omega \in \Omega.$$

4.6 Proportional Geschichtete Stichproben

Bemerkungen

- Bei proportionalen Stichproben ist der "Schichtenschätzer" $\bar{X}_S$ des Erwartungswerts gleich dem "Einfachstichprobenschätzer" $\bar{X}$! Aber die Güte ist im Allg. nicht gleich.
- Vor einem Gütevergleich dieser beiden Schätzer wollen wir noch versuchen, den Schichtenschätzer mit der kleinsten Varianz, d.h. mit der größten Güte, zu konstruieren. Eine solche Frage macht allerdings nur Sinn für **eine feste Größe N der Gesamtstichprobe**.

4.6 Optimal Geschichtete Stichproben

Definition

optimale geschichtete Zufallsstichprob

Eine geschichtete Zufallsstichprobe heißt **optimale geschichtete Zufallsstichprobe** bzgl. einer vorgegebenen Gesamtstichprobengröße N, wenn die Varianz des "Schichtenschätzers" $\bar{X}_S$ für die Größen N_l der Schichtstichproben minimal ist und wenn gilt:

$$\sum_{l=1}^{L} N_l = N.$$

4.6 Optimal Geschichtete Stichproben

Satz
Für eine optimale geschichtete Zufallsstichprobe gilt:

$$\frac{N_l}{N} = \frac{w_l \sigma_l}{\sum_{k=1}^{L} w_k \sigma_k}$$

wobei

$$\sigma_k^2 := \text{var}(X|Y = k) = \text{ Varianz in der } k\text{-ten Schicht} k = 1, \ldots, L.$$

4.6 Optimal Geschichtete Stichproben

Bemerkungen

- Je größer σ_l= Standardabweichung in der l-ten Schicht ist, desto größer sollte also $N_l :=$ Größe der Stichprobe in der l-ten Schicht gewählt werden.
- Zum Abschluss der Diskussion der geschichteten Stichproben soll diskutiert werden, ob das Ziel erreicht wurde, geschichtete Stichproben zu konstruieren, so dass $\bar{X}_S$ kleinere Varianz als $\bar{X}$ aus einer Zufallsstichprobe hat.

4.6 Optimal Geschichtete Stichproben

Satz:
Es gilt: $\text{var}(\bar{X}_{So}) \leq var(\bar{X}_{Sp}) \leq var(\bar{X})$.

Folgerung:

1. Der proportionale Schichtenschätzer ist nur optimal, wenn sämtliche Schichtvarianzen gleich sind, da dann $\sigma_1 = \sigma$.
2. Der Einfachstichprobenschätzer ist nur dann optimal, wenn sämtliche Schichterwartungswerte und sämtliche Schichtvarianzen gleich sind.

4.6 Optimal Geschichtete Stichproben

Bemerkungen:

- Offenbar können geschichtete Stichproben tatsächlich bessere Schätzer für den Erwartungswert einer Population liefern. Leider wird dazu **Information über** die **Schichten und** / oder die **untersuchte Zufallsvariable** benötigt.
- Für die proportionale Schichtung benötigt man die Schichtgewichte w_l.
- Für die optimale Schichtung sogar die Schichtvarianzen σ_l^2 oder zumindest eine gute Schätzung davon.