

Missing Values

Konzepte und statistische Literatur

Dipl.-Wirtsch.-Ing. Matthias Runte
Universität Kiel, Lehrstuhl für Marketing
Westring 425, 24098 Kiel
Tel 0431/880-1535
Email: matthias@runte.de
URL: <http://www.runte.de/matthias>

Inhaltsverzeichnis

<i>Tabellenverzeichnis</i>	<i>III</i>
<i>Abkürzungsverzeichnis</i>	<i>III</i>
<i>Symbolverzeichnis</i>	<i>III</i>
1 Problemstellung	1
2 Ursachen fehlender Daten	1
3 Ausfallmechanismen	2
4 Strukturanalyse	4
4.1 Deskriptive Analyse	5
4.2 Explorative Analyse	6
4.3 Induktive Analyse	7
5 Behandlung fehlender Daten	8
5.1 Eliminierungsverfahren	9
5.2 Imputationsverfahren	11
5.3 Parameterschätzverfahren	13
5.4 Multivariate Analyseverfahren	14
5.5 Sensitivitätsanalysen	15
6 Zusammenfassung	15
<i>Quellenverzeichnis</i>	<i>16</i>

Tabellenverzeichnis

<i>Tabelle 1 - Voraussetzungen für unsystematische Ausfallmechanismen</i>	3
<i>Tabelle 2 – Explorative Analyseansätze</i>	6
<i>Tabelle 3 – Missing-Data-Verfahren</i>	8

Abkürzungsverzeichnis

MD	Missing Data
MV	Missing Value(s)

Symbolverzeichnis

A	Datenmatrix
a_{ik}	Element der Datenmatrix
M	Merkmalsmenge
m	Anzahl Merkmale
N	Objektmenge
n	Anzahl Objekte
V	Indikatormatrix
v_{ik}	Element der Indikatormatrix

1 Problemstellung

Die Statistik und insbesondere die Multivariate Statistik wird überall dort eingesetzt, wo komplexes Datenmaterial multivariat ausgewertet werden muß (Hartung/Elpelt 1995, S. XIII). Dabei bilden insbesondere die multivariaten Analysemethoden die Fundamente der empirischen Forschung (Backhaus 1995, S. X). Fehlende Daten stellen dabei ein häufig anzutreffendes Problem dar. Die herkömmlichen, auf vollständigem Datenmaterial basierenden Analyseverfahren können in diesem Fall nicht mehr ohne weiteres angewendet werden. Daher ist es notwendig, sich mit der Auswirkung von fehlenden Daten zu beschäftigen und die daraus entstehenden Folgen in der angewandten Statistik zu berücksichtigen.

Die in dieser Arbeit zugrundegelegte Datenbasis stellt dabei eine endliche Objektmenge $N=\{1,\dots,n\}$ dar, deren Elemente anhand der Ausprägungen einer Merkmalsmenge $M=\{1..m\}$ charakterisiert werden. Alle Merkmalsausprägungen der Objekte werden in einer Datenmatrix $A=(a_{ik})_{n,m}$ zusammengefaßt, deren Zeilen durch die Objekte und deren Spalten durch die Merkmale repräsentiert werden. Die Matrix A kann fehlende Werte enthalten. Fehlende Werte werden durch einen punktförmigen Platzhalter (\bullet) dargestellt (Toutenburg 1992, S. 197).

In diesem Aufsatz sollen zunächst kurz die Ursachen für fehlende Daten umrissen werden. Hier stellt sich vor allen Dingen die Frage, ob mit einem systematischen oder unsystematischen Ausfallmechanismus gerechnet werden kann, in anderen Worten also, ob von einem zufälligen oder nicht zufälligen Fehlen der Daten ausgegangen werden muß. Da nur auf Basis eines bekannten Ausfallmechanismus eine angemessene Behandlung der fehlenden Werte durchgeführt werden kann, werden nachfolgend einige Ansätze zur Strukturanalyse der Datenmatrix dargestellt. Anhand der hierdurch gewonnenen Erkenntnisse über das unvollständige Datenmaterial werden nachfolgend Verfahren und Konzepte vorgestellt, mit denen die adäquate Aufbereitung der Datenbasis bzw. die Datenanalyse durchgeführt werden kann.

2 Ursachen fehlender Daten

Die Gründe für das Fehlen von Daten können vielfältig sein. In Anlehnung an Schnell (1986, S. 24-58) sind insbesondere die folgenden Gründe hervorzuheben:

- Fehlerhaftes oder mangelhaftes Untersuchungsdesign,
- Antwortverweigerung im Rahmen einer Befragung,

- Mangelndes Wissen oder unzureichende Antwortmotivation des Befragten,
- Unaufmerksamkeit eines Beobachters,
- Unvollständigkeit von Sekundärdaten,
- Codierungs- und Übertragungsfehler der Daten.

Beispielhaft für ein fehlerhaftes Untersuchungsdesign läßt sich das Merkmal „Alter der Kinder“ im Rahmen einer demografischen Erhebung nennen. Bei kinderlosen Befragten tritt hier zwingenderweise ein fehlender Wert auf, wenn diese Möglichkeit nicht anders berücksichtigt wird. Mangelhafte Untersuchungsdesigns liegen beispielsweise bei mißverständlichen Fragen oder ungebräuchlichen Redewendungen bei einer schriftlichen Erhebung vor. Ein sorgfältiges Untersuchungsdesign kann bereits im Vorfeld der Erhebung zu einer deutlichen Reduzierung oder gänzlichen Vermeidung von fehlenden Werten beitragen.

In diesem Zusammenhang lassen sich auch fehlende Werte durch die Verweigerung der Angaben durch die Befragten vermeiden, indem keine sensitiven oder zu persönlichen Fragen gestellt werden, ohne daß dem Befragten Anonymität zugesichert wird. Für Befragungen in besonders sensiblen Bereichen sind spezielle Verfahren wie z.B. die Randomized Response Technik entwickelt worden, auf die hier jedoch nicht näher eingegangen werden kann (vgl. Schnell et al. 1988, S. 312).

Im Vorwege ist ebenfalls sicherzustellen, daß der Kreis der befragten Personen sowohl über die notwendige Kompetenz als auch über eine ausreichende Motivation zur Beantwortung der Fragen verfügt.

3 Ausfallmechanismen

Die im vorigen Abschnitt beispielhaft angesprochenen Ausfallursachen können zu einer unvollständigen Datenbasis führen, also einer Datenmatrix mit fehlenden Werten. Den fehlenden Werten muß dabei durch Einbeziehen des jeweiligen Ausfallmechanismus Rechnung getragen werden. In der Literatur wird hier von **missing data mechanism** (Rubin 1976), **Ausfallmechanismus** (Bankhofer/Praxmarer 1998, S. 110), **response mechanism** oder **Antwortmechanismus** (Schwab 1991, S. 6) gesprochen.

Grundsätzlich sind zwei Arten von Ausfallmechanismen zu unterscheiden:

- Systematisch fehlende Daten (d.h. nicht zufällig fehlend)
- Unsystematisch fehlende Daten (d.h. zufällig fehlend)

Die Definitionen von systematischen bzw. unsystematischen Ausfallmechanismen gehen auf einen Aufsatz von Rubin (1976) zurück. Rubin bezeichnet Daten als **missing at random** (MAR), wenn das Fehlen der Daten unabhängig von der Ausprägung des Merkmals selbst ist. Daten gelten als **observed at random** (OAR), wenn das Fehlen der Daten unabhängig von den Ausprägungen anderer Merkmale im selben Objekt ist. Die Eigenschaft **missing completely at random** (MCAR) gilt schließlich für Daten, für die sowohl MAR als auch OAR gilt (Schwab 1991, S. 7). In diesem Falle hängt die Existenz der Daten weder von fehlenden noch von existierenden Merkmalen ab. Die Ausfallswahrscheinlichkeit der Daten besitzt also keinerlei Relation zur Existenz oder den Ausprägungen anderer Daten.

MAR	Missing at random. Antwortrate ist unabhängig von der Ausprägung des Merkmals selbst.
OAR	Observed at random. Antwortrate ist unabhängig von den Ausprägungen anderer Merkmale.
MCAR	Missing completely at random. MAR und OAR treffen zu.

Tabelle 1 - Voraussetzungen für unsystematische Ausfallmechanismen

Zur Verdeutlichung folgen einige Beispiele. In einer Befragung werde auch nach dem Einkommen gefragt. Einige Befragte verweigern dabei die Antwort. Die Daten sind MAR, falls die Ausfallswahrscheinlichkeit nicht von der Höhe des Einkommens selbst abhängt. Ist jedoch die Antwortverweigerungsrate bei Personen mit einem überdurchschnittlichen Einkommen höher als bei Personen mit niedrigem Einkommen, so gilt die Eigenschaft MAR nicht.

Möglich wäre auch, daß die Ausfallswahrscheinlichkeit zwar unabhängig vom Einkommen ist, aber von anderen Merkmalen wie z.B. dem Alter abhängt. In diesem Falle gilt zwar MAR; allerdings ist das Fehlen der Daten von den Ausprägungen anderer Merkmale abhängig, womit die Eigenschaft OAR verletzt wäre. Ist die Antwortrate für das Merkmal „Einkommen“ vollkommen unabhängig sowohl vom Einkommen selbst als auch von allen anderen Merkmalsausprägungen der befragten Person, dann und nur dann gilt MCAR.

Unsystematische Ausfallmechanismen liegen in Fällen von Einflußfaktoren vor, welche nicht unmittelbar mit einzelnen Merkmalen oder Objekten zusammenhängen. Zu den typischen Einflußfaktoren für unsystematische Ausfallmechanismen zählen beispielsweise Unaufmerksamkeit beim Ausfüllen von Fragebögen oder des Untersuchungspersonals bei der Datenaus-

wertung (Bankhofer/Praxmarer 1998). Zufällig fehlende Daten verzerren das Untersuchungsergebnis in der Regel kaum oder gar nicht.

Anders kann die mangelnde Berücksichtigung von systematischen Ausfallmechanismen erhebliche Verzerrungen im Rahmen der Datenauswertung zur Folge haben. Beispielsweise fiele die Schätzung des Einkommens auf Basis des Mittelwertes der vorhandenen Daten zu niedrig aus, wenn relativ gesehen mehr Personen mit höherem Einkommen die Antwort verweigert hätten. Man erkennt, daß die Berücksichtigung des Ausfallmechanismus in der Datenanalyse notwendig ist, falls ein systematischer Ausfallmechanismus vorliegt.

Einschränkend kann ein systematischer Ausfallmechanismus nur dann adäquat berücksichtigt werden, wenn dieser bekannt ist. So wäre es in dem Beispiel oben notwendig, die Ausfallwahrscheinlichkeit des Merkmals „Einkommen“ in Abhängigkeit von der Höhe des Einkommens festzustellen und möglichst genau zu modellieren.

Es wird deutlich, daß im Rahmen der Datenanalyse zunächst genaue Kenntnisse über den zugrunde liegenden Ausfallmechanismus gesammelt werden müssen. Nur so kann eine angemessene Berücksichtigung des Mechanismus erzielt werden.

4 Strukturanalyse

Im Rahmen der Strukturanalyse wird versucht, Erkenntnisse über die Abhängigkeit der Missing-Value-Struktur (MV-Struktur) vom Fehlen bzw. den Ausprägungen anderer Merkmale zu gewinnen.

Im Rahmen der Strukturanalyse ist jedoch nur ein Teil der Abhängigkeiten innerhalb des Datenmaterials überhaupt ermittelbar. Somit ergeben sich stets lediglich notwendige, aber keine hinreichenden Bedingungen für die Existenz eines unsystematischen Ausfallmechanismus. So ist es beispielsweise nur unter bestimmten Voraussetzungen möglich, Anhaltspunkte über das Vorliegen der Eigenschaft MAR zu gewinnen. Trotzdem lassen sich durch die Strukturanalyse wertvolle Erkenntnisse für eine ggf. notwendige Modellierung des Ausfallmechanismus oder die für Auswahl eines geeigneten Missing-Data-Verfahrens (MD-Verfahrens) gewinnen.

Bei der Strukturanalyse lassen sich prinzipiell drei Ansätze unterscheiden (Bankhofer 1995, S. 29 ff.):

1. Bei der **deskriptiven Analyse** lassen sich sogenannte MD-Maße (Missing-Data-Maße) berechnen, anhand derer das Verhältnis von fehlenden zu existierenden Werten und ggf. Konzentrationstendenzen von Missing Values innerhalb der Matrix ermittelt werden kön-

nen. Die Aussagekraft der deskriptiven Analyse beschränkt sich somit im wesentlichen auf das Ermitteln von Kennzahlen über die Struktur der fehlenden Werte.

2. Mit der **explorativen Analyse** lassen sich Abhängigkeiten innerhalb der Datenmatrix aufdecken. Dabei wird nach Zusammenhängen innerhalb der Daten gesucht, anhand derer auf Abhängigkeiten von Missing Values geschlossen werden kann.
3. Im Rahmen der **induktiven Analyse** werden statistische Tests durchgeführt. Dabei können u.a. Tests auf Konzentrationen von Missing Values in bestimmten Bereichen der Matrix sowie Tests auf einen unsystematischen Ausfallsmechanismus durchgeführt werden. Anzumerken ist, daß die induktive Analyse trotz allem nur notwendige Kriterien für das Vorliegen eines unsystematischen Ausfallsmechanismus liefern kann. Eine positive Bestätigung eines unsystematischen Ausfallsmechanismus ist nur in Sonderfällen und dann auch nur mit externem Wissen über die Verteilung der Missing Values möglich.

4.1 Deskriptive Analyse

Bei der deskriptiven Analyse sollen möglichst aussagefähige Kennzahlen in bezug auf die fehlenden Daten ermittelt werden. Im Rahmen dieser Arbeit kann nur kurz auf die prinzipielle Vorgehensweise eingegangen werden. Für detailliertere Darstellungen von Missing-Datamaßen wird auf die Literatur verwiesen (z.B. Bankhofer 1995, S. 31 f.).

Zur Berechnung der MD-Maße dient eine sog. Indikatormatrix V , die sich auf die Datenmatrix A bezieht und wie folgt definiert wird (Rubin 1976, S. 583):

$$V = \begin{bmatrix} v_{11} & \dots & v_{1m} \\ \dots & & \dots \\ v_{n1} & \dots & v_{nm} \end{bmatrix} \text{ mit } v_{ik}=1 \text{ falls } a_{ik} \text{ vorhanden, sonst } 0.$$

Die Indikatormatrix besteht also aus Binärvariablen, die die Existenz (1) oder das Fehlen (0) der Elemente a_{ik} der Datenmatrix A anzeigen. Aus der Indikatormatrix lassen sich eine Reihe von Kennzahlen ermitteln. Die Liste der in der Literatur gebräuchlichen Kennzahlen ist recht umfangreich. Exemplarisch seien aufgeführt:

- Absolute und relative Anzahl der fehlenden Daten bei Objekt i oder Merkmal k (Rummel 1970, S. 266),
- absolute und relative Anzahl fehlender Daten in der Datenmatrix,

- Indikatorvariablen, wie z.B. „Objekt i vollständig“, d.h. es liegen bei Objekt i keine fehlenden Daten vor (Cohen und Cohen 1975, S. 265 ff.),
- Korrelationskoeffizienten, die den Zusammenhang von paarweise fehlenden Objekten oder Merkmalen beschreiben (Brown 1983, S. 282 ff.). Hierfür werden zusätzliche Indikatorvariablen erzeugt, die die paarweise Existenz oder Nicht-Existenz von Daten indizieren.

4.2 Explorative Analyse

Im Rahmen der explorativen Analyse wird nach Zusammenhängen innerhalb der unvollständigen Datenmatrix gesucht. Die Untersuchung hat zum Ziel, ggf. Abhängigkeitsbeziehungen der fehlenden Werte aufzudecken. Wie auch bei der deskriptiven Analyse spielt in diesem Zusammenhang die Indikatormatrix V die dominierende Rolle. Tabelle 2 faßt die Ansätze der explorativen Analyse zusammen (Bankhofer 1995, S. 47).

Zielsetzung	Untersuchte Eigenschaft	Ausgangspunkt	Methoden
Abhängigkeit vom Fehlen der Daten (v_{ik}) bei anderen Merkmalen bzw. Objekten	MAR, OAR	Indikatormatrix V	Korrelationsanalyse Faktorenanalyse Clusteranalyse Dependenzanalyse
Abhängigkeit von von den Merkmalsausprägungen (a_{ik}) bei anderen Merkmalen bzw. Objekten	OAR	Datenmatrix A , Indikatormatrix V	Korrelationsanalyse Dependenzanalyse

Tabelle 2 – Explorative Analyseansätze

Bei den korrelationsanalytischen Ansätzen werden die Korrelationsbeziehungen der Indikatormatrix V mit sich selbst bzw. mit der Datenmatrix A untersucht. Lösel und Wüstendorfer (1974, S. 345) schlagen hierzu die Durchführung einer Interkorrelationsanalyse zwischen den Matrizen vor. Im Rahmen der induktiven Analyse (s. folgender Abschnitt) läßt sich dann testen, ob die Korrelationskoeffizienten signifikant von Null abweichen. In diesem Fall ist von einer Abhängigkeit auszugehen.

Das Ergebnis der Korrelationsanalyse ist eine Korrelationsmatrix. Diese Matrix dient wiederum als Basis für anschließende Analysen wie die Faktorenanalyse. Bei der Faktorenanalyse dürfen die errechneten Faktorladungen nicht substantiell hoch sein. In diesem Falle ist von

einer Abhängigkeit der fehlenden Werte und den Faktoren auszugehen. Bei niedrigen Faktorladungen kann hingegen nicht direkt auf eine Abhängigkeit geschlossen werden.

Weiterhin kann eine Clusterung von Objekte bzw. Spalten mit ähnlichen Mustern fehlender Werte durchgeführt werden. Bilden sich große Cluster, in denen die Objekte bzw. Merkmale große Ähnlichkeiten bzgl. ihrer MD-Mustern aufweisen, kann ebenfalls von einer Abhängigkeit innerhalb der Cluster und damit von einem systematischen Ausfallmechanismus ausgegangen werden.

Bei der Dependenzanalyse schließlich wird eine Abhängigkeit einzelner Objekte bzw. Merkmale von den Ausprägungen oder MD-Mustern anderer Objekte bzw. Merkmale modelliert.

4.3 Induktive Analyse

Die Möglichkeiten der deskriptiven und explorativen Analyse im Hinblick auf die Erfassung von Abhängigkeitsbeziehungen der fehlenden Werte sind beschränkt. Die induktive Analyse stellt sich die Aufgabe, Hypothesen aufzustellen und zu testen. In bezug auf fehlende Werte können unterschiedliche Hypothesen aufgestellt werden.

So können beispielsweise Tests auf Häufungen fehlender Daten durchgeführt werden. Dabei lassen sich fehlende Daten als „seltene Ereignisse“ in der Datenmatrix interpretieren (Lösel und Wüstendorfer 1974, S. 344). Damit wird das hypothetische Modell einer Poisson-Verteilung zugrunde gelegt. Bei dem Test wird geprüft, ob die Anzahl der fehlenden Daten in den Objekten bzw. Merkmalen poissonverteilt ist.

Zudem bietet sich im Rahmen der deskriptiven und explorativen Analyse keine Möglichkeit, eine Abhängigkeit der fehlenden Werte von den Ausprägungen eben dieser Werte zu ermitteln. Dies ist nur mit Hilfe der induktiven Analyse möglich, und auch nur dann, wenn die Verteilung der Grundgesamtheit bekannt ist. In diesem Falle können Anpassungstests der restlichen Werte an eine hypothetische Grundgesamtheit durchgeführt werden. Als mögliche Testverfahren werden in der Literatur der χ^2 -Anpassungstest und der Kolmogoroff-Smirnoff-Test erwähnt (Bankhofer 1995, S. 72). Wenn die Anpassungshypothese nicht abgelehnt werden kann, kann weiterhin von einem unsystematischen Ausfallmechanismus ausgegangen werden.

Zusammenfassend bietet die Strukturanalyse die Möglichkeit, systematische Ausfallmechanismen aufzudecken, die bei Nichtberücksichtigung zu erheblichen Verzerrungen bei der Auswertung des unvollständigen Datenmaterials führen können. Nur bei genauer Kenntnis des Ausfallmechanismus ist eine Modellierung des Mechanismus denkbar (Schnell 1986, S.

10). Dies ist jedoch meist schwierig und in praktischen Anwendungen oft kaum durchführbar (Bankhofer 1995, S. 85).

Einfacher zu handhaben sind hingegen die Fälle, in denen von unsystematischen Ausfallmechanismen ausgegangen werden kann. Hierbei stellen die fehlenden Daten eine Zufallsstichprobe der Gesamtstichprobe dar. Der Großteil der in der Literatur zu findenden Ansätze und Verfahren zur Behandlung von Missing Values geht von einem unsystematischen Ausfallmechanismus aus (Bankhofer 1995, S. 85).

5 Behandlung fehlender Daten

Auf Basis der Ergebnisse der Strukturanalyse können verschiedene Verfahren zur Behandlung fehlender Daten zum Einsatz kommen. Schwab (1991, S. 4) unterteilt die möglichen Strategien in die drei Klassen Eliminierungsverfahren, Parameterschätzverfahren und Imputationsverfahren.

Missing-Data-Verfahren	Beschreibung	Bedingung
Eliminierungsverfahren	Ausschluß unvollständiger Objekte bzw. Merkmale	MCAR
Imputationsverfahren	Vervollständigung der Datenmatrix durch Schätzung der fehlenden Werte	MCAR oder MAR, ggf. Modellierung des Ausfallmechanismus
Parameterschätzverfahren	Schätzung von Parametern aus der unvollständigen Datenmatrix	MCAR oder MAR
Multivariate Verfahren	Modifikation multivariater Analyseverfahren	MCAR
Sensitivitätsanalysen	Vergleich verschiedener Ausfallmechanismen oder MD-Verfahren	Je nach Ausfallmechanismus oder MD-Verfahren unterschiedlich

Tabelle 3 – Missing-Data-Verfahren

Bei den **Eliminierungsverfahren** werden Objekte oder Merkmale mit fehlenden Werte aus der Datenmatrix entfernt. Sie werden im Rahmen der Analyse nicht beachtet. Bei den Parameterschätzverfahren werden bestimmte Parameter wie z.B. Mittelwert oder Regressionskoeffizienten geschätzt. Die Schätzung der Parameter erfolgt mit Algorithmen, die alle Objekte bzw. Merkmale, also auch die mit Missing Values, berücksichtigen und jeweils auf spezielle Schätzprobleme abgestimmt sind. Imputationsverfahren füllen die Datenmatrix mit Hil-

fe geeigneter Verfahren auf, so daß mit einer vollständigen Datenmatrix weitergearbeitet werden kann.

Bankhofer (1995, S. 89 ff.) erweitert die Spannweite einsetzbarer Verfahren noch um Multivariate Analyseverfahren und Sensitivitätsbetrachtungen (Beale und Little 1975, S. 130; Frane 1976, S. 409). Die multivariaten Analyseverfahren basieren direkt auf der unvollständigen Datenmatrix. Es handelt sich dabei um in der Regel um geringfügig modifizierte, auf vollständigen Daten basierenden multivariate Verfahren. Im Rahmen von Sensitivitätsanalysen werden schließlich die Spannweite möglicher Analyseergebnisse aufgezeigt, die durch das Fehlen und damit die Unbestimmtheit von Teilen der Datenmatrix entsteht. Eine Übersicht der Verfahren findet sich in Tabelle 3 (Bankhofer 1995, S. 89).

5.1 Eliminierungsverfahren

Die Klasse der Eliminierungsverfahren umfaßt Vorgehensweisen, die unvollständige Merkmale (Merkmalseliminierung) oder Objekte (Objekteliminierung) von der weiteren Untersuchung ausschließen. Die anschließende Auswertung der Daten kann auf Basis einer reduzierten, aber vollständigen Datenmatrix durchgeführt werden.

Eliminierungsverfahren sind nur unter der restriktiven Annahme MCAR uneingeschränkt anwendbar (Little/Rubin 1987, S. 39 f.). Wenn die Eliminierung ohne genaue Kenntnis des Ausfallmechanismus durchgeführt wird, kann es zu gravierenden Verzerrungen kommen, wenn ein systematischer Ausfallmechanismus vorliegt.

Als Beispiel läßt sich wieder das Beispiel der Antwortverweigerung bei höheren Einkommen anführen. Liegt bei höheren Einkommen die Antwortwahrscheinlichkeit für das Einkommen niedriger, kommt es in Falle einer Eliminierung aller Personen mit Missing Values zu einer Unterschätzung des durchschnittlichen Einkommens. Auch multivariate Verfahren wie die Faktorenanalyse oder die Clusteranalyse können in diesen Fällen verfälschte Ergebnisse liefern, wenn überproportional viele Personen mit überdurchschnittlichem Einkommen eliminiert wurden.

Sowohl bei der Objekteliminierung als auch bei der Merkmalseliminierung lassen sich prinzipiell zwei Ansätze unterscheiden. Bei der **complete-case analysis** werden nur die Objekte bzw. Merkmale weiterverwendet, die vollständig vorliegen (Toutenburg 1992, S. 198). Bei der **available-case analysis** werden die Objekte bzw. Merkmale mit Missing Values von der Analyse nicht ausgeschlossen, sondern die jeweils verfügbaren Merkmale bzw. Objekte für die Auswertung verwendet.

Das Verfahren der **complete-case analysis** besitzt vor allem drei Vorteile. Erstens ist es einfach anzuwenden, da lediglich Zeilen bzw. Spalten aus der Datenmatrix gestrichen werden. Zweitens kann anschließend mit einer vollständigen Datenmatrix weitergearbeitet werden. Dies erlaubt die uneingeschränkte Verwendung aller auf vollständigen Daten basierenden multivariaten Standardverfahren. Drittens wird die Vergleichbarkeit insbesondere univariater Statistiken gewährleistet, da die Analyse jeweils auf derselben Stichprobengröße basiert (Little/Rubin 1987, S. 40).

Nachteilig wirkt sich bei der complete-case analysis allerdings das Streichen existierender Werte aus. Der dadurch entstehende Informationsverlust kann in Relation zur Anzahl fehlender Werte vergleichsweise hoch ausfallen. Daher ist die complete-case analysis oft unbefriedigend, da die Untersuchungseinheiten in der Regel nicht grundlos gewählt wurden, und bei dieser Form der Eliminierung insgesamt nicht mehr für eine weitere Auswertung zur Verfügung stehen (Bankhofer 1998, S. 113). Aus diesem Grunde lassen sich complete-case-Analysen oft nur bei einer geringen Anzahl von Missing Values oder bei einer ausreichend großen Datenbasis sinnvoll einsetzen.

Aus diesem Grunde ist der zweite Ansatz oft vorteilhafter. Bei der **available-case analysis** werden die Objekte bzw. Merkmale mit Missing Values von der Analyse nicht ausgeschlossen, sondern jeweils alle verfügbaren Merkmale bzw. Objekte für die Auswertung verwendet.

Beispielhaft für die available-case analysis sind die Berechnung von univariaten Statistiken und Distanzen bei Objekten unter Missing Values, bei der nur die jeweils paarweise vorliegenden Werte verwendet werden (Bankhofer 1995, S. 93 ff.). Ebenso kann bei der Berechnung von Korrelationskoeffizienten verfahren werden. Zur Berechnung dienen hier ebenfalls nur paarweise vorliegende Werte. Ein wesentliches Problem bei diesem Vorgehen ist die schwierige Vergleichbarkeit der auf diese Weise berechneten Statistiken, da jeweils eine unterschiedliche Stichprobengröße bei der Berechnung zu Grunde gelegt wird.

Interessant ist in diesem Zusammenhang, daß die Eliminierung von Objekten bzw. Merkmalen als Optimierungsproblem aufgefaßt werden kann (Dempster 1971, S. 343 in: Bankhofer 1995, S. 103). Als Zielfunktion dient hierbei der entstehende Informationsverlust durch die Eliminierung von Objekten oder Merkmalen. Um möglichst wenige vorhandene Ausprägungen zu eliminieren, wird ein optimaler „Eliminierungsplan“ aufgestellt, in dem die zu eliminierenden Zeilen und Spalten festgelegt werden. Durch Einbindung geeigneter Nebenbedingungen wird erreicht, daß nach der Eliminierung keine Missing Values mehr existieren. Das

entstehende Optimierungsproblem ist durch Branch-and-Bound-Verfahren lösbar (z.B. Domschke/Drexl 1998, S. 125 ff.).

5.2 Imputationsverfahren

Das Problem des teilweise hohen Informationsverlustes bei Eliminierungsverfahren wird durch die Anwendung sogenannter Imputationsverfahren versucht zu beseitigen. Mit Hilfe dieser Verfahrensklasse werden fehlende Werte geschätzt und die Datenmatrix vervollständigt. Da auch hier eine vollständige Datenmatrix erzeugt wird, kann anschließend ohne Einschränkung mit allen Standardverfahren weitergearbeitet werden.

Als Imputationsschätzer kommen im einfachsten Fall je nach Skalenniveau der Mittelwert, der Median oder der Modus in Betracht. In diesen Fällen muß stets die Voraussetzung MCAR vorliegen. Zu weiteren **einfachen Imputationstechniken** gehören die folgenden (Bankhofer/Praxmarer 1998, S. 114):

- Imputation durch den **Verhältnisschätzer**. Bei diesem Verfahren wird für ein möglichst hochkorreliertes Merkmal ermittelt, mit dessen Hilfe der fehlende Wert unter Annahme einer Proportionalitätsbeziehung geschätzt wird. In der Regel haben sowohl das erklärende als auch das fehlende Merkmal kardinales Skalenniveau (Ford 1976, S. 324). Für das Verfahren muß MAR gelten.
- Imputation per **Zufallsauswahl**. Bei diesem Verfahren lassen sich wiederum zwei Varianten unterscheiden: Erstens kann für jeden Missing Value eine zufällige Auswahl aus den existierenden Werten bei anderen Objekten bzw. Merkmalen verwendet werden. Der gezogene Wert wird dann für die Imputation verwendet. Zweitens kann ein Zufallsgenerator direkt einen Wert ermitteln, der für die Imputation verwendet wird. Der Zufallsgenerator wird dabei so eingestellt, daß er die Verteilung des Merkmals korrekt abbildet (Schnell 1986, S. 95). Hier muß die Bedingung MCAR gelten, da systematisch fehlende Daten keinem zufälligen Ausfallmechanismus unterliegen.
- Imputation durch **Expertenratings**. Zur Imputation werden die Werte subjektiv durch Experten geschätzt, oder es werden Modelle verwendet, deren Parameter durch Experten ebenfalls subjektiv geschätzt werden. Wenn ein systematischer Ausfallmechanismus dem Experten bekannt ist, kann das Verfahren auch in diesen Fällen eingesetzt werden. Allerdings muß selbstverständlich auch auf die Probleme hingewiesen werden, die sich durch die subjektiven Festlegungen des Experten ergeben können.

Ein weitere Klasse von Imputationsverfahren bilden die **Imputationsverfahren innerhalb von Klassen**. Das Prinzip dieser Verfahren basiert auf der Hypothese, daß die Objekte innerhalb einer Klasse eine hohe Ähnlichkeit aufweisen. Damit basiert die Festlegung der fehlenden Werte auf den Ähnlichkeitsbeziehungen zwischen den Objekten. Ein wesentlicher Vorteil dieser Vorgehensweise ergibt sich aus der Forderung, daß die Eigenschaft MCAR nur innerhalb der Klassen gelten muß. Man spricht hier von MCARC (**missing completely at random in classes**; Bankhofer 1995, S. 17 f.).

Zur Festlegung der Klassen können zum einen herkömmliche Klassifikationsverfahren verwendet werden, bei denen die jeweils vollständig vorliegenden Daten zur Clusterung verwendet werden. Zum anderen kann die Klassifikation auch auf Basis externer Informationen erfolgen. Anschließend können die oben vorgestellten simplen Imputationsverfahren auf die einzelnen Klassen angewendet werden.

Zusätzlich existieren noch speziell auf Imputationsklassen zugeschnittene Verfahrensvarianten, die in der Literatur als **Cold-Deck** und **Hot-Deck**-Techniken bezeichnet werden (Schnell 1986, S. 108 ff.). Beim Cold-Deck-Verfahren wird ausgehend von den einzelnen Klassen aus einer externen Quelle imputiert. Beim Hot-Deck-Verfahren werden Werte aus der Datenmatrix selbst imputiert. Hot-Deck-Verfahren lassen sich grundsätzlich in sequentielle und simultane Verfahren einteilen. Während bei den sequentiellen Hot-Deck-Verfahren die Imputation der Missing Values eines bestimmten Objektes oder Merkmals jeweils aus verschiedenen Objekten bzw. Merkmalen erfolgen kann, werden bei den simultanen Verfahren aus jeweils einem einzigen vollständig vorliegenden Objekt oder Merkmal sämtliche Werte übernommen.

Sowohl simultane als auch sequentielle Verfahren basieren dabei auf der Ähnlichkeit der Objekte untereinander. Entweder wird aus dem jeweils ähnlichsten Objekt imputiert, oder ein in der entsprechenden Klasse befindliches Imputationsobjekt wird zufällig ausgewählt. Der Imputationsvektor muß dabei jedoch mindestens in allen zu imputierenden Werten vollständig vorliegen, um beim simultanen Hot-Deck-Verfahren überhaupt ausgewählt werden zu können (Ford 1976, S. 326).

Neben den vorgestellten einfachen Imputationstechniken existieren auch komplexere Verfahren: Die Klasse der **multivariaten Imputationstechniken**. Dabei wird die Imputation je nach Skalenniveau anhand verschiedener multivariater Verfahren vorgenommen, wie z.B. mittels Regressions-, Varianz-, Diskriminanz- oder Faktorenanalyse. Für alle multivariaten Imputationstechniken existiert eine Vielzahl von Varianten, die sich in bezug auf das Modell, die

Schätzmethode und die zur Schätzung verwendeten Daten unterscheiden. Eine Übersicht gibt Bankhofer (1995, S. 125 ff.).

Prinzipiell basieren auch die multivariaten Imputationstechniken auf der Annahme MCAR. Wenn jedoch alle Abhängigkeitsbeziehungen zwischen den Merkmalen bzw. Objekten durch das Verfahren erfaßt und ausreichend berücksichtigt werden, ist auch die weniger restriktive Eigenschaft MAR ausreichend (Little/Rubin 1987, S. 45).

Zu den an dieser Stelle ebenfalls zu nennenden Verfahren gehören auch die **Imputationstechniken bei systematischen Ausfallmechanismen**. Bei diesen Verfahren wird der zugrunde liegende systematische Ausfallmechanismus modelliert. Aus diesem Grunde muß genau bekannt sein, wie hoch die Ausfallwahrscheinlichkeit und die Verteilung der Ausprägungen der ausgefallenen Werte a_{ik} in Abhängigkeit der fehlenden und vorhandenen Daten ist. Basierend auf der Modellierung lassen sich dann entweder Erwartungswerte oder mit Hilfe des Modells transformierte Zufallswerte imputieren, die der Verteilung des Ausfallmechanismus entsprechen (Greenless et al. 1982, S. 253). Im Gegensatz zu anderen Imputationstechniken sind bei der expliziten Modellierung des Ausfallmechanismus die Bedingungen MAR, OAR und MCAR nicht erforderlich, da der Ausfallmechanismus durch die Modellierung bereits berücksichtigt wird.

Einer der wesentlichen Vorteile von Imputationsverfahren ist, daß anschließend mit einer vollständigen Datenmatrix weitergearbeitet werden kann. Außerdem führt die Anwendung der Verfahren zu keinem Informationsverlust.

Die Verwendung bestimmter Schätzwerte kann jedoch auch bei Vorliegen der notwendigen Eigenschaften (z.B. MCAR) zu Verzerrungen führen. Beispielsweise führt die mehrmalige Imputation von Erwartungswerten wie z.B. dem Mittelwert zu einer Unterschätzung der Varianz. Verfahren, die sich statt am Erwartungswert an der Verteilung der zu imputierenden Werte orientieren, liefern diesbezüglich bessere Ergebnisse.

5.3 Parameterschätzverfahren

Parameterschätzverfahren stellen eine weitere Verfahrensklasse zur Behandlung fehlender Daten dar. In diese Verfahrensklasse fallen Methoden, bei denen direkt aus der unvollständigen Datenmatrix bestimmte Parameter geschätzt werden. Geschätzt werden meistens Mittel-

werte, Varianzen und Kovarianzen. Die Ergebnisse können anschließend in multivariaten Verfahren wie der Faktoren- oder Diskriminanzanalyse verwendet werden (Bankhofer/Praxmarer 1998, S. 115).

Parameterschätzverfahren grenzen sich zu Eliminierungstechniken insofern ab, als daß kein Informationsverlust bei der Schätzung der Parameter auftritt. Der Unterschied dieser Verfahren zu den Imputationstechniken liegt in der Tatsache, daß zur Schätzung der Parameter Korrekturen durchgeführt werden können. Diese Korrektur erlaubt die Anwendung der Verfahren auch bei Vorliegen lediglich der Eigenschaft MAR, wenn eine quantitative Datenmatrix mit multinormalverteilten Daten vorausgesetzt gesetzt werden kann. Verteilungsfreie Verfahren benötigen hingegen in der Regel die Eigenschaft MCAR, wie beispielsweise der EM-Algorithmus von Orchard und Woodbury (Schwab 1991, S. 152). Weitere Parameterschätzverfahren können Bankhofer (1995, S. 156 ff.) entnommen werden.

In den weniger restriktiven Voraussetzungen liegt der Vorteil der Parameterschätzverfahren. Nachteilig ist, daß zur weiteren Auswertung nur Verfahren in Frage kommen, die auf den ermittelten Parametern basieren.

5.4 Multivariate Analyseverfahren

Bei den multivariaten Analyseverfahren zur Behandlung von Missing Values handelt es sich in der Regel um geringfügige Modifikationen der entsprechend auf vollständigen Daten basierenden multivariaten Verfahren. Beispielsweise führen Schader und Gaul (1991) das **Missing Value Linkage**-Verfahren ein. Im Vergleich zum entsprechenden Verfahren mit vollständigen Daten ergeben sich zwei Unterschiede. Zum einen kann bei der Distanzberechnung innerhalb und zwischen den Klassen lediglich auf die vorhandenen Werte zurückgegriffen werden. Zum anderen muß beim Abbruchkriterium berücksichtigt werden, daß die zur Abbruchbestimmung herangezogenen Parameter ebenfalls fehlende Werte darstellen können.

Auch für die Faktorenanalyse, Regressionsanalyse und Multidimensionale Skalierung existieren in der Literatur Modifikationen, um dem Problem der Missing Values gerecht werden zu können. Eine Übersicht mit Literaturhinweisen liefert Bankhofer (1995, S. 168-181). Die Verfahren benötigen die MCAR-Eigenschaft.

Der Vorteil multivariater Analyseverfahren liegt in der vollständig integrierten Behandlung der unvollständigen Datenmatrix, bei dem auch das Analyseziel direkt erreicht wird. Die erzielten Ergebnisse können bei einer verhältnismäßig hohen Anzahl von fehlenden Werten jedoch verzerrt sein.

5.5 Sensitivitätsanalysen

Die bis hier angesprochenen Verfahren führen stets zu jeweils einem Analyseergebnis. Fehlende Werte in Datenmatrizen führen jedoch auch immer zu einer gesteigerten Varianz bzw. Unsicherheit. Mit Hilfe von Sensitivitätsanalysen versucht man, mehrere mögliche Analyseergebnisse in Abhängigkeit unterschiedlicher Ausfallmechanismen und MD-Verfahren für die fehlenden Daten aufzuzeigen.

Ein viel zitierter Ansatz aus der Literatur ist das Konzept der **multiplen Imputation** (Rubin 1987). Die Idee ist hierbei, daß jedes MD-Verfahren **immer** ein Modell des Ausfallmechanismus heranzieht, wie in den meisten Fällen der unsystematische Ausfallmechanismus. Die multiple Imputation verwendet hingegen verschiedene Ausfallmechanismen und gelangt so zu unterschiedlichen Ergebnissen. Damit kann die Sensitivität der Analyseergebnisse in Abhängigkeit des unterstellten Ausfallmechanismus aufgezeigt werden. Statt eines einzelnen Resultates wird also vielmehr die Spannweite möglicher Ergebnisse aufgezeigt.

Selbstverständlich können neben den nebeneinander untersuchten Ausfallmechanismen auch unterschiedliche MD-Verfahren angewendet werden, deren Ergebnisse dann im Rahmen einer Sensitivitätsanalyse verglichen werden können.

6 Zusammenfassung

Fehlende Daten stellen ein häufig anzutreffendes Problem in der angewandten Statistik dar. Zu untersuchen ist dabei jeweils, welche Auswirkungen sich durch die Missing Values auf Analyseergebnisse ergeben können. Hierzu sind zunächst Kenntnisse über den Ausfallmechanismus der Daten zu gewinnen, was im Rahmen einer Strukturanalyse geschehen kann. Anschließend können unterschiedliche Verfahren eingesetzt werden, um dem Problem der Missing Values gerecht zu werden. Neben Eliminierungsverfahren eignet sich die Klasse der Imputationsverfahren oft am besten, da hier kein Informationsverlust auftritt.

Zur Frage, welches MD-Verfahren am besten geeignet ist, existieren eine Anzahl von Studien in der Literatur. Ein durchgängig am besten geeignetes MD-Verfahren existiert jedoch nicht. Die Auswahl einer geeigneten Strategie bzw. eines geeigneten MD-Verfahrens hat vielmehr unter Abwägung der an die Daten gestellten Anforderungen, der Zielsetzung der Analyse sowie der Vor- und Nachteile der einzelnen Strategien zu erfolgen (Bankhofer 1995, S. 190).

Quellenverzeichnis

- Backhaus; Erichson; Plinke; Weiber (1994): *Multivariate Analysemethoden*, 7. Auflage, Berlin, Heidelberg, New York, Tokio
- Bankhofer, Udo (1995): *Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse*, Bergisch-Gladbach, Eul
- Bankhofer, Udo; Praxmarer, Sandra (1998): Zur Behandlung fehlender Daten in der Marktforschungspraxis, *Marketing ZFP*, Heft 2, 2. Quartal 1998
- Beale, E.M.L.; Little R.J.A. (1975): Missing Values in Multivariate Analysis, *Journal of the Royal Statistical Society*, 37, Series B, S. 129-145
- Brown, C.H. (1983): Asymptotic Comparison of Missing Data Procedures for Estimating Factor Loadings, *Psychometrika*, 48, S. 269-291
- Cohen, Jacob; Cohen, Patricia (1975): *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Hillsdale, NJ, Erlbaum
- Dempster, A.P. (1971): An Overview of Multivariate Data Analysis, *Journal of Multivariate Analysis*, 1, S. 316-346
- Domschke, Drexl (1998): *Einführung in Operations Research*, 4. Auflage, Berlin, Springer.
- Frane, J.W. (1976): Missing Data and BMDP: Some Pragmatic Approaches, *ASA Proceedings of the Statistical Computing Section*, S. 27-33
- Greenless, W.S.; Reece, J.S.; Zieschang, K.D. (1982): Imputation of Missing Values when the Probability of Response Depends on the Variables Being Imputed, *Journal of the American Statistical Association*, Vol. 77, 1982, S. 251-261
- Hartung, Joachim; Elpelt, Bärbel (1995): *Multivariate Statistik: Lehr- und Handbuch der angewandten Statistik*, 5. Auflage, München, Wien, Oldenbourg
- Little, Roderick J.A.; Rubin, Donald B. (1987): *Statistical Analysis with Missing Data*, New York, Wiley
- Lösel, F; Wüstendorfer, W. (1974): Zum Problem unvollständiger Datenmatrizen in der empirischen Sozialforschung, *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 26, S. 342-357.
- Rubin, Donald B. (1976): Inference and missing data, in: *Biometrika*, Vol. 63, S. 581-592
- Rubin, Donald B. (1987): *Multiple Imputation for Nonresponse in Surveys*, New York, Wiley
- Rummel, R.J. (1970): *Applied Factor Analysis*, Evanston, Northwestern University Press
- Schnell, Rainer (1986): *Missing-Data-Probleme in der empirischen Sozialforschung*, Bochum

Schnell, Rainer; Hill, Paul B.; Esser, Elker (1988): Methoden der empirischen Sozialforschung, München

Schwab, Georg (1991): Fehlende Werte in der angewandten Statistik, Wiesbaden, DUV

Toutenburg, Helge (1992): Lineare Modelle, Heidelberg, Physica