M. Wirtz

Über das Problem fehlender Werte: Wie der Einfluss fehlender Informationen auf Analyseergebnisse entdeckt und reduziert werden kann

On the Problem of Missing Data: How to Identify and Reduce the Impact of Missing Data on Findings of Data Analysis

Zusammenfassung

Fehlende Werte stellen ein häufig unterschätztes Problem bei der Analyse empirischer Datensätze dar. Neben der effektiven Verringerung der Stichprobenumfänge muss damit gerechnet werden, dass durch fehlende Angaben statistische Ergebnisse verfälscht werden. Die beiden zentralen Aussagen des Beitrags bestehen darin, dass zum einen die Analyse der Ursachen fehlender Angaben in empirischen Datensätzen fester Bestandteil jeder Datenanalyse sein sollte und zum anderen der Umgang mit fehlenden Werten auf angemessenen Annahmen beruhen sollte, um fehlerhafte Ergebnisse und Probleme bei der Interpretation empirischer Befunde zu vermeiden.

Schlüsselwörter
Missing-Data-Diagnose · Missing-Data-Prozesse · fallweiser Ausschluss · paarweiser Ausschluss · Imputationsverfahren

Abstract

The impact of missing data on the analysis of empirical data is a frequently unrecognized problem. Missing data may not only result in a decrease in the actual sample size but potentially biasing effects on statistical findings have to be considered as well. Two important points are made in this article: Firstly, it is shown why the identification of potential causes of missing data should be an inherent part of any data analysis; secondly, the handling of missing data should be based on appropriate assumptions in order to avoid biased results and problems concerning the interpretation of empirical findings.

Key words
Missing data diagnosis · missing data processes · listwise deletion · pairwise deletion · imputation procedure

Fehlende Werte stellen stets ein Problem bei der Auswertung empirischer Daten dar: Wie soll damit umgegangen werden, wenn beispielsweise Personen Angaben verweigern oder ab einem bestimmten Zeitpunkt nicht mehr an der Datenerhebung teilnehmen? Die Bedeutung dieses Problems wird jedoch häufig unterschätzt und Standards für ein optimales Vorgehen sind wenig verbreitet. Der Anwender ist zumeist auf sich selbst gestellt, so dass nicht selten die Voreinstellungen des Analyseprogramms unhinterfragt die Vorgehensweise determinieren. Dies ist verwunderlich, da die Auswahl der Verfahren zum Umgang mit feh-

Koordinatoren der Reihe „Methoden in der Rehabilitationsforschung“:
Prof. Dr. Dr. Hermann Faller, Würzburg; Prof. Dr. Thomas Kohlmann, Greifswald;
Dr. Christian Zwingmann, Frankfurt/Main
Interessenten, die einen Beitrag zur Reihe beisteuern möchten, werden gebeten, vorab Kontakt aufzunehmen, E-mail: christian.zwingmann@vdr.de

Institutsangabe
Methodenzentrum des Rehabilitationswissenschaftlichen Forschungsverbunds Freiburg/Bad Säckingen, Freiburg

Korrespondenzadresse
Dr. phil. Dipl.-Psych. Markus Wirtz · Methodenzentrum des RFV Freiburg/Bad Säckingen · Abt. für Rehabilitationspsychologie · Institut für Psychologie · Universität Freiburg · Engelbergerstraße 41 · 79085 Freiburg · E-mail: wirtz@psychologie.uni-freiburg.de

Bibliografie
Rehabilitation 2004; 43: 109 – 115 © Georg Thieme Verlag KG Stuttgart · New York
DOI 10.1055/s-2003-814839
ISSN 0034-3536

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

lenden Werten erheblichen Einfluss auf das Ergebnis der statistischen Analyse haben kann. Nur wenn die Voraussetzungen der Verfahren erfüllt sind, können Fehler bei der Datenauswertung und Ergebnisinterpretation vermieden werden. Im Folgenden wird skizziert, welche Probleme durch fehlende Werte entstehen können und unter welchen Bedingungen welche Verfahren geeignet sind. Hierbei werden unter „fehlenden Werten" nur fehlende Daten verstanden, von denen angenommen werden kann, dass diese empirisch existieren, obwohl sie nicht erhoben wurden.[1]

Was ist gegen die Praxis einzuwenden, das Problem fehlender Werte weitestgehend zu ignorieren? Auf Basis empirischer Daten soll in der Forschung beispielsweise über die Gültigkeit von Hypothesen entschieden werden: Wenn Daten nicht vorliegen, können diese folglich nicht für diesen Entscheidungsprozess genutzt werden. Nun, das Problem besteht darin, dass Daten nicht notwendigerweise zufällig fehlen müssen, sondern dass der Datenausfall systematisch bedingt sein kann: „Systematisch" bedeutet hier, dass es einen Grund für das Nichtvorliegen der Information gibt, der bei der Ergebnisinterpretation nicht vernachlässigt werden darf.

Ein Beispiel soll diese Problematik verdeutlichen: In einer Befragung soll u. a. erfasst werden, ob die Befragten in der Vergangenheit bereits einmal selbst illegale Drogen konsumiert haben. Ein Teil der Personen verweigert die Angaben zu der betreffenden Frage. Wenn nun tendenziell eher diejenigen Personen, die illegale Drogen konsumiert haben, die Auskunft verweigern, wird der Anteil von Konsumenten durch die Stichprobendaten systematisch unterschätzt.

Hier gilt, dass die Befragtenstichprobe in Bezug auf die Informationen zum Drogenkonsum allein deswegen nicht repräsentativ für die Gesamtpopulation sein kann, weil das zu erhebende Merkmal „Drogenkonsum" selbst mit der Tatsache zusammenhängt, ob die betreffende Frage beantwortet wird. Nur für die Population der „Auskunftswilligen" – hier also eher diejenigen, für die das Merkmal nicht zutrifft – kann das Befragungsergebnis unverfälscht oder repräsentativ sein.[2] Es wäre fahrlässig, aus diesen Ergebnissen auf die Merkmalsverteilung in der Stichprobe aller Befragten zu verallgemeinern.

Komplexer stellt sich die Problematik fehlender Werte in multivariaten Analysen dar, bei denen mehrere Variablen simultan analysiert werden (z. B. multiple Regression oder Faktorenanalyse). Um einen Datensatz mit fehlenden Angaben angemessen auswerten zu können, müssen zwei zentrale Fragen im Rahmen einer Missing-Data-Analyse geklärt werden [1, 2]:

1. Wodurch sind fehlende Angaben bedingt? Treten fehlende Angaben zufällig oder systematisch auf? (Missing-Data-Diagnose)

Tab. **1** Ein fiktiver Datensatz für 11 Personen, in dem 3 Personen die Angaben zur „Zufriedenheit mit dem Beruf" (X1) und 4 Personen die Angabe zur „Zufriedenheit mit dem Einkommen" (X2) verweigert haben. Für die „allgemeine Lebenszufriedenheit" (Y) liegen die Angaben aller Personen vor. Die Indikatorvariablen Ind(X1) und Ind(X2) enthalten die Information, ob die entsprechende Angabe in X1 bzw. X2 vorliegt (0 = fehlend; 1 = vorhanden)

Zufriedenheit mit dem Beruf X1	*Zufriedenheit mit dem Einkommen X2*	*allgemeine Lebenszufriedenheit Y*	*Ind(X1)*	*Ind(X2)*
.	2	1	0	1
.	1	2	0	1
.	3	3	0	1
6	5	4	1	1
4	4	5	1	1
4	5	6	1	1
6	4	7	1	1
7	.	8	1	0
8	.	9	1	0
9	.	10	1	0
9	.	11	1	0

2. An welche Annahmen sind die zur Verfügung stehenden Verfahren zum Umgang mit Missing Data gekoppelt?

Um das Vorgehen bei der Missing-Data-Analyse zu beschreiben, soll das überschaubare Datenbeispiel in Tab. **1** verwendet werden. Es seien von 11 Personen die Merkmale „Zufriedenheit mit dem Beruf" (X1), „Zufriedenheit mit dem Einkommen" (X2) und „allgemeine Lebenszufriedenheit" (Y) erhoben worden. Drei Personen haben die Angaben zur „Zufriedenheit mit dem Beruf", vier Personen die Angaben zur „Zufriedenheit mit dem Einkommen" verweigert. Durch dieses Beispiel werden die Gefahren nachvollziehbar, die durch einen falschen Umgang mit fehlenden Werten entstehen können. Bei allen Analysen muss beachtet werden, dass die teststatistischen Ergebnisse und die abgeleiteten Aussagen für diesen Datensatz nur bei genügend großen Stichproben eine ausreichende Teststärke besitzen würden.

Missing-Data-Diagnose: Welche Zusammenhänge bestehen zwischen fehlenden Werten und den übrigen Informationen im Datensatz?

Zunächst sollte untersucht werden, ob einzelne Personen oder Variablen hohe Anteile fehlender Werte aufweisen: Hierdurch wird zum einen deutlich, welche Personen oder Personengruppen Antworten typischerweise verweigern oder Probleme mit der Beantwortung haben. Zum anderen zeigt sich, welche Fragen/Items u. U. ungeeignet sind. Oft führt die Elimination weniger Personen und/oder einzelner Fragen zu einer wesentlich vollständigeren Datenmatrix. Entschließt man sich zu einer solchen Elimination, muss diese natürlich dokumentiert werden und die Implikationen müssen bei der Interpretation berücksichtigt werden. Als Richtlinie gilt, dass durch Personen oder Variablen mit mehr als 30 % fehlenden Werten mehr Unsicherheiten und Fehler

[1] Angenommen, eine Person lebt momentan nicht mit einem/r Partner/in zusammen, so würden fehlende Angaben zu der Zufriedenheit mit der Partnerschaft nicht als „fehlende Werte" in diesem Sinne gelten, da diese logisch nicht möglich sind.

[2] Es muss natürlich zusätzlich u. a. vorausgesetzt werden, dass die Angaben in einer solchen Befragung wahrheitsgemäß erfolgen.

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

Tab. **2** Mittelwerte und Ergebnisse des t-Tests mit den Werten von Y als abhängige Variable und den Indikatorvariablen als unabhängige Variablen (Daten s. Tab. **1**)

unabhängige Variable	*abhängige Variable* $\bar{Y}$, *wenn Ind 0*	$\bar{Y}$, *wenn Ind 1*	$t_{df=8}$ *(Signifikanz)*[1]
Ind(X1)	2,0	7,5	– 3,67 ($p = 0{,}005 < \alpha = 0{,}05$)
Ind(X2)	9,0	3,5	5,20 ($p = 0{,}001 < \alpha = 0{,}05$)

[1] Das Ergebnis des Signifikanztests wird hier angegeben, obwohl bei den vorliegenden Stichprobengrößen das Ergebnis des Signifikanztests mit Vorbehalt interpretiert werden muss. Die Ergebnisse des Tests werden dargestellt, um das allgemein empfohlene Vorgehen in großen Datensätzen zu erläutern

für die statistischen Analysen erkauft werden, als dass substanzielle Informationen gewonnen werden.[3]

Im Rahmen einer Missing-Data-Diagnose wird weiterhin analysiert, ob fehlende Werte systematisch im Zusammenhang mit den Werten anderer Variablen auftreten oder Muster fehlender Werte gehäuft vorliegen. Eine wichtige Auswertung soll exemplarisch beschrieben werden. In Tab. **1** wurden zwei Indikatorvariablen – Ind(X1) und Ind(X2) – gebildet, in denen kodiert ist, ob die Variablen X1 bzw. X2 fehlende Werte enthalten. Es wird für die beiden Stufen (0, 1) der beiden Indikatorvariablen überprüft, ob die *vorhandenen* Angaben in allen anderen Variablen systematisch differieren. Tab. **2** zeigt die Ergebnisse, wenn die Systematik der Unterschiede in der Variable Y in Abhängigkeit von den beiden Indikatorvariablen mittels eines t-Tests überprüft wird. Die Personen mit fehlenden Werten in der „Zufriedenheit mit dem Beruf", also Ind(X1) = 0, haben mit 2,0 einen signifikant kleineren Wert in der „allgemeinen Lebenszufriedenheit" als die übrigen Personen (7,5). Hingegen haben die Personen mit fehlenden Werten in der Variable „Zufriedenheit mit dem Einkommen" eine signifikant höhere „allgemeine Lebenszufriedenheit" als die übrigen Personen. Diese Ergebnisse deuten darauf hin, dass die fehlenden Angaben nicht zufällig auftreten, da deren „Auftreten" durch die „allgemeine Lebenszufriedenheit" vorhergesagt werden kann.

In einer Missing-Data-Diagnose wird des Weiteren untersucht, ob Muster fehlender Werte gehäuft auftreten [1–3]. Muster bedeutet hier, dass in einer Gruppe von Variablen fehlende Werte gemeinsam auftreten: Immer, wenn beispielsweise Variable A nicht beantwortet wird, fehlen auch eher die Werte in den Variablen D und F.

[3] In Tab. **1** müsste nach diesem Kriterium Variable X2 wegen 36 % fehlender Werte ausgeschlossen werden. Da das sehr vereinfachende Beispiel zur Demonstration typischer Probleme bei der Missing-Data-Analyse dient, wird die Variable dennoch weiter analysiert.

[4] In der Literatur wird MAR manchmal fälschlicherweise so definiert, dass die übrigen Variablen das Fehlen von Angaben kausal bedingen müssen. Diese Annahme ist nicht notwendig: Es wird nur gefordert, dass ein hinreichender korrelativer Zusammenhang besteht [5, 6].

Die Missing-Data-Diagnose liefert einen sehr guten Überblick über Datenstrukturen, die mit den fehlenden Werten in Verbindung stehen: Der Untersucher wird dafür sensibilisiert, welche Gefahren bei der Interpretation der Ergebnisse durch nicht vorhandene Angaben entstehen können. Eine Missing-Data-Diagnose sollte stets durchgeführt werden, wenn für einzelne Variablen mehr als 5 % der Werte fehlen, da dann ein systematischer Ausfall von Daten erhebliche Verzerrungen der Ergebnisse bewirken kann. Zudem muss beachtet werden, dass bei einer bedeutsamen Systematik der fehlenden Angaben Standardmethoden zum Umgang mit fehlenden Werten nicht mehr angewendet werden sollten. Dieser Punkt wird im Zusammenhang mit Tab. **3** verdeutlicht werden.

Ursachen für fehlende Werte: Missing-Data-Prozesse

Unter Missing-Data-Prozessen versteht man die Ursachen, die das Auftreten von fehlenden Werten bedingen. Es werden drei Arten von Prozessen unterschieden [4], die für den Umgang mit Missing Data entscheidend sind. Man sagt, dass Daten „vollständig zufällig fehlen" (MCAR = *M*issing *C*ompletely *A*t *R*andom), wenn das Auftreten fehlender Angaben (a) weder von den Ausprägungen anderer Variablen noch (b) von der Ausprägung der (nicht angegebenen) Werte der Variablen selbst abhängt. Beantworten Personen einzelne Fragen nicht, weil sie insgesamt schwach motiviert sind, an einer Untersuchung teilzunehmen, so wäre Bedingung a verletzt. Geben – wie im einleitenden Beispiel – Personen tendenziell eher keine Auskunft zum Drogenkonsum, wenn sie Drogen konsumieren, so würde Bedingung b nicht gelten.

Dass Werte vollständig zufällig fehlen, ist in rehabilitationswissenschaftlichen Studien wohl in den seltensten Fällen gegeben, weil es zumeist einen Grund dafür gibt, weshalb Befragte eine Angabe verweigern oder durch erhebungsbedingte Probleme bestimmte Personen zu bestimmten Zeitpunkten nicht befragt wurden. Typische Beispiele für „vollständig zufälliges Fehlen" sind Fehlwerte durch Dateneingabefehler oder Irrläufer beim postalischen Rücklauf.

Der zweite Missing-Data-Prozess wird als „zufälliges Fehlen" (MAR = *M*issing *A*t *R*andom) bezeichnet. „Zufälliges Fehlen" bedeutet, dass das Fehlen von Werten vollständig durch die übrigen Informationen im Datensatz vorhergesagt werden kann.[4] Der Begriff „zufälliges Fehlen" ist hier leider etwas irreführend: „Bedingt zufälliges Fehlen" würde den Sachverhalt besser beschreiben, da gefordert wird, dass nach Berücksichtigung der Informationen im Datensatz keine weiteren Informationen zur Vorhersage des Fehlens von Werten erforderlich sind. Ist beispielsweise die Motivation zur Teilnahme an einer Studie der wesentliche Prädiktor dafür, dass Angaben nicht vollständig vorliegen, so fehlen die Daten genau dann zufällig (MAR), wenn die Motivation zur Teilnahme als Messdatum für die Befragten erhoben wurde.

Wenn jedoch die im Datensatz fehlenden Angaben bekannt sein müssten, damit das Fehlen von Werten hinreichend vorhergesagt werden kann, fehlen die Daten „nicht zufällig" (NRM = *N*on *R*andom *M*issing). Im einleitenden Beispiel der Befragung zum

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

Tab. **3** Die Korrelationsmatrizen und die Mittelwerte und Standardabweichungen für die Daten in Tab. **1** bei fallweisem oder paarweisem Ausschluss und bei Mittelwertersetzung

		fallweise (n = 3)			*paarweise*			*MW-Ersetzung (n = 11)*		
		X1	*X2*	*Y*	*X1*	*X2*	*Y*	*X1*	*X2*	*Y*
r	**X1**	1	0,00	0,00	1[1]	0,00[4]	0,86[1]	1	– 0,36	0,53
r	**X2**		1	– 0,45		1[2]	0,77[2]		1	0,39
r	**Y**			1			1[3]			1
MW		5,00	4,50	5,00	6,63[1]	3,43[2]	6,00[3]	6,63	3,43	6,00
SD		1,15	0,58	1,29	2,00	1,51	3,32	1,67	1,17	3,32

[1] n = 8, [2] n = 7, [3] n = 11, [4] n = 4

Drogenkonsum würde „nicht zufälliges Fehlen“ gelten, da die Werte der zu messenden Eigenschaft deren Angabe mitdeterminieren. Würde man in der Untersuchung jedoch andere Variablen erfassen, die mit dem Konsum illegaler Drogen und somit mit dem Fehlen der Information in gutem Zusammenhang stehen, so würden die Werte lediglich „zufällig fehlen“ (MAR).

Wie kann festgestellt werden, welcher dieser drei Prozesse vorliegt? Dass die Bedingung „vollständig zufällig fehlend“ (MCAR) verletzt ist, ergibt sich als Ergebnis der Missing-Data-Diagnose, wenn bedeutsame Zusammenhänge der fehlenden Werte mit den übrigen Informationen im Datensatz bestehen. Neben dieser eher subjektiven Grundlage existiert ein Testverfahren (MCAR-Test nach Little [2]), das überprüft, ob die Bedingung „vollständig zufällig fehlend“ verletzt ist. In unserem Datenbeispiel wird der Test mit $Chi^2(4) = 9{,}55$ ($p = 0{,}049 < Alpha = 0{,}25$)[5] signifikant: Die Annahme, dass die Daten „vollständig zufällig fehlen“ muss verworfen werden, dem Ausfall der Daten liegt also eine Systematik zugrunde.

Ob die Daten jedoch „zufällig“ (MAR) oder „nicht zufällig“ (NRM) fehlen, kann empirisch nicht entschieden werden. Um dies entscheiden zu können, müsste die Ausprägung der fehlenden Daten bekannt sein. Denn nur dann könnte geklärt werden, ob die nicht vorliegenden Informationen notwendig sind, um den Ausfall der Daten erklären zu können. Die Beantwortung dieser Frage basiert somit vorwiegend auf inhaltlichem Wissen über den Forschungsgegenstand. In der Regel ist keine eindeutige Entscheidung möglich.

Welches Verfahren zum Umgang mit fehlenden Werten sollte gewählt werden?

Zunächst sollen die Eigenschaften der Methoden „fallweiser“ und „paarweiser Ausschluss“ sowie „Ersetzung durch den Variablenmittelwert“, die in der Praxis am häufigsten angewendet werden, diskutiert werden. Beim „fallweisen Ausschluss“ („listwise deletion“, Complete-information-Ansatz) fällt eine Person aus der Berechnung *aller* Statistiken heraus, wenn in einer der zu analysierenden Variablen für diese Person ein Wert fehlt. Für die Beispieldaten in Tab. **1** würde dies bedeuten, dass nur die Daten der Fälle 4 – 7 in die Auswertung eingehen.

Bei „paarweisem Ausschluss“ („pairwise deletion“, All-available-information-Ansatz) werden für die Berechnung jeder Statistik alle verfügbaren Dateninformationen verwertet. Eine Person wird beispielsweise nur bei Berechnung der Korrelationen nicht berücksichtigt, wenn einer der Werte der direkt betroffenen Variablen nicht vorliegt. Die Daten von den Personen 1 – 7 gehen mit in die Berechnung der Korrelation von X2 und Y ein. Da die Werte in X1 für die Personen 1 – 3 aber nicht vorliegen, können diese drei Personen nicht berücksichtigt werden, wenn die Korrelationen von X1 und X2 bzw. X1 und Y bestimmt werden.

In Tab. **3** sind links für den fallweisen Ausschluss die Korrelationen (r), Mittelwerte (MW) und die Standardabweichungen (SD) angegeben. In der Mitte finden sich die Werte für den paarweisen Ausschluss. Die beiden Verfahren führen zu deutlich unterschiedlichen Ergebnissen. Da durch die Systematik des Datenausfalls beim fallweisen Ausschluss nur Personen mit mittlerer Ausprägung in Y berücksichtigt werden, ergeben sich hier insgesamt niedrigere Zusammenhänge. Wenn Werte nicht „vollständig zufällig fehlen“, ist dieses Ergebnis typisch: Da eher Personen mit extremen Ausprägungen auf bestimmten Variablen nicht antworten, wird die Stichprobe homogener. Je homogener eine Stichprobe ist, desto niedriger sind die Korrelationen im Allgemeinen ausgeprägt [7]. Diese Homogenisierung der Stichprobe zeigt sich in den geringeren Standardabweichungen bei fallweisem Ausschluss.

Grundsätzlich lässt sich für den „fallweisen Ausschluss“ festhalten: Durch den Ausschluss von Personen mit fehlenden Werten aus der Analysestichprobe können die Eigenschaften der Stichprobe verändert werden. Nur wenn die fehlenden Werte „vollständig zufällig fehlen“ würden, wäre die durch den fallweisen Ausschluss neu definierte, reduzierte Stichprobe eine repräsentative Auswahl der ursprünglichen Stichprobe. Fehlen die Daten nicht „vollständig zufällig“, so werden die Statistiken verzerrt. Ein weiteres Problem des „fallweisen Ausschlusses“ entsteht durch den Stichprobenschwund. Je größer die Anzahl der Variablen, die fehlende Werte aufweisen, desto gravierender ist das Problem der Stichprobenreduktion. Häufig wird der effektive Stichprobenumfang mehr als halbiert [8]. Hierdurch wird die Zu-

[5] Da hier im Allgemeinen die Nullhypothese präferiert wird und das Beta-Fehlerrisiko minimiert werden soll, sollte die Alpha-Fehlerwahrscheinlichkeit höher als üblich gewählt werden.

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

verlässigkeit und Teststärke der statistischen Analysen erheblich reduziert [9].

In Bezug auf das Problem des Stichprobenumfangs ist der „paarweise Fallausschluss“ die günstigere Alternative, da keine erhobenen Daten unberücksichtigt bleiben. Weil alle Personen mit in die Auswertung eingehen, ist auch das Problem der Homogenisierung der Stichprobe zumindest entschärft: Beispielsweise wird die Streuung des Merkmals Y mit 3,32 (s. Tab. **3**) durch den Umgang mit den fehlenden Werten nicht beeinflusst. Bei der Berechnung der Korrelation von X1 und Y wird immerhin das obere Spektrum der Merkmalsausprägungen von Y vollständig abgedeckt, und bei der Berechnung der Korrelation von X2 und Y wird das untere Spektrum vollständig berücksichtigt.

Neben diesen beiden Vorteilen des „paarweisen Ausschlusses“ existiert aber ein wichtiger Nachteil. Die Stichprobenstatistiken in Tab. **3** (Mitte) stammen effektiv aus verschiedenen Substichproben. Die Korrelation von X1 und Y wird an Personen mit tendenziell geringerer Lebenszufriedenheit bestimmt, die Korrelation von X1 und X2 an Personen mit mittlerer Lebenszufriedenheit und die Korrelation von X2 und Y an Personen mit hoher Lebenszufriedenheit. Diese Problematik führt im Beispiel sogar dazu, dass die resultierende Korrelationsmatrix in Tab. **3** in einem vollständigen Datensatz empirisch unmöglich ist: Wenn die „allgemeine Lebenszufriedenheit“ mit der „Zufriedenheit mit dem Beruf“ zu 0,86 und mit der „Zufriedenheit mit dem Einkommen“ zu 0,77 korreliert, dann müssen auch die „Zufriedenheit mit dem Beruf“ und die „Zufriedenheit mit dem Einkommen“ positiv korreliert sein. Es lässt sich nachweisen, dass die Korrelation größer als 0,34 sein muss, eine kleinere Korrelation wäre in einem Datensatz ohne fehlende Werte nicht möglich [10].[6]

Aber auch in weniger extremen Anwendungsfällen, in denen eine empirisch mögliche Informationsmatrix resultiert, entsteht ein großes Problem, wenn die Substichproben, die für jede zu berechnende Statistik herangezogen werden, nicht vergleichbar sind: Das resultierende *Gesamt*ergebnis gilt dann für keine der „Substichproben“, und es lässt sich keine Gesamtstichprobe definieren, für die die „gefundene“ Struktur Gültigkeit besitzt. Es ergeben sich Puzzlestücke von Informationen, die nicht mehr zu einem sinnvollen Ganzen zusammengefügt werden können, da jedes Puzzlestück zu einem anderen Gesamtbild gehört. Dies ist genau dann der Fall, wenn die Daten nicht „vollständig zufällig fehlen“ (MCAR) [1].

Bei der „Ersetzung durch den Variablenmittelwert“ werden in Tab. **1** die fehlenden Werte in Variable X1 durch den Mittelwert der Fälle 4 – 11 (6,63) und in Variable X2 durch den Mittelwert der Fälle 1 – 7 (3,43) ersetzt. Dieser Ansatz erscheint auf den ersten Blick attraktiv, weil die Datenmatrix anschließend keine Lücken mehr aufweist. Trotzdem kann diese Ersetzung nicht empfohlen werden.[7] Damit der Mittelwert der übrigen Personen eine unverzerrte Schätzung des Mittelwerts der Gruppe der Personen mit fehlenden Angaben liefert, müssten die fehlenden Informationen eine repräsentative Auswahl derselben Verteilung darstellen, die den vorhandenen Informationen im Datensatz zugrunde liegt („vollständig zufälliges Fehlen“, MCAR). Aber auch, wenn diese Bedingung erfüllt ist, wird durch die eingesetzten Mittelwerte die Datenverteilung künstlich homogenisiert. In Tab. **3** zeigt sich dies darin, dass die Streuung der Variablen X1 und X2 für die Mittelwertersetzung niedriger ist als bei paarweisem Ausschluss. Die negative Korrelation zwischen X1 und X2 ist in diesem Beispiel ein weiteres Artefakt, das durch die Mittelwertersetzung zustande kommt. Wie weiter unten gezeigt wird (s. Tab. **4**), muss davon ausgegangen werden, dass der Mittelwert in X1 die fehlenden Werte in X1 überschätzt und der Mittelwert in X2 die fehlenden Werte in X2 unterschätzt. Wie man gut nachvollziehen kann, sind diese Schätzfehler die alleinige Ursache für den negativen Zusammenhang nach der Mittelwertersetzung. Ein besserer Ansatz, der auf einem ähnlichen Prinzip wie die Mittelwertersetzung beruht, ist die Regressionsersetzung [3,11], bei der zusätzlich die Zufallsvariabilität der Daten berücksichtigt werden kann. Diese kann im Falle „zufälligen Fehlens“ (MAR) verwendet werden. Da diese aber den im Folgenden vorgestellten Verfahren unterlegen sind, wird auf deren Darstellung hier verzichtet.

Tab. **4** Ergebnis der Ersetzung mittels des EM-Algorithmus. Die ersetzten Werte sind links kursiv eingetragen, rechts sind die imputierte Korrelationsmatrix, die Mittelwerte und Standardabweichungen angegeben

X1	*X2*	*Y*			*X1*	*X2*	*Y*
1,81	2	1	**r**	**X1**	1	0,90	0,92
1,86	1	2	**r**	**X2**		1	0,90
3,24	3	3	**r**	**Y**			1
3	5	4	**MW**		5,45	4,63	6,00
4	4	5	**SD**		2,69	2,14	3,32
6	5	6					
5	4	7					
7	*5,81*	8					
8	*6,49*	9					
9	*7,16*	10					
9	*7,46*	11					

Es bleibt festzuhalten, dass die „Ersetzung durch den Mittelwert“ nicht verwendet werden sollte und die Verfahren „fallweiser“ und „paarweiser Ausschluss“ nur bei „vollständig zufälligem Fehlen“ (MCAR) angewendet werden dürfen. Da dies aber in der empirischen Forschung zumeist eine unrealistische Annahme ist, soll kurz die Logik des E(xpectation)-M(aximation)-Algorithmus und der multiplen Ersetzung erläutert werden, die auch im Falle von „zufälligem Fehlen“ (MAR) eingesetzt werden können, ohne dass eine Verzerrung der Ergebnisse resultiert.

Der EM-Algorithmus ist ein Schätzalgorithmus, der die fehlenden Daten so ersetzt, dass die gesamte Information im Datensatz

[6] Eine Informationsmatrix, die eine solche nicht mögliche Gesamtstruktur enthält, bezeichnet man in der Mathematik als „nicht positiv definit“.

[7] Eine Ausnahme stellt die Berechnung von Summenwerten aus mehreren Einzelvariablen dar. Fehlen die Werte von Einzelvariablen, so ist die Ersetzung durch den Variablenmittelwert im Allgemeinen zulässig [5].

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

in sich widerspruchsfrei und maximal plausibel ist [1,2,5,6]. Dies geschieht nach folgendem Prinzip:

1. Es wird angenommen, dass die im Datensatz vorliegende Informationsstruktur (Mittelwerte, Standardabweichungen und Korrelationen) gültig ist. Die fehlenden Werte werden gemäß dieser Informationen geschätzt und ersetzt.
2. Für diesen neuen Datensatz werden dann wieder die Mittelwerte, Standardabweichungen und Korrelationen berechnet. Haben diese sich durch diese Ersetzung im Vergleich zum Originaldatensatz verändert, so kann die ursprüngliche Informationsstruktur nicht gültig sein: *denn eine plausible Informationsstruktur muss sowohl für die vorhandenen als auch die fehlenden Werte gelten*. Die neue Informationsstruktur des vervollständigten Datensatzes wird nun beibehalten und die fehlenden Werte werden *gemäß dieser modifizierten Informationen* erneut ersetzt. Anschließend wird wieder überprüft, ob sich die Informationsstruktur verändert hat.
3. Diese Schleife wird so lange durchlaufen, bis eine Informationsstruktur identifiziert wurde, die sich nach der Ersetzung nicht mehr verändert.

Wie kann man sich nun dieses Vorgehen für den Datensatz in Tab. **1** vorstellen? Der Algorithmus geht von den Dateninformationen bei paarweisem Ausschluss (Tab. **2**, Mitte) aus. X1 und X2 korrelieren demnach hoch mit Y. Dies bedeutet, dass für die fehlenden Werte in X1 tendenziell niedrige Werte ersetzt werden, da Y für die Personen 1 – 3 unterdurchschnittlich ausgeprägt ist. Für X2 werden tendenziell höhere Werte eingesetzt, da in Y für die Fälle 8 – 11 hohe Werte vorliegen. Wie man nun gut nachvollziehen kann, entsteht durch diese Ersetzung eine positive Korrelation in X1 und X2, die ursprünglich unkorreliert waren.[8] Nach dieser ersten Simulation hat sich die Korrelationsstruktur also verändert. Deswegen werden ausgehend von dieser neuen Informationsgrundlage die fehlenden Werte erneut ersetzt. In unserem Beispiel ergibt sich eine stabile Lösung erst, nachdem diese Schleife 97-mal durchlaufen wurde. Tab. **4** enthält die durch den EM-Algorithmus ersetzten Werte und die resultierende Informationsstruktur.

Das Ergebnis zeigt, dass davon ausgegangen werden muss, dass die „Zufriedenheit mit dem Beruf" (X1) und die „Zufriedenheit mit dem Einkommen" (X2) mit $r = 0{,}90$ hoch korreliert sind. Die Mittelwerte haben sich zudem mit $M_1 = 5{,}54$ und $M_2 = 4{,}63$ einander angenähert, und die Standardabweichungen sind mit $SD_1 = 2{,}69$ und $SD_2 = 2{,}14$ deutlich größer geworden. Die ursprünglich fehlende Korrelation von X1 und X2, der hohe Mittelwertsunterschied und die zu geringen Standardabweichungen dieser beiden Merkmale können demnach plausibel als Artefakte angesehen werden, die durch die Verweigerung der Datenangabe verursacht wurden.

Es lässt sich also festhalten, dass hier – neben den Datenangaben selbst – Informationen darüber verwertet werden, weshalb Antworten verweigert wurden: Die Informationen werden ausgenutzt, um die fehlende Information bestmöglich zu rekonstruieren. Der EM-Algorithmus erreicht dies, indem eine Informationsstruktur identifiziert wird, die eine Ersetzung gewährleistet, deren Ergebnis die zugrunde gelegte Struktur bestätigt. Im Falle „zufälligen Fehlens" (MAR) werden hierdurch systematische Verzerrungen der Ergebnisse vermieden, und das Problem des durch fehlende Werte effektiv verringerten Stichprobenumfangs wird zumindest reduziert [1,2,6]. Dieser zweite Vorteil gilt auch bei „vollständig zufälligem Fehlen" (MCAR) im Vergleich zu „paarweisem" und „fallweisem Ausschluss".

In Simulationen hat sich gezeigt, dass der EM-Algorithmus im Falle von „vollständig zufälligem Fehlen" (MCAR) und „zufälligem Fehlen" (MAR) zuverlässige Ergebnisse liefert, wenn bis zu 30% der Werte in einzelnen Variablen fehlen und genügend große Stichprobenumfänge vorliegen [2]. Dabei muss beachtet werden, dass die Stichprobenumfänge größer als üblich[9] sein müssen, da die Daten aufgrund der fehlenden Werte nicht vollständig sind.

Obwohl der EM-Algorithmus theoretisch nicht bei „nicht zufällig fehlenden" (NRM) Daten angewendet werden sollte, erweist sich die Prozedur als sehr robust bei moderaten Verletzungen der Bedingung „zufällig fehlend" (MAR) [12]. Gerade in gesundheitswissenschaftlichen Studien liegen in Datensätzen zumeist zu jedem Merkmal substanziell korrelierte Merkmale vor, so dass diese Ersetzungsmethode im Allgemeinen die Methode der Wahl darstellt. Für „nicht zufällig fehlende" (NRM) Daten wurden spezielle moderne Methoden entwickelt, die jedoch nicht standardisiert umgesetzt werden können und eine hohe inhaltliche und methodische Kompetenz bei der Auswertung voraussetzen [5,6].

Eine Erweiterung der Idee des EM-Algorithmus stellt die so genannte „multiple Ersetzung" („multiple imputation", [13]) dar. Die Daten werden hier nach einem ähnlichen Prinzip ersetzt, jedoch nicht nur einmal, sondern mehrfach. Es entstehen mehrere vervollständigte Datensätze, die unabhängig voneinander statistisch analysiert werden. Die Variabilität der resultierenden Lösungen wird dann zusätzlich untersucht. Je geringer die Unterschiede zwischen den Analyseergebnissen für die verschiedenen Datensätze sind, desto höher ist die Genauigkeit und Zuverlässigkeit, mit der die fehlenden Werte ersetzt werden können.

Mit welcher Software können die empfohlenen Methoden umgesetzt werden?

Das Zusatzmodul Missing-Value-Analysis (MVA) im Statistikprogramm SPSS ermöglicht die Durchführung einer differenzierten Missing-Data-Diagnose und die Ersetzung mittels des EM-Algorithmus. Die Programme LISREL, AMOS und Mplus verwenden einen ähnlichen Algorithmus (Full-Information-ML), der insbesondere bei der Analyse von Strukturgleichungsmodellen leichte

[8] Zwar berücksichtigt der Algorithmus auch die Information, dass X1 und X2 ursprünglich unkorreliert waren, diese Information wird aber allein deswegen geringer gewichtet, weil die Berechnung nur auf vier Fällen beruht.

[9] „Üblicherweise" sollte der Stichprobenumfang so gewählt werden, dass die als empirisch relevant erachteten Effekte zuverlässig entdeckt werden können [9]. Um die Generalisierbarkeit von Effekten sicherzustellen, sollte beispielsweise in der multiplen linearen Regression das Verhältnis „Personen- zu Prädiktorenanzahl" mindestens 15 betragen, wenn keine fehlenden Werte vorliegen [3]. Da dem EM-Algorithmus eine Maximum-Likelihood-Schätzung zugrunde liegt, sollte das Verfahren grundsätzlich erst ab mindestens 100 Personen angewendet werden [5].

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

Vorteile gegenüber dem EM-Algorithmus besitzt [6,14]. Der EM-Algorithmus ist ebenso wie die multiple Imputation (PROC MI) in SAS verfügbar [15]. Zur multiplen Imputation existiert eine Homepage von Joe Schafer, auf der weitere Programme dokumentiert sind.[10]

Zusammenfassende Schlussfolgerungen

Fehlende Werte stellen ein oft unterschätztes Problem der empirischen Forschung dar, da sie den effektiven Stichprobenumfang verringern und die resultierenden Teststatistiken erheblich verfälschen können. In einer Missing-Data-Diagnose können Hinweise auf die Ursachen fehlender Werte gewonnen werden, und es wird deutlich, welche Gefahren bei der Interpretation der Daten durch fehlende Angaben entstehen können. Standardmethoden zum Umgang mit fehlenden Werten (paarweiser oder fallweiser Ausschluss) basieren auf der zumeist nicht haltbaren Annahme „vollständig zufällig fehlender" (MCAR) Werte. Fehlen die Werte lediglich „zufällig" (MAR), so korrigiert der EM-Algorithmus die Einflüsse fehlender Informationen auf die Korrelationen, Mittelwerte und Standardabweichungen der Variablen im Datensatz. Der EM-Algorithmus ist somit den klassischen Verfahren überlegen und sollte bevorzugt angewendet werden. Bei der „multiplen Ersetzung" wird zusätzlich die Genauigkeit und Zuverlässigkeit der Datenersetzung überprüft.

Diese beiden Methoden können auch angewendet werden, wenn die Bedingung „zufällig fehlend" (MAR) nur moderat verletzt ist, die Daten also eigentlich „nicht zufällig fehlen" (NRM). Je mehr Merkmale erhoben werden, die mit Variablen mit fehlenden Angaben korrelieren, und je stärker diese korrelieren, desto gesicherter kann davon ausgegangen werden, dass die Bedingung „zufällig fehlend" (MAR) approximativ gilt.

Dies hat für die Planung von Studien wichtige Implikationen. Man sollte bei der Datenerhebung nicht nur die Merkmale berücksichtigen, die für die Überprüfung der inhaltlichen Fragestellungen notwendig sind, sondern es sollte zusätzlich explizit überlegt werden, weshalb Personen einzelne Fragen u.U. nicht beantworten könnten oder – in Studien mit mehreren Messzeitpunkten – weshalb Personen aus der Studie ausscheiden könnten. Werden diese Informationen erhoben, so wird das Fehlen von Daten vorhersagbar (MAR statt NRM), und die fehlenden Informationen können durch die vorgeschlagenen Methoden effizient ersetzt werden.

Die im vorliegenden Text behandelten Verfahren sind für multivariat normalverteilte intervallskalierte Merkmale geeignet. Fehlen in einem Messwiederholungsdesign Werte zu bestimmten Messzeitpunkten, so können die behandelten Verfahren ebenfalls eingesetzt werden. Eine Adaptation für ordinale Daten ist auf Basis der Theorie von Agresti [16] möglich. Eine gute Zusammenfassung von speziellen Anwendungssituationen und Problemlösungsansätzen gibt Allison [6].

Es sei abschließend nochmals darauf hingewiesen, dass alle Verfahren an Voraussetzungen geknüpft sind, deren Gültigkeit zumeist nicht explizit überprüft werden kann. Deswegen sind fehlende Werte stets ein Problem: Alle vorgestellten Verfahren sollten besser als „Methoden zur Übelminimierung" denn als „Lösung" bezeichnet werden. Es soll nicht der Eindruck erweckt werden, dass nicht alles dafür getan werden müsse, um fehlende Werte zu vermeiden, denn „the only really good solution to the missing data problem is not to have any" ([6], S. 2). Können fehlende Werte jedoch nicht vermieden werden, so birgt die Anwendung üblicher Standardmethoden zum Umgang mit fehlenden Werten große Gefahren, die durch die in diesem Text dargestellten Methoden, die an schwächere Annahmen gebunden sind, zumindest entschärft werden können.

Literatur

1 Schafer J. Analysis of incomplete multivariate data. London: Chapman & Hall, 1997
2 Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley, 2002
3 Hair JF, Anderson RE, Tatham RL, Black W. Multivariate data analysis. New Jersey: Prentice Hall, 1998
4 Rubin DB. Inference and Missing Data. Biometrika 1976; 63: 581 – 592
5 Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychological Methods 2002; 7: 147 – 177
6 Allison PD. Missing data. Thousand Oaks: Sage, 2001
7 Rost J. Testtheorie und Testkonstruktion. Göttingen: Huber, 1996
8 King G, Honaker J, Joseph A, Scheve K. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. American Political Science Review 2001; 95: 49 – 69
9 Bock J. Bestimmung des Stichprobenumfangs. München: Oldenbourg, 1998
10 Wirtz M, Nachtigall C. Deskriptive Statistik. Weinheim: Juventa, 2002
11 Müller JM. Umgang mit fehlenden Werten. In: Reusch A, Zwingmann C, Faller H (Hrsg). Empfehlungen zum Umgang mit Daten in der Rehabilitationsforschung. Regensburg: Roderer, 2002: 109 – 125
12 Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing-data procedures. Psychological Methods 2001; 6: 330 – 350
13 Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research 1998; 33: 545 – 571
14 Arbuckle JL, Wothke W. AMOS 4.0 User's Guide. Chicago: Smallwaters, 1995
15 Yuan YC. Multiple imputation for missing data. In: Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference (Paper No. 267). Cary, NC: SAS Institute, 2000
16 Agresti A. An introduction to categorical data analysis. New York: Wiley, 1996

[10] „Software for multiple imputation", URL: www.stat.psu.edu/%7Ejls/misoftwa.html (zuletzt aufgerufen Februar 2004).

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.