Halle/Saale, 08.06.2004

Behandlung fehlender Werte

Dipl.-Psych. Wilmar Igl
- Methodenberatung -

Rehabilitationswissenschaftlicher
Forschungsverbund
Bayern

Einleitung (1)

- Fehlende Werte als allgegenwärtiges Problem der Forschung, auch in den Gesundheitswissenschaften

- Probleme durch fehlende Werte:
  - Fehlerquellen in der Studie möglich
  - Verzerrung (*bias*) der Ergebnisse
  - Verringerung der Effizienz von statistischen Verfahren

Einleitung (2)

- Vorteile der Beschäftigung mit fehlenden Werten:
  - Analyse von Fehlwerten zur Verbesserung der Studie
  - Erkennen von möglichen Verzerrungen
  - Auswahl geeigneter Methoden zur Behandlung von Fehlwerten
  - Steigerung der Effizienz der statistischen Auswertung
- Aktueller Stand:
  - noch geringes Problembewusstsein bei vielen Statistik-Anwendern
  - geeignete Methoden sind noch nicht standardmäßig Statistik-Programmen enthalten (z.B. SPSS, SAS) bzw. sind Zusatz-module und -programme notwendig

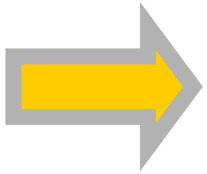
Ziele des Vortrags

- Schaffung von Problembewusstsein
- Grundlagen zum Thema
- Methoden zur Analyse von Fehlwerten
- Methoden zur Behandlung von Fehlwerten
- Anwendungsempfehlungen

I. Grundlagen zu Fehlwerten

Was ist ein fehlender Wert?

- Erwartete fehlende Werte (*intentional missing*):
  Merkmal existiert nicht in der Realität
  Beispiel: Beurteilung der Arbeitszufriedenheit bei Arbeitslosen
- Unerwartete fehlende Werte ("echte Fehlwerte"):
  Merkmal existiert in der Realität, aber keine Daten vorhanden
  Beispiel: Frage zur Arbeitszufriedenheit wurde von Probanden, der Arbeit hat, übersehen



In diesem Vortrag geht es um unerwartete Fehlwerte (missing data (MD), missing values) !

Ursachen für fehlende Werte

- Untersucher z.B. Unterschätzung des Umfangs der Untersuchung/ Belastung des Patienten,...
- Instrument z.B. unklare Fragen, unpassende Antworten, mangelnde Übersichtlichkeit,...
- Proband z.B. mangelnde Compliance/ Aufmerksamkeit, Scham, ...
- Dateneingabe z.B. unzuverlässige Hilfskräfte,...
- Auswertung z.B. "Divison by Zero Error", Ausschluss von Ausreissern,...
- Sonstiges z.B. Datenverlust durch EDV-Probleme, Fehler der Post,...

Missing Data Prozesse

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Non-Missing At Random (NMAR)/ Informative Drop-Out
  (vgl. Little & Rubin, 2002)

⇒ Abhängig von zugrundeliegendem Missing Data Prozess können Verzerrungen der Daten/Ergebnisse auftreten!

⇒ Vorbedingung für die Auswahl von Methoden zur Behandlung von Fehlwerten!

Missing Completely At Random (MCAR)

- Annahme: Das Auftreten eines fehlenden Wertes in der Variable Y ist nicht abhängig
  a) von den Ausprägung der Variable Y selbst oder
  b) den restlichen Variablen X_1 bis X_n im Datensatz.

- Beispiele:
  - Fehler der Post, Dateneingabefehler ("Vertipper"),... => MCAR
  - Fehlwerte abhängig von der "Motivation zur Studienteilnahme", fehlende Angaben zum "Ausmaß des Suchtmittelgebrauchs" bei stark Drogenabhängigen => kein MCAR

MCAR - Diagramm

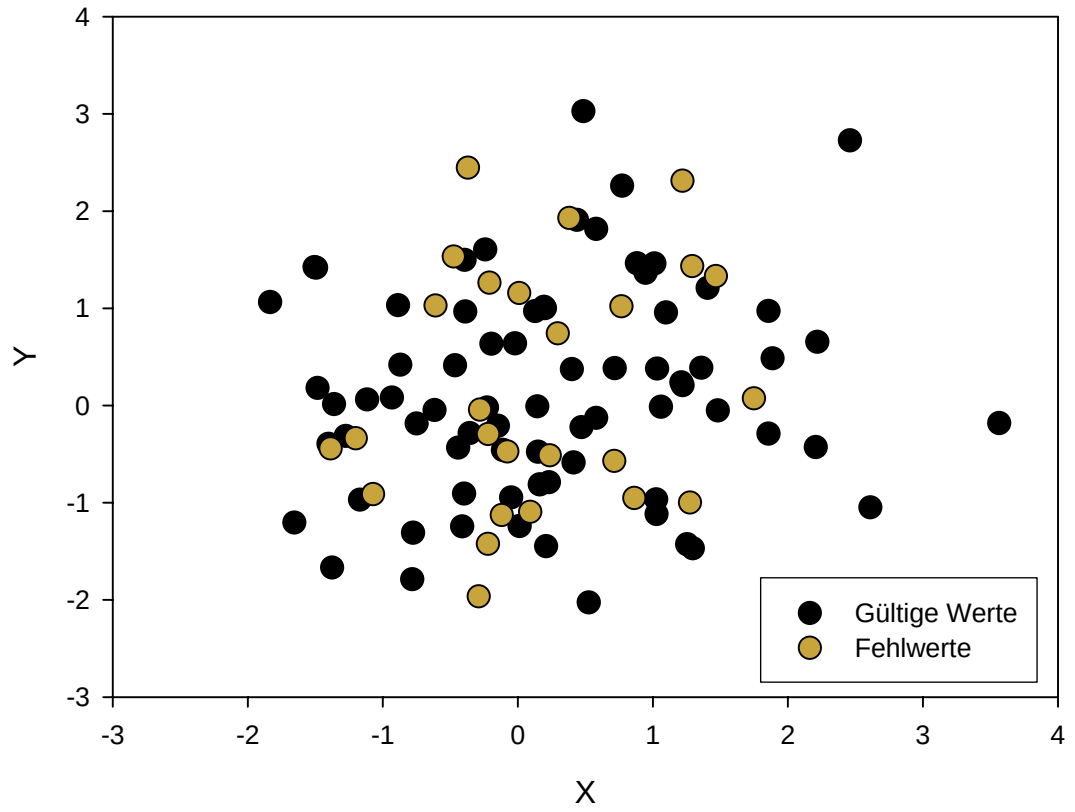


Abb. 1: Streudiagramm mit Fehlwerten (braun) unter der Bedingung MCAR

Missing At Random (MAR)

- Annahme: Das Auftreten eines fehlenden Wertes in einer Variable Y ist vollständig durch die Ausprägungen der restlichen Variablen X_1 bis X_n erklärbar.

- Beispiele:
  - Fehlwerte abhängig von der "Motivation zur Studienteilnahme" => MAR
  - fehlende Angaben zum Ausmaß des Drogenkonsums kann nicht aus anderen Variablen erklärt werden, sondern nur aus dem Drogenkonsum selbst => kein MAR

MAR - Diagramm

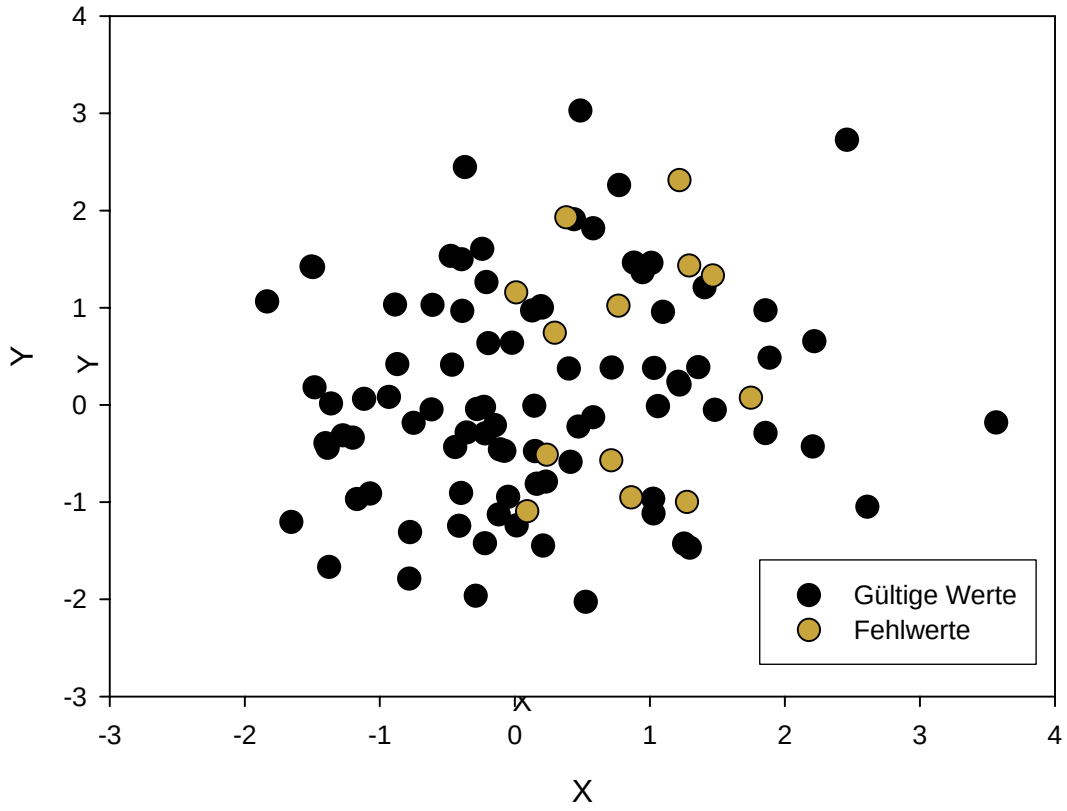


Abb. 2: Streudiagramm mit Fehlwerten (braun) unter der Bedingung MAR

Not Missing At Random (NMAR)/ Informative Drop-Out

- Annahme: Das Auftreten von fehlenden Werten in der Variable Y ist
  a) von der (unbekannten) Ausprägung der Variable Y abhängig
  b) ist nicht durch die Ausprägungen der übrigen Variabeln X_1 bis X_n erklärbar.

- Beispiele:
  - fehlende Angaben zum Einkommen von besonders gut verdienenden Personen, fehlende Angaben von Drogenabhängigen zu Fragen zum Suchtmittelgebrauch => NMAR
  - Fehlende Angaben zum Einkommen können aus dem Bildungsstand ermittelt werden => Kein NMAR (sondern MAR)

NMAR - Diagramme

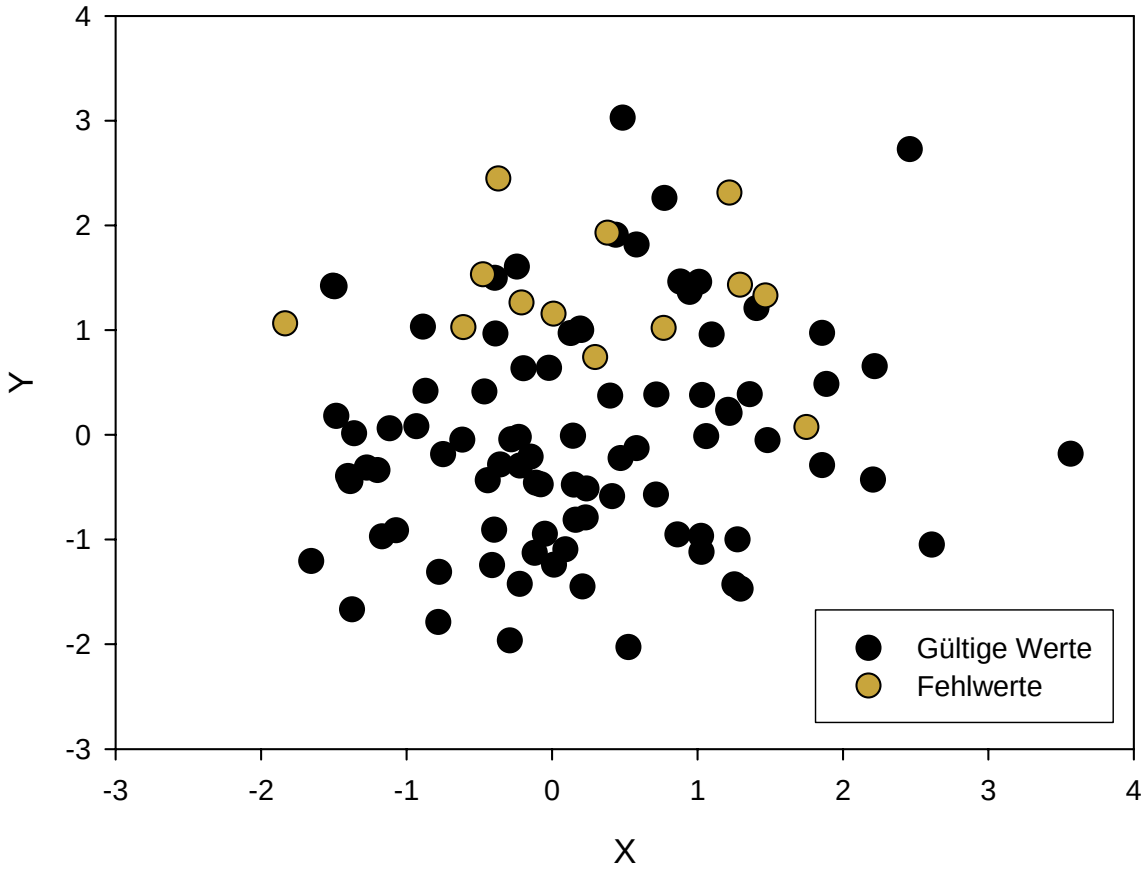


Abb. 2: Streudiagramm mit Fehlwerten (braun) unter der Bedingung NMAR

II. Analyse von Fehlwerten und zugrundeliegender Prozesse

Diagnoseschritte

1. Erstellung einer Indikatormatrix der Fehlwerte
2. Berechnung des Anteils fehlender Werte
   - pro Fall
   - pro Variable
3. Untersuchung häufig auftretender Muster fehlender Werte
4. Untersuchung von Gruppenunterschieden zwischen Responder vs. Non-Respondern
5. Untersuchung von Korrelationen zwischen Indikatorvariablen

Datenbeispiel

	Daten		
Fall	**X1**	**X2**	**Y**
1	?	2	1
2	?	1	2
3	?	3	3
4	6	5	4
5	4	4	5
6	4	5	6
7	6	4	7
8	7	?	8
9	8	?	9
10	9	?	10
11	9	?	11

Zufriedenheit mit...

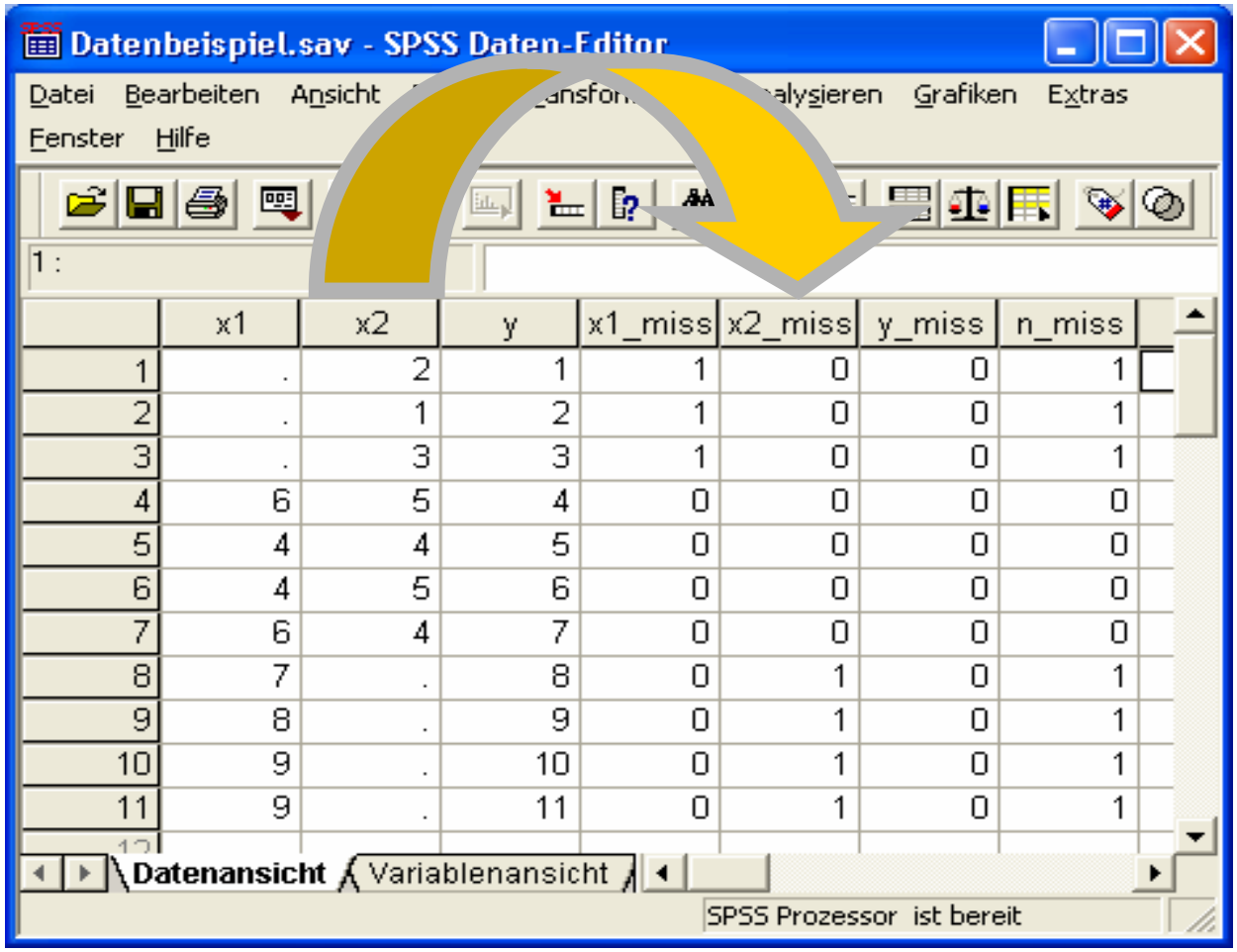
X1 = Beruf

X2 = Einkommen

Y= Leben allgemein

Tab. 1: Datenbeispiel aus Wirtz (2004)

Erstellung der Indikatormatrix



	x1	x2	y	x1_miss	x2_miss	y_miss	n_miss
1	.	2	1	1	0	0	1
2	.	1	2	1	0	0	1
3	.	3	3	1	0	0	1
4	6	5	4	0	0	0	0
5	4	4	5	0	0	0	0
6	4	5	6	0	0	0	0
7	6	4	7	0	0	0	0
8	7	.	8	0	1	0	1
9	8	.	9	0	1	0	1
10	9	.	10	0	1	0	1
11	9	.	11	0	1	0	1

Umkodierung:

Fehlwert = 1

gültiger Wert = 0

Abb. 3: Datenbeispiel aus Wirtz (2004)

Anteil fehlender Werte

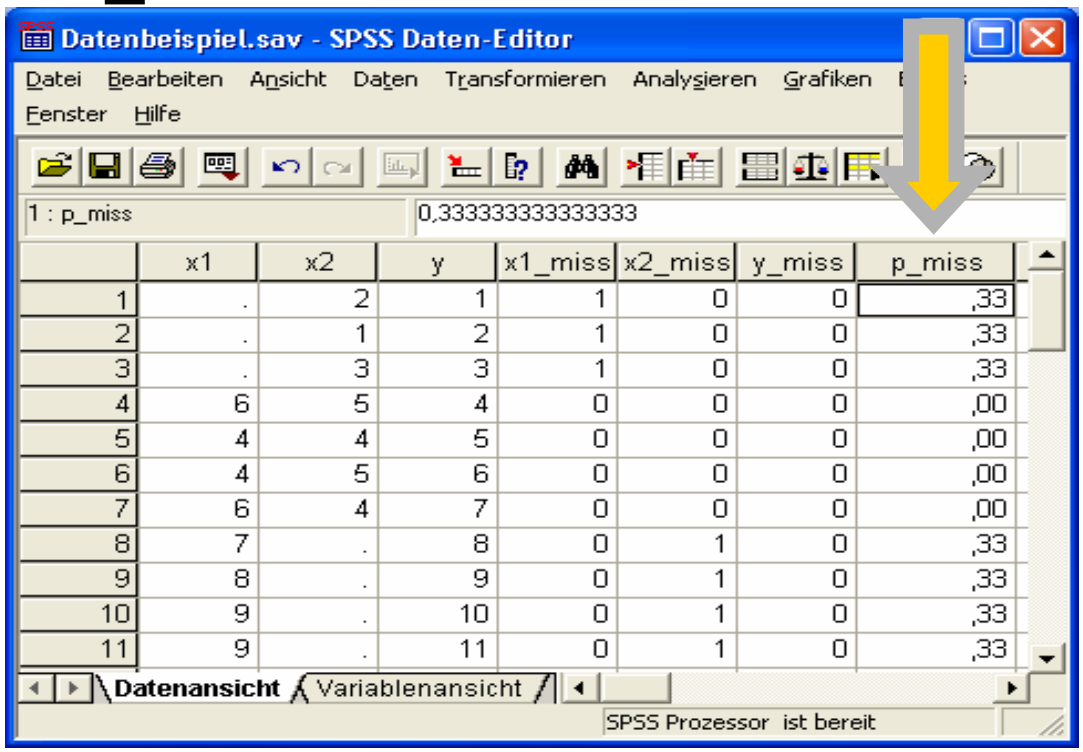


Abb. 4: Anteil fehlender Werte pro Fall

X1_MISS

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	0	8	72,7	72,7	72,7
	1	3	27,3	27,3	100,0
	Gesamt	11	100,0	100,0	

Tab. 2: Anteil fehlender Werte pro Variable

Erkennen häufiger Muster fehlender Werte

x1_miss	x2_miss	x3_miss	y_miss	**f(Muster)**
0	0	0	0	4
0	1	0	0	3
1	0	1	0	3
0	1	1	0	1

Zufriedenheit mit...

X1 = Beruf

X2 = Einkommen

X3 = Kollegen

Y= Leben allgemein

Tab. 3: Muster von fehlenden Werten und ihre Häufigkeit

Besonders häufig auftretende Muster weisen auf systematisches Auftreten fehlender Werte hin!

Gruppenunterschiede

- Testung von Gruppenunterschieden in den Variablen x1, x2, y basierend auf den Indikatorvariablen x1_miss, x2_miss, y_miss

Unabhängige Variable	**Abhängige Variable**		
	M(Y), wenn Miss = 0	**M(Y), wenn Miss=1**	**$t_{df=8}$ (Signifikanz)**
X1_miss	2,0	7,5	-3,67 (p=,005**)
X2_miss	9,0	3,5	5,20 (p=,001**)

Tab. 4: Ergebnisse der Testung auf Gruppenunterschiede

Signifikante Gruppenunterschiede weisen auf systematisches Auftreten fehlender Werte hin!

Korrelationen zwischen Fehlwerten

Korrelationen

		X1_MISS	X2_MISS	Y_MISS
X1_MISS	Korrelation nach Pearson	1	-,463	.[a]
	Signifikanz (2-seitig)	.	,152	.
	N	11	11	11
X2_MISS	Korrelation nach Pearson	-,463	1	.[a]
	Signifikanz (2-seitig)	,152	.	.
	N	11	11	11
Y_MISS	Korrelation nach Pearson	.[a]	.[a]	.[a]
	Signifikanz (2-seitig)	.	.	.
	N	11	11	11

a. Kann nicht berechnet werden, da mindestens eine der Variablen konstant ist.

Tab. 4: Korrelationen der Indikatormatrix

Signifikante Korrelationen weisen auf systematisches Auftreten fehlender Werte hin!

III. Methoden zur Behandlung von fehlenden Werten

Allgemeine Ansätze

- Fallausschluss:
  - Listenweiser Fallausschluss (*listwise deletion, complete case approach*)
  - Paarweiser Fallausschluss (*pairwise deletion, available case approach*)
- Einfache Imputation (=Ersetzung):
  - Mittelwertersetzung (*mean substitution*)
  - EM/FIML-Algorithmus (=>*state of the art!*)
- Multiple Imputation (=>*state of the art!*)

Datenbeispiel

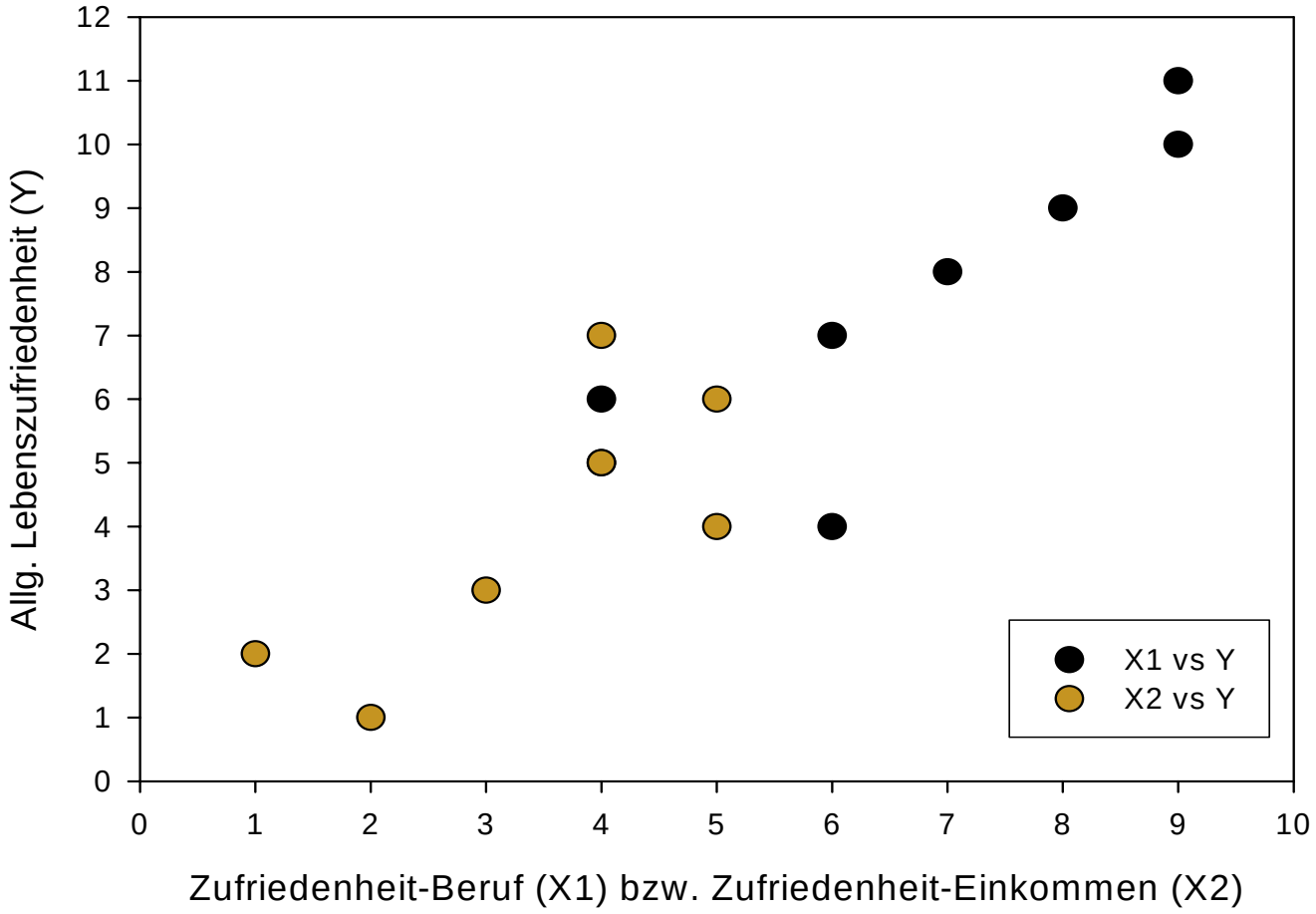


Abb. 5: Datenbeispiel aus Wirtz (2004)

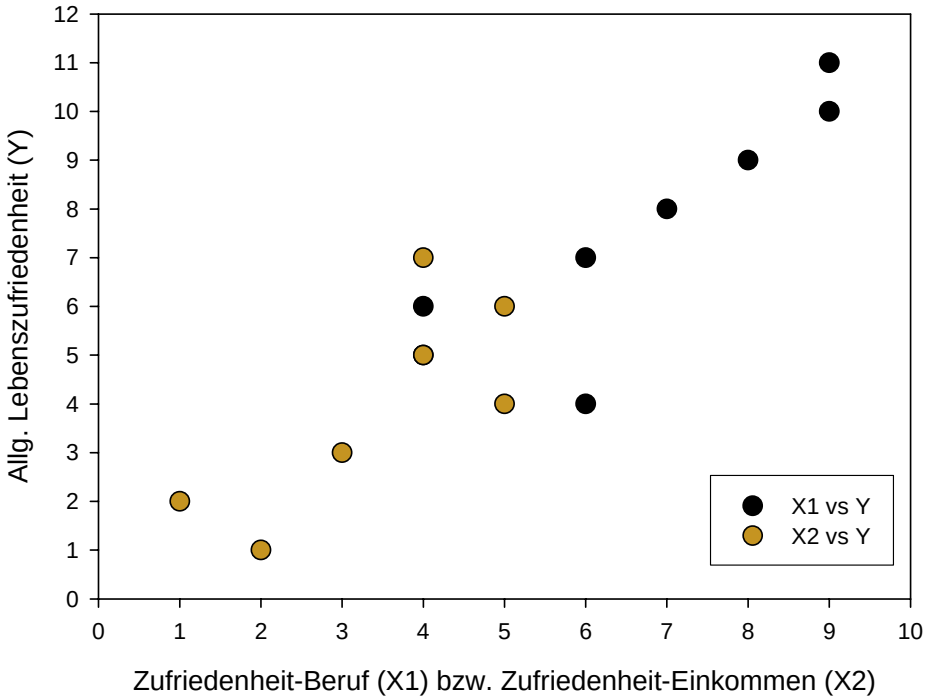
Listenweiser Fallausschluss

- Vorgehen: Ausschluss aller unvollständiger Fälle/ Variablen („*complete case approach*“)

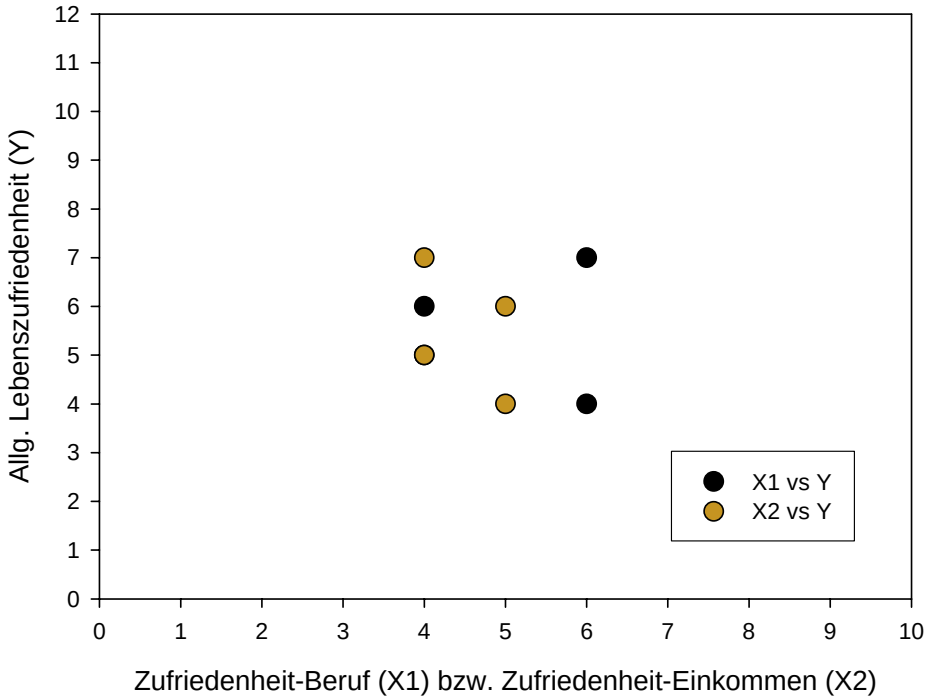
- Anwendung bei:
  - MCAR
  - große Stichprobe
  - starke Effekte

- Nachteile:
  - Reduktion der Stichprobe bis zur Unbrauchbarkeit möglich
  - MCAR ist in der Forschungspraxis selten gegeben

LW-Fallausschluss: Datenbeispiel



Originaldaten



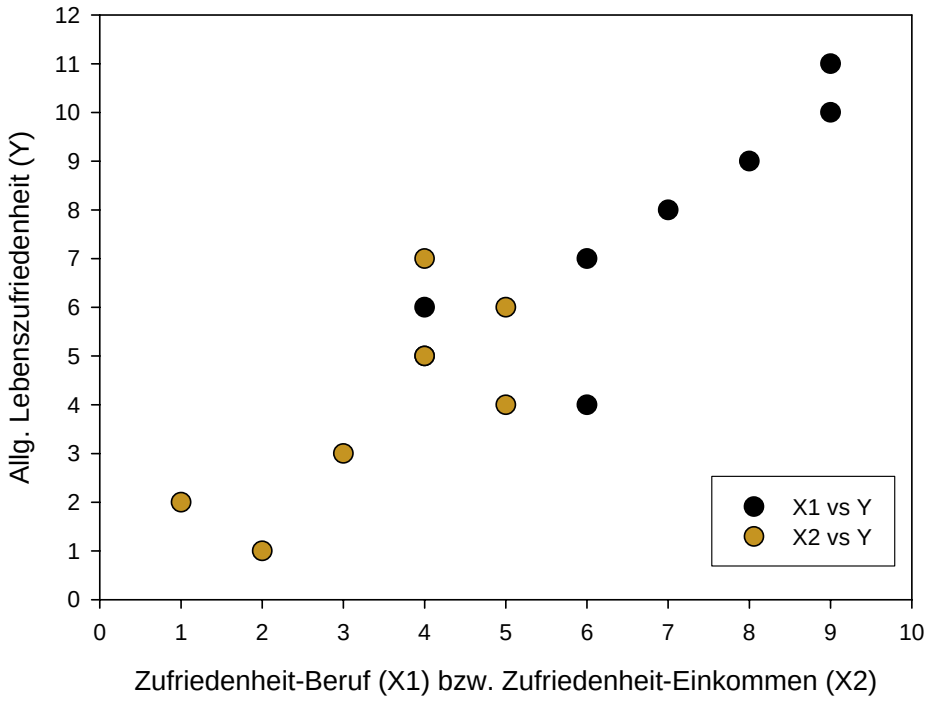
Listenweiser Fallausschluss

Abb. 6: Listenweiser Fallausschluss am Datenbeispiel von Wirtz (2004)

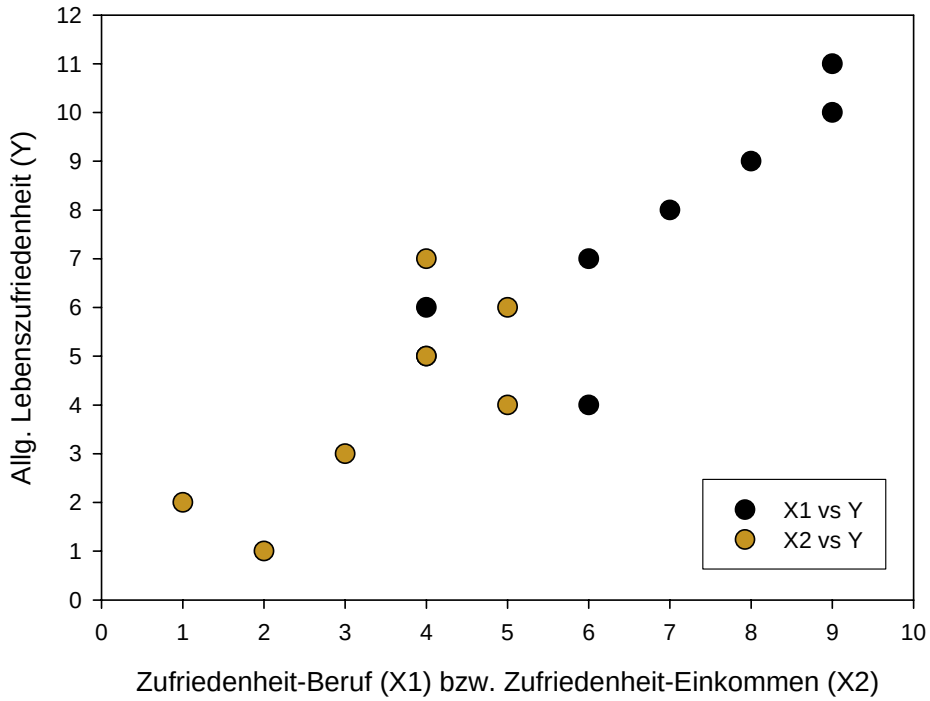
Paarweiser Fallausschluss

- **Vorgehen**: Alle gültigen Fälle, der in die Berechnung eingehenden Variablen, werden ausgewertet. Verteilungscharakteristika der gültigen Werte werden übernommen ("*available cases approach*")
- **Anwendung bei:**
  - MCAR
  - Berechnung von Korrelationen, Mittelwerten, Streuungen
- **Nachteile:**
  - Statistiken können auf unterschiedlichen Stichproben von Beobachtungen basieren (unterschiedliches N !)
  - mathematische Inkonsistenzen möglich (z.B. zwischen Korrelationen zweier Variablen X, Y und deren Partialkorrelationen mit Z)

PW-Fallausschluss: Datenbeispiel



Originaldaten



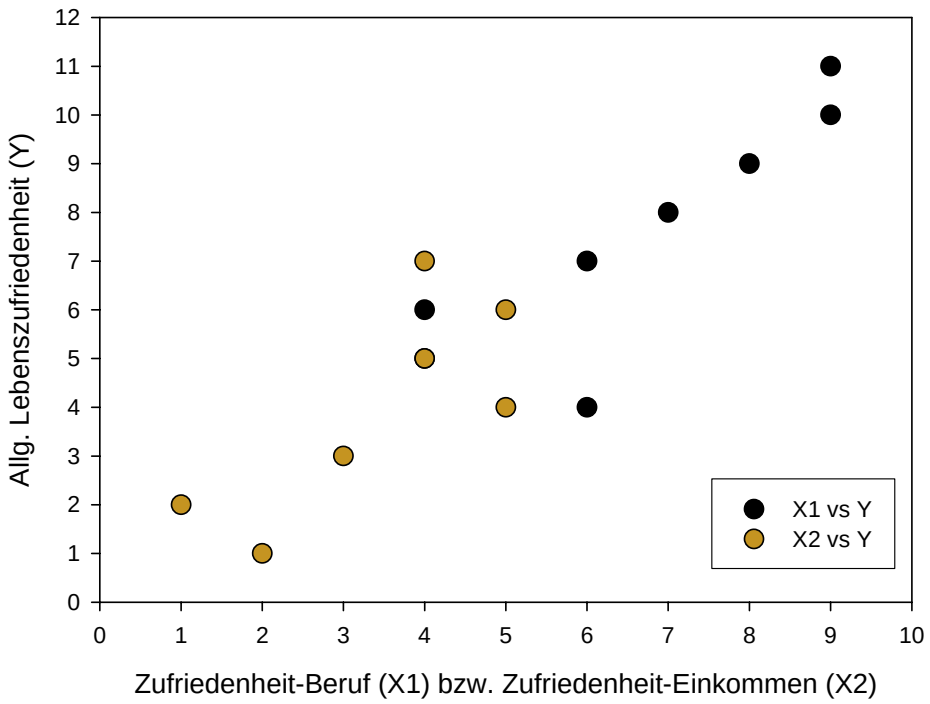
Paarweiser Fallausschluss

Abb. 7: Paarweiser Fallausschluss am Datenbeispiel von Wirtz (2004)

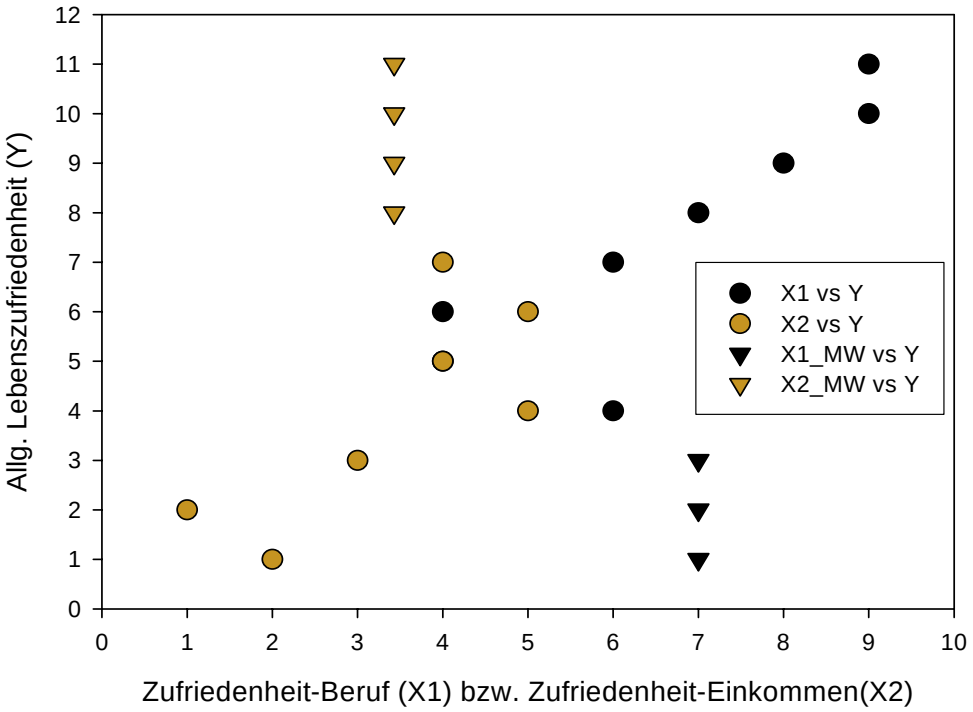
Mittelwertersetzung (MWE)

- Vorgehen: Ersetzen der fehlenden Werten pro Variable durch den Mittelwert der gültigen Werte der jeweiligen Variable
- Vorteile:
  - einfache Anwendung
  - vollständiger Datensatz
- Nachteile (auch bei MCAR!):
  - MWE ist nur unter MCAR bei der Berechnung von Summen und Mittelwerten verzerrungsfrei!
  - Verzerrung der wahren Verteilung
  - Unterschätzung der wahren Varianz
  - Unterschätzung der wahren Zusammenhänge

MWE: Datenbeispiel



Originaldaten



Mittelwertersetzung

Abb. 8: Mittelwertersetzung am Datenbeispiel von Wirtz (2004)

EM-Algorithmus

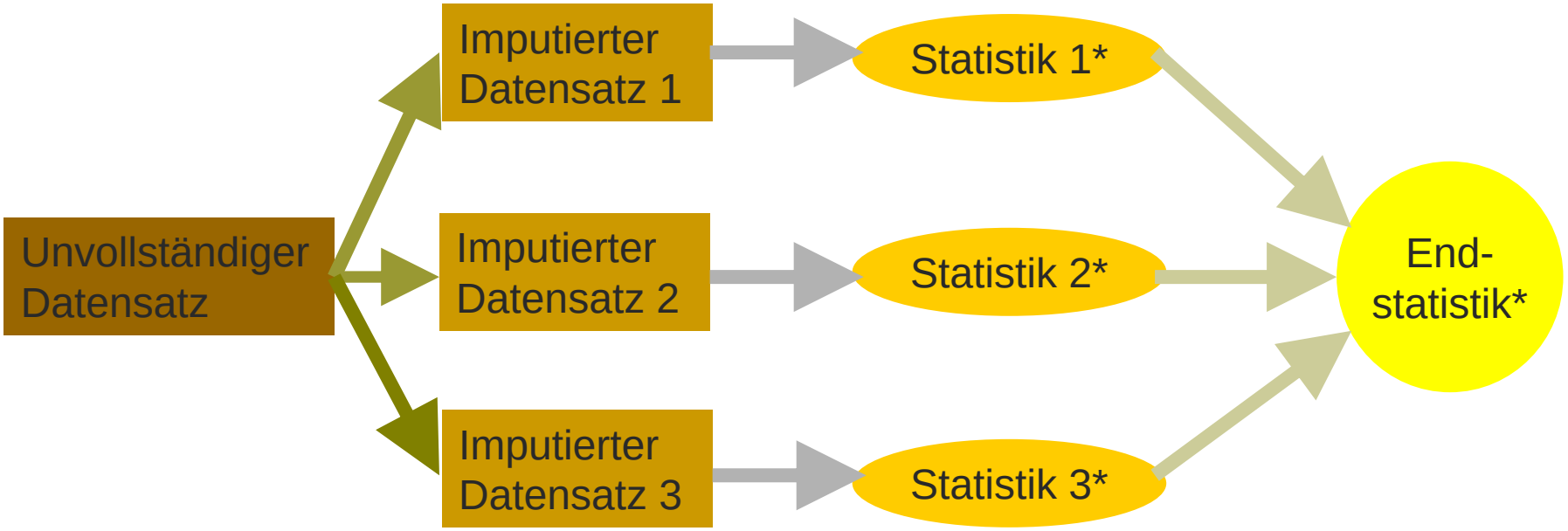
1. Schätzung fehlender Werte auf Grundlage der Parameter der zuletzt geschätzten Daten (M,SD,r)
2. Neue Berechnung von Parametern (M,SD,r) für neue (vollständige) Verteilung
3. Vergleich der Parameter

   a) Parameter(neuer Datensatz) ≠ Parameter(letzter Datensatz), => zurück zu Schritt 1)

   b) Parameter(neuer Datensatz) = Parameter(letzter Datensatz), => Schritt 4)
4. Ersetzung der fehlenden Werte entsprechend den zuletzt berechneten Parametern (M, SD, r).

Multiple Imputation

1) IMPUTATION ⇨ 2) ANALYSE ⇨ 3) INTEGRATION



*Punktschätzer und Standardfehler

Anwendungsvoraussetzungen

- EM-Algorithmus:
  - Multivariate Normalverteilung
  - MAR/MCAR
  - große Stichprobe (>> N=100)
  - ca. 30% Fehlwerte akzeptabel
- Multiple Imputation:
  - Multivariate Normalverteilung
  - MAR/MCAR
  - (große Stichprobe)
  - akzeptabler Anteil fehlender Werte abhängig vom "Anteil fehlender Information"

Beide Verfahren (EM, MI) sind relativ stabil gegen (moderate) Verletzungen ihrer Voraussetzungen!

MI: Datenbeispiel

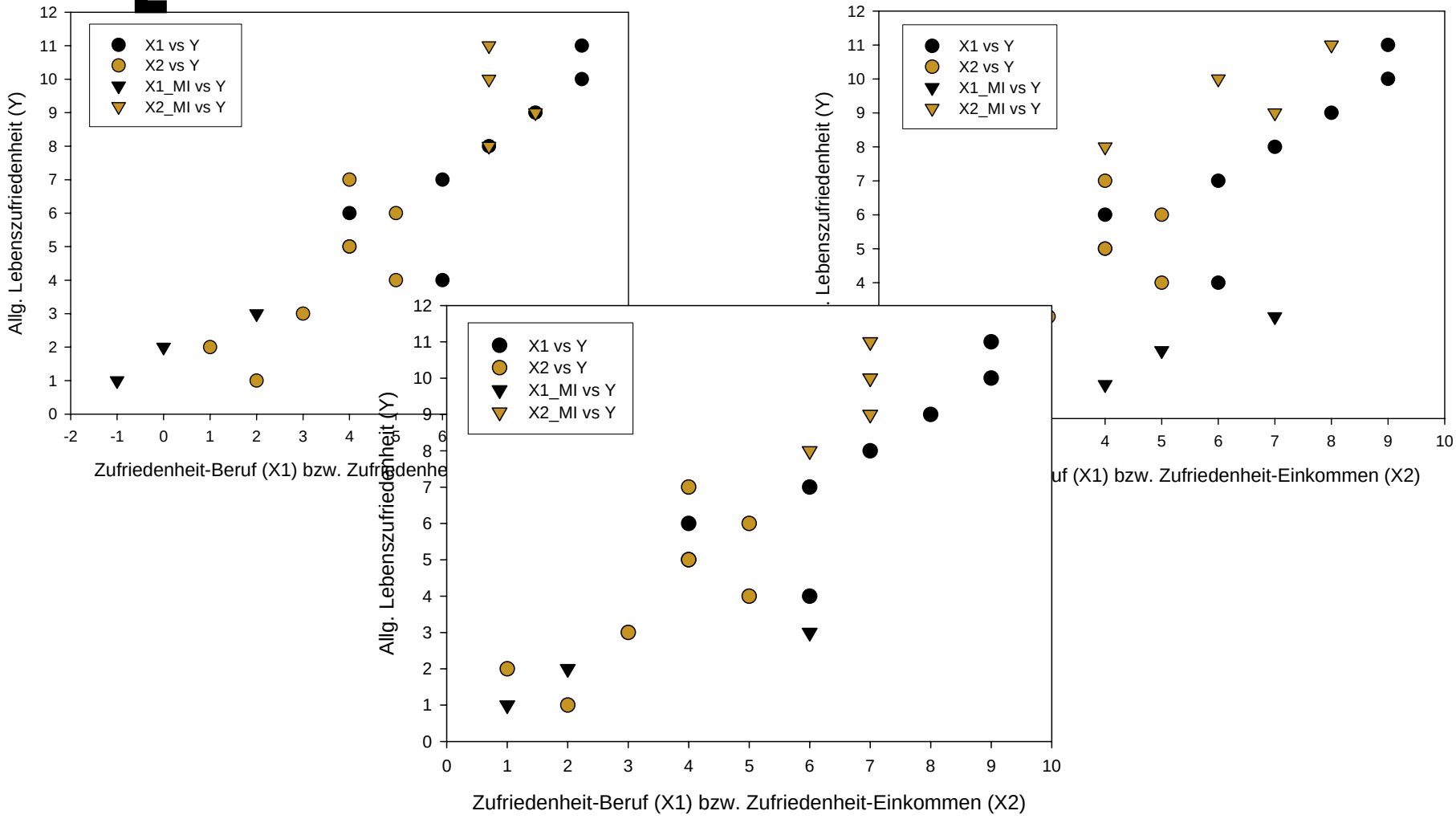
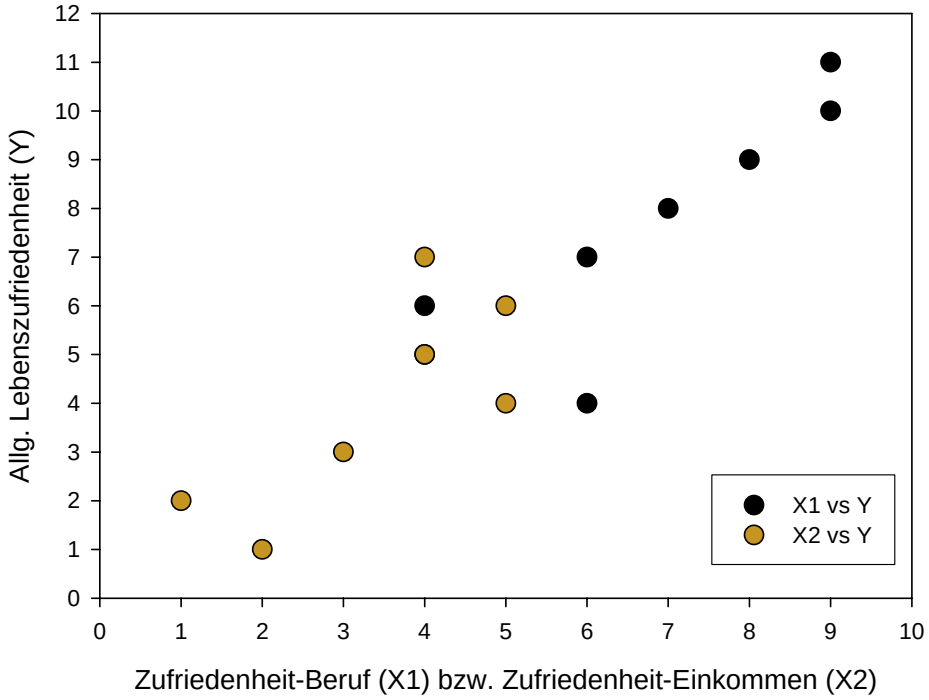
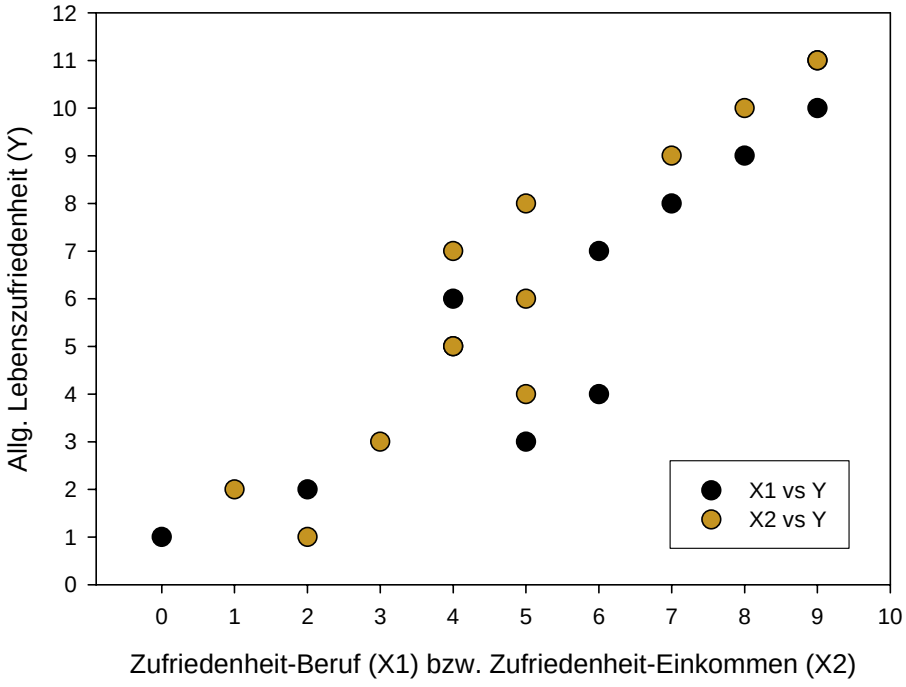


Abb. 9: Multiple Imputation am Datenbeispiel von Wirtz (2004)

EM-A: Datenbeispiel



Originaldaten



Daten ersetzt mit EM-Algorithmus

Abb. 9: EM-Algorithmus am Datenbeispiel von Wirtz (2004)

IV. Allgemeine Empfehlungen

Anwendungsempfehlungen (1)

- Grundregel:
  "Die einzige gute Lösung für das Problem fehlender Werte ist kein Problem zu haben!" (Allison,2001)
- Besser höheren Aufwand für einen möglichst vollständigen "schlanken" Datensatz als nach umfangreicher Datenerhebung nicht genau fassbare Verzerrungen durch fehlende Werte zu erhalten
- Erhebung von Kovariaten, die dem Fehlen von Daten korrelieren, ermöglichen Rückschlüsse auf die fehlenden Daten

Anwendungsempfehlungen (2)

- Mittelwertsersetzung sollte nicht zur Behandlung fehlender Werte eingesetzt werden (Ausnahme: Summen/Mittelwerte)
- LW-Fallausschluss und PW-Fallausschluss sollte nur unter der MCAR-Annahme, welche in der Realität selten vorliegt, angewendet werden
- *State-of-the-art*-Verfahren (EM-Algorithmus, Multiple Imputation) sind den konventionellen Verfahren überlegen
  - schwächere Voraussetzungen (MAR statt MCAR)
  - geringere Verzerrungen (Erhalt der Datenstruktur)
  - größere Effizienz (kein Fallausschluss)

Anwendungsempfehlungen (3)

- SPSS: Zusatzmodul "Missing Value Analysis (MVA)"
  - EM-Algorithmus
  - Möglichkeiten zur MD-Diagnose
- SAS: Prozeduren PROC MI und MIANALYZE
  - Multiple Imputation
  - noch experimentell bis Version 8.2
- Freeware-Programme:
  EMCOV [DOS] (Graham, 1995) und NORM [Windows] (Schafer, 2000) verwenden den EM-Algorithmus und können Multiple Imputation
- u.a.

Literatur

1. Wirtz, M. (2004). Über das Problem fehlender Werte: Wie der Einfluss fehlender Informationen auf Analyseergebnisse entdeckt und reduziert werden kann. *Rehabilitation*, 43, 109-115.
2. Müller, J. M. (2002). Umgang mit fehlenden Werten. In: Reusch, A., Zwingmann, Ch. & Faller, H. (Hrsg.). *Empfehlungen zum Umgang mit Daten in der Rehabilitationsforschung*. Regensburg: Roderer.
3. Schafer, J.L. & Graham, J.W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, 7(2),147-177.
4. Schafer, J.L. (2000). NORM 2.03 for Windows 95/98/NT [Software]. Quelle: http://www.stat.psu.edu/~jls
5. Graham, J. W. (1995). EMCOV for DOS [Software]. Quelle: http://methodology.psu.edu/downloads/EMCOV.html

Vielen Dank

für Ihre Aufmerksamkeit!

Kontakt: wilmar.igl@mail.uni-wuerzburg.de