

13. Rehabilitationswissenschaftliche Kolloquium –
Selbstkompetenz: Weg und Ziel der Rehabilitation
Düsseldorf, 8.-10.3.2004

Multiple Imputation

*Behandlung von fehlenden Werten
mit Multipler Imputation*

Dipl.-Psych. Wilmar Igl

Rehabilitationswissenschaftlicher
Forschungsverbund
Bayern

[Einleitung]


- **Multiple Imputation (MI):**
 - mehrfache Ersetzung (=Imputation) von fehlenden Werten durch $m > 1$ plausible Werte
 - Erweiterung von einfachen Imputationsmethoden (z.B. Mittelwert, Regression,...)
- MI als ***state of the art-Methode*** zur Behandlung von fehlenden Werten (neben *maximum-likelihood-Methode*) (vgl. Schafer & Graham, 2002)

[Vorteile von MI (I)]

- Nutzung der verfügbaren Information in beobachteten Daten
- Komfortable Auswertung von vollständigen Datensätzen möglich
- universeller Einsatz für verschiedenste Fragestellungen möglich
- Berücksichtigung der Unsicherheit aufgrund von fehlenden Werten

Vorteile von MI (II)

- **Hohe Effizienz** bei kleiner Anzahl von Imputationen
- Anwendung von MI mit **frei erhältlicher Software** (z.B. NORM, Schafer 2000) und **Standard-Statistik-Software** (z.B. SPSS, SAS) möglich



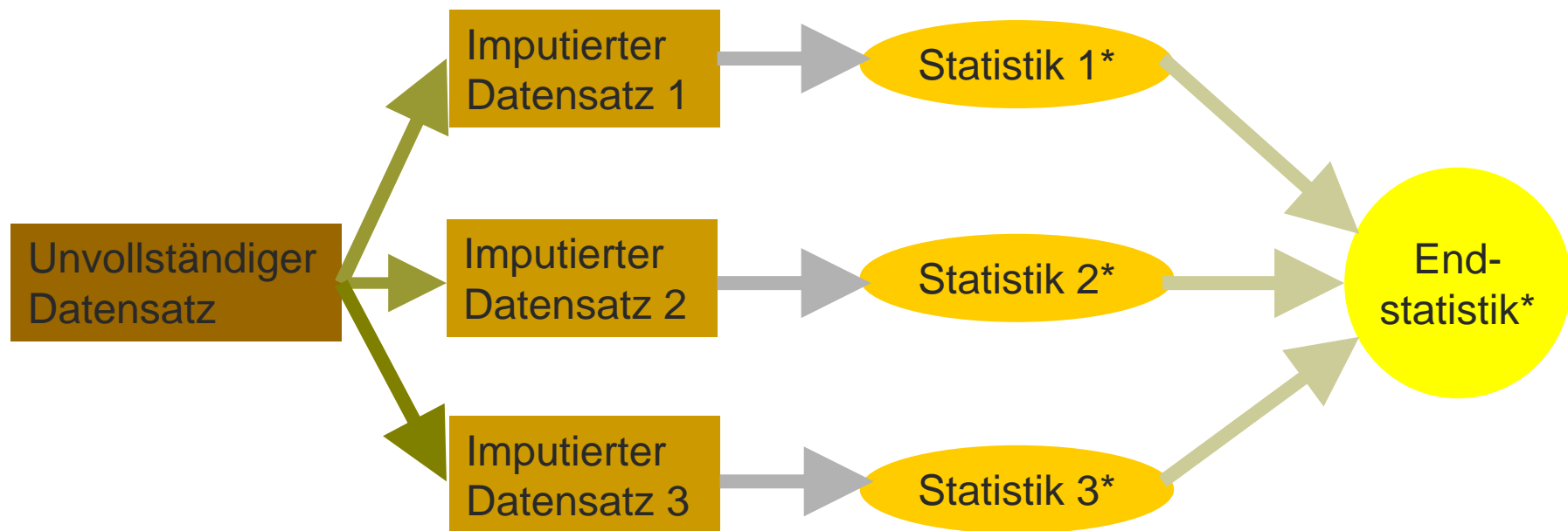
I.
Grundlagen von
Multipler Imputation

[MI: Die Grundidee]

1) IMPUTATION ⇒

2) ANALYSE ⇒

3) INTEGRATION



*Punktschätzer und Standardfehler

[1) Imputation: Datenbeispiel]

Beobachtete Daten				Imputationen			
				1	2	...	m
3	?	4	2	2	1	...	2
4	3	6	3				
2	1	?	2	4	4	...	5
4	2	6	4				
?	3	5	4	2	3	...	3
3	1	4	2				
2	2	5	?	3	3	...	4

1) Imputation: DA-Algorithmus

- Iterative „Ziehung“ von **Parametern** (z.B. $NV(\mu, \sigma)$) und **Variablenwerten** aus einer prädiktiven Bayes-Verteilung

$$P(Y_{mis}, \theta | Y_{obs})$$

- **Durchführung von k Schleifen** (z.B. 1000) bis der Algorithmus „konvergiert“ (bis zur Unabhängigkeit der Schätzungen)

[2) Analyse]

- Berechnung **statistischer Parameter** (Punktschätzer und ihre Standardfehler) mit Hilfe von Standard-Statistik-Software (SPSS, SAS,...)

Beispiele: Mittelwerte, Regressionskoeffizienten, Kovarianzen und Korrelationen, ...

- Berechnung der zugehörigen **Standardfehler (SE)** **notwendig**

[3) Integration (nach Rubin, 1987)]

- **MI Punktschätzer:**
Berechnung des arithmetischen Mittels der m Statistiken (z.B. Mittelwerte) aus m imputierten Datensätzen
- **Varianz (gesamt)** = Varianz (innerhalb der m Datensätze) + Varianz (zwischen den m Datensätzen)
- Berechnung von **weiteren Statistiken**, z. B. Freiheitsgrade, t-Werte, p-Werte, Konfidenzintervalle (95%)



II.
Umsetzung von
Multipler Imputation
mit NORM

[NORM 2.03 (Schafer, 2000)]

Wesentliche Komponenten:

- **EM-Algorithmus:**
Generierung von Startwerten für den DA-Algorithmus
- **DA-Algorithmus**
(Data Augmentation):
Ersetzen der fehlenden Werte durch m Imputationen
- **MI-Inferenz** (nach Rubin, 1987): Integration der m Datenanalysen

The screenshot shows the NORM software interface. The title bar reads 'NORM'. The menu bar includes 'File', 'Display', 'Series', 'Analyze', 'Window', and 'Help'. The main window title is 'NORM session: LQDaten'. Below the title bar, there are tabs for 'Data', 'EM algorithm', 'Data augmentation', and 'Impute from parameters'. Underneath, there are sub-tabs for 'Data file', 'Variables', and 'Summarize'. The main display area shows the following information:

File: D:\Daten\LQDaten.dat
No. of variables = 5 No. of cases = 30 Missing value code =

001	2	43	-9.00	56.34
002	2	50	-9.00	-9.00
003	2	44	-9.00	-9.00
004	2	38	42.48	-9.00
005	2	44	40.38	-9.00
006	2	42	36.70	-9.00
007	2	57	35.44	-9.00
008	2	42	37.43	-9.00
009	1	43	36.29	-9.00
010	2	42	33.10	-9.00

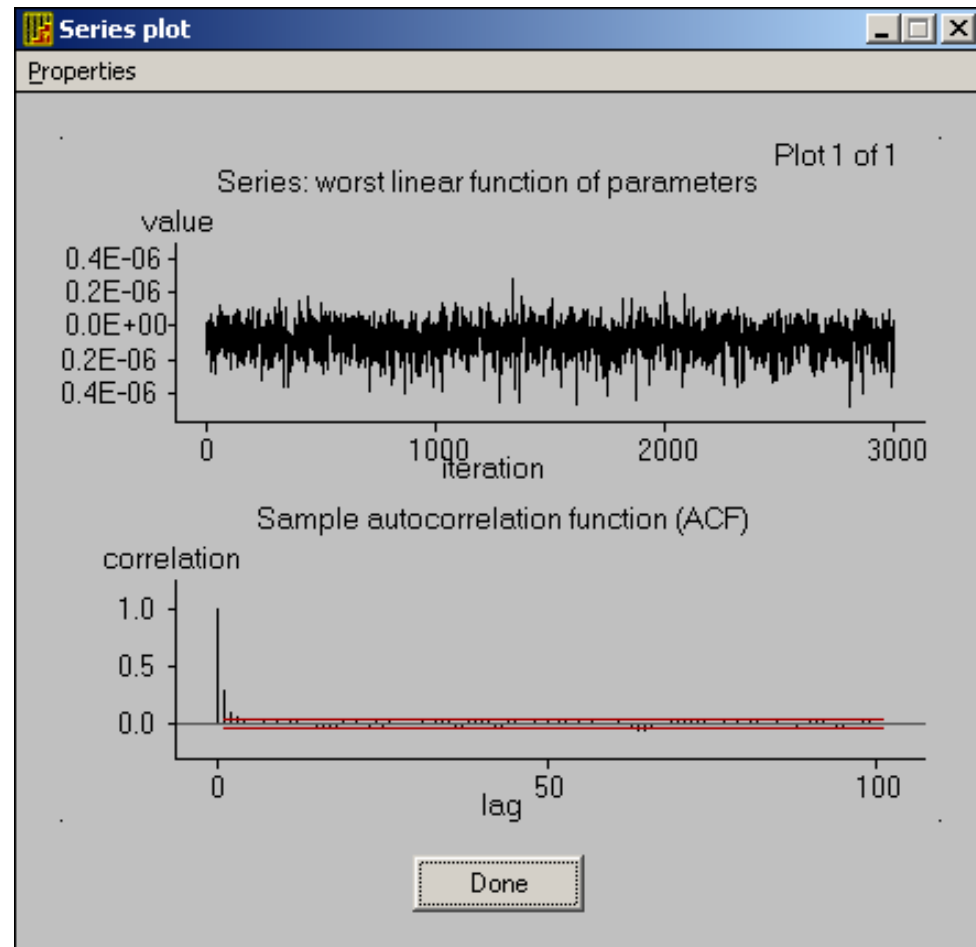
At the bottom of the window, a status bar indicates 'Session open.'

[Anwendung]

1. **EM-Algorithmus**: Startwerte und erforderliche Durchläufe k schätzen
2. **DA-Algorithmus**: Schätzung von m Imputationen mit $m \cdot k$ Durchläufen
3. Beurteilung der **Konvergenz**
4. **Analyse** der m Datensätze (Berechnung von m Punktschätzern und m Standardfehlern)
5. Berechnung der **MI-Inferenz**:
1 Punktschätzer mit Konfidenzintervall

[Beurteilung der Konvergenz]

- Worst Linear Function:
keine Trends
- AutoCorrelation Function (ACF):
keine bedeutsame Korrelation
($p < .05$)



[Datenbeispiel]

Vollständiger Datensatz

NORM - File: LQDaten_komplett.out

File Display Series Analyze Window Help

NUMBER OF OBSERVATIONS = 30
NUMBER OF VARIABLES = 4

	NUMBER MISSING	% MISSING
sex	0	0.00
alter	0	0.00
KSK	0	0.00
PSK	0	0.00

MEANS AND STANDARD DEVIATIONS OF OBSERVED DATA

	MEAN	ST.DEV.
sex	1.80000	0.406838
alter	18.8333	7.91369
KSK	36.7693	7.26314
PSK	45.1557	11.3721

Session open.

Datensatz mit Missings

NORM - File: Daten_mit_Missings.out

File Display Series Analyze Window Help

NUMBER OF OBSERVATIONS = 30
NUMBER OF VARIABLES = 4

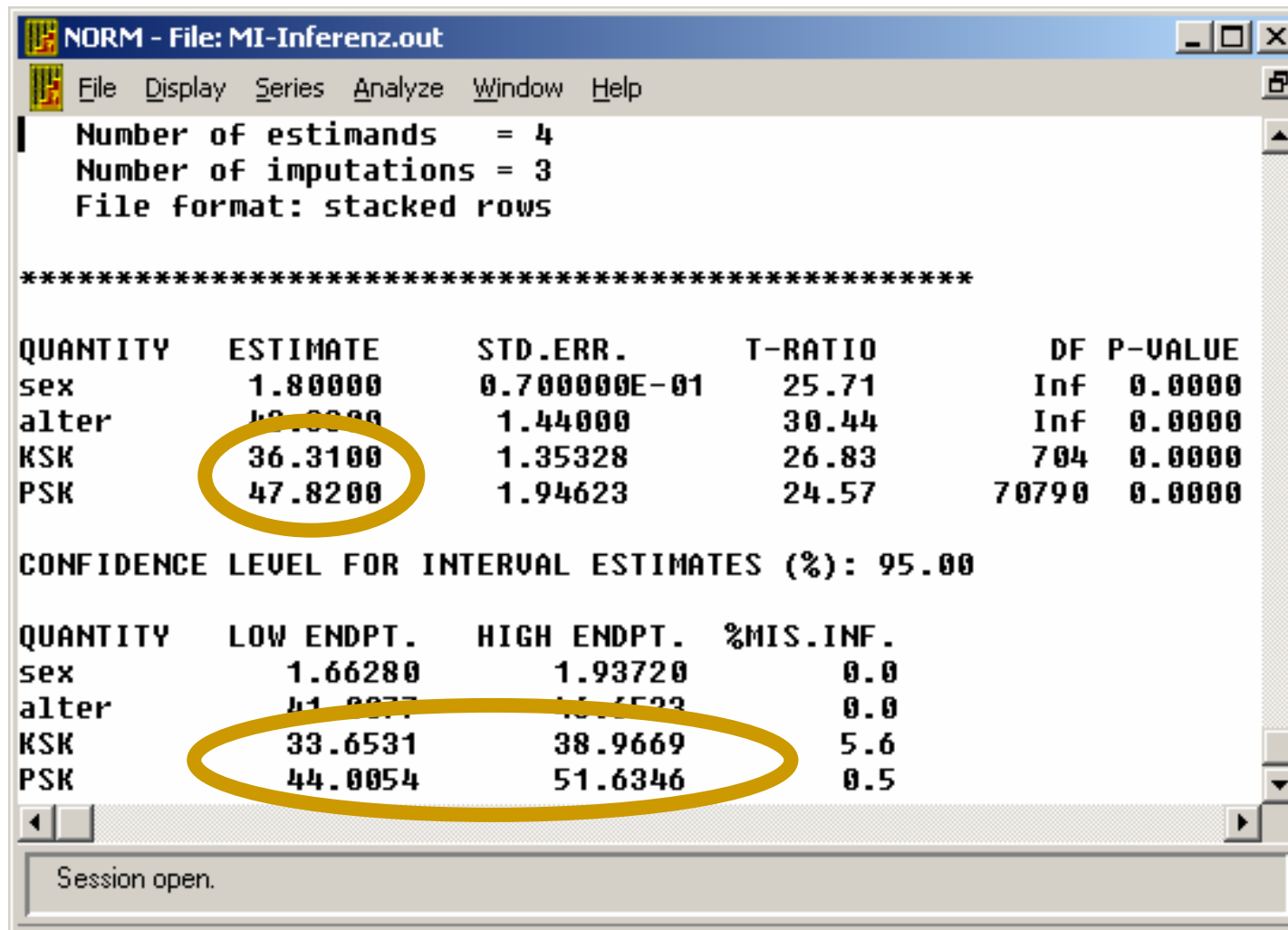
	NUMBER MISSING	% MISSING
sex	0	0.00
alter	0	0.00
KSK	3	10.00
PSK	10	33.33

MEANS AND STANDARD DEVIATIONS OF OBSERVED DATA

	MEAN	ST.DEV.
sex	1.80000	0.406838
alter	18.8333	7.91369
KSK	36.2678	7.45385
PSK	48.1350	10.6234

Session open.

Datenbeispiel: MI-Inferenz



Number of estimands = 4
Number of imputations = 3
File format: stacked rows

QUANTITY	ESTIMATE	STD.ERR.	T-RATIO	DF	P-VALUE
sex	1.80000	0.700000E-01	25.71	Inf	0.0000
alter	42.0000	1.44000	30.44	Inf	0.0000
KSK	36.3100	1.35328	26.83	704	0.0000
PSK	47.8200	1.94623	24.57	70790	0.0000

CONFIDENCE LEVEL FOR INTERVAL ESTIMATES (%): 95.00

QUANTITY	LOW ENDPT.	HIGH ENDPT.	%MIS.INF.
sex	1.66280	1.93720	0.0
alter	41.0077	43.0023	0.0
KSK	33.6531	38.9669	5.6
PSK	44.0054	51.6346	0.5

Session open.



III.

Anwendungsempfehlungen
für Multiple Imputation

[Anwendungsempfehlungen]

- **Multivariate Normalverteilung** der Daten
- Missing-Data-Mechanismus: **MAR**
- **Anzahl von Imputationen** abhängig vom Anteil fehlender Information (vgl. Tab. 1)
- **Imputations-Modell** sollte mit **Analyse-Modell** kompatibel sein (z.B. IM sollte auch die Variablen des AMs enthalten)
- **Beachte:** MI ist im Allgemeinen **robust** gegenüber Abweichungen von Voraussetzungen z.B. kein MAR, ungenaues Parametermodell

[Effizienz]

Imp m	Anteil fehlender Information (γ)				
	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	96	93	92
20	100	99	98	97	96

Tab. 1: Effizienz (in %) von MI abhängig von fehlender Information (γ) und Anzahl der Imputationen

[Anwendungsempfehlungen]

- **Multivariate Normalverteilung** der Daten
- Missing-Data-Mechanismus: **MAR**
- **Anzahl von Imputationen** abhängig vom Anteil fehlender Information (vgl. Tab. 1)
- **Imputations-Modell** sollte mit **Analyse-Modell** kompatibel sein (z.B. IM sollte auch die Variablen des AMs enthalten)
- **Beachte:** MI ist im Allgemeinen **robust** gegenüber Abweichungen von Voraussetzungen z.B. kein MAR, ungenaues Parametermodell

Literatur

1. Schafer, J.L. & Graham, J.W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, 7(2), 147-177
2. Sinharay, S., Stern, H.S. & Russell, D. (2001). The Use of Multiple Imputation for the Analysis of Missing Data. *Psychological Methods*, 6(4), 317-329.
3. Little, R.J.A. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2. Aufl.). Hoboken, NJ: Wiley
4. Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall
5. Schafer, J.L. (2000). *NORM 2.03 for Windows 95/98/NT* [Software]. Quelle: <http://www.stat.psu.edu/~jls>
6. <http://www.multiple-imputation.com>

A decorative graphic consisting of a horizontal line with a gradient from light green to white. A black left square bracket is positioned on the left side of the line, and a gold right square bracket is on the right side.

*Vielen Dank
für Ihre Aufmerksamkeit!*

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

[

Anhang

]

1) Imputation: DA-Algorithmus (II)

- I-Schritt: $Y_{mis}^{t+1} \sim P(Y_{mis} | Y_{obs}, \theta^{(t)})$
- P-Schritt: $\theta^{(t+1)} \sim P(\theta | Y_{obs}, Y_{mis}^{(t+1)})$
- Markov-Kette: $Y_{mis}^{(1)}, \theta^{(1)}, Y_{mis}^{(2)}, \theta^{(2)}, \dots, Y_{mis}^{(t)}, \theta^{(t)}$
- Endverteilung: $P(Y_{mis}, \theta | Y_{obs})$

[3) Integration (nach Rubin, 1987)]

- MI Punktschätzer: $\bar{Q} = \frac{1}{m} \sum_{i=1}^m \hat{Q}_i$
- Varianz(innerhalb): $\bar{U} = \frac{1}{m} \sum_{i=1}^m U_i$
- Varianz(zwischen): $B = \frac{1}{m-1} \sum_{i=1}^m (\hat{Q}_i - \bar{Q})^2$
- Varianz(gesamt): $T = \bar{U} + (1 + m^{-1})B$

[3) Integration (II)]

- MI Konfidenzintervall: $\bar{Q} \pm t_{df} \sqrt{T}$
- Freiheitsgrade (df): $df = (m - 1) \left(1 + \frac{m\bar{U}}{(m + 1)B} \right)^2$
- Anteil fehlender Information:
$$\gamma = \frac{r + 2 / (df + 3)}{r + 1}$$
$$r = (T - \bar{U}) / \bar{U}$$