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 Imputations-methoden (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)

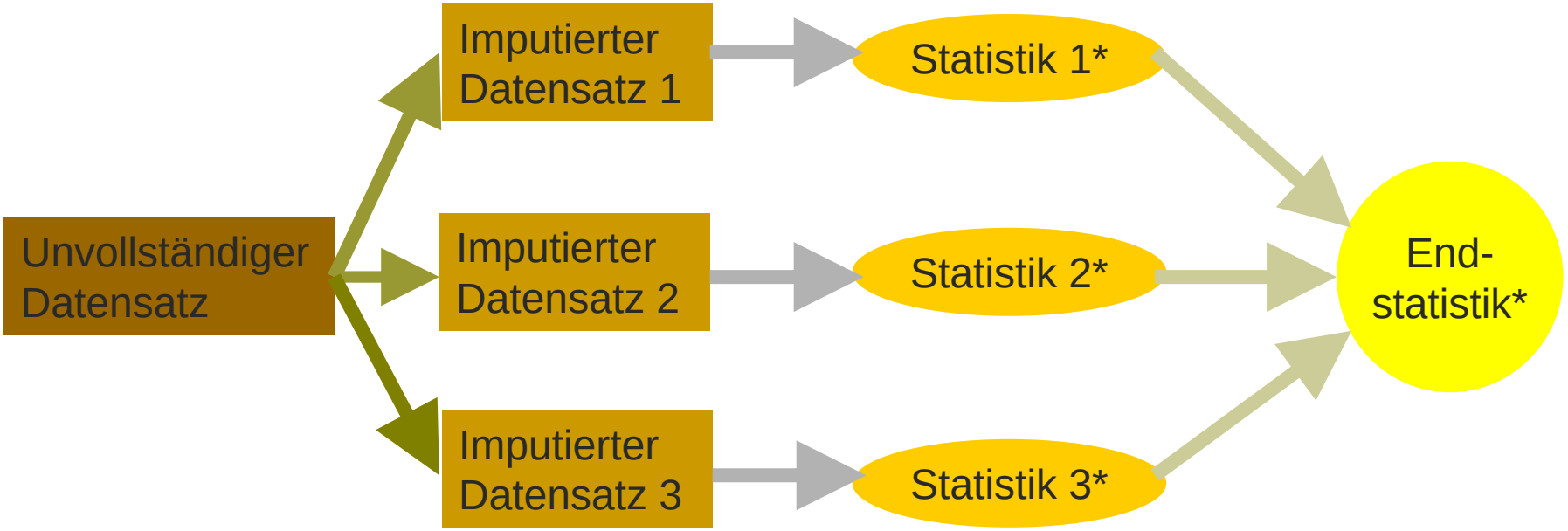
- Hohe Effizienz bei kleiner Anzahl von Imputationen
- Anwendung von MI mit frei erhältlicher Software (z.B. NORM, Schafer 2000) und Standard-Statistik-Software (z.B. SPSS, SAS) möglich

I.

Grundlagen von Multipler Imputation

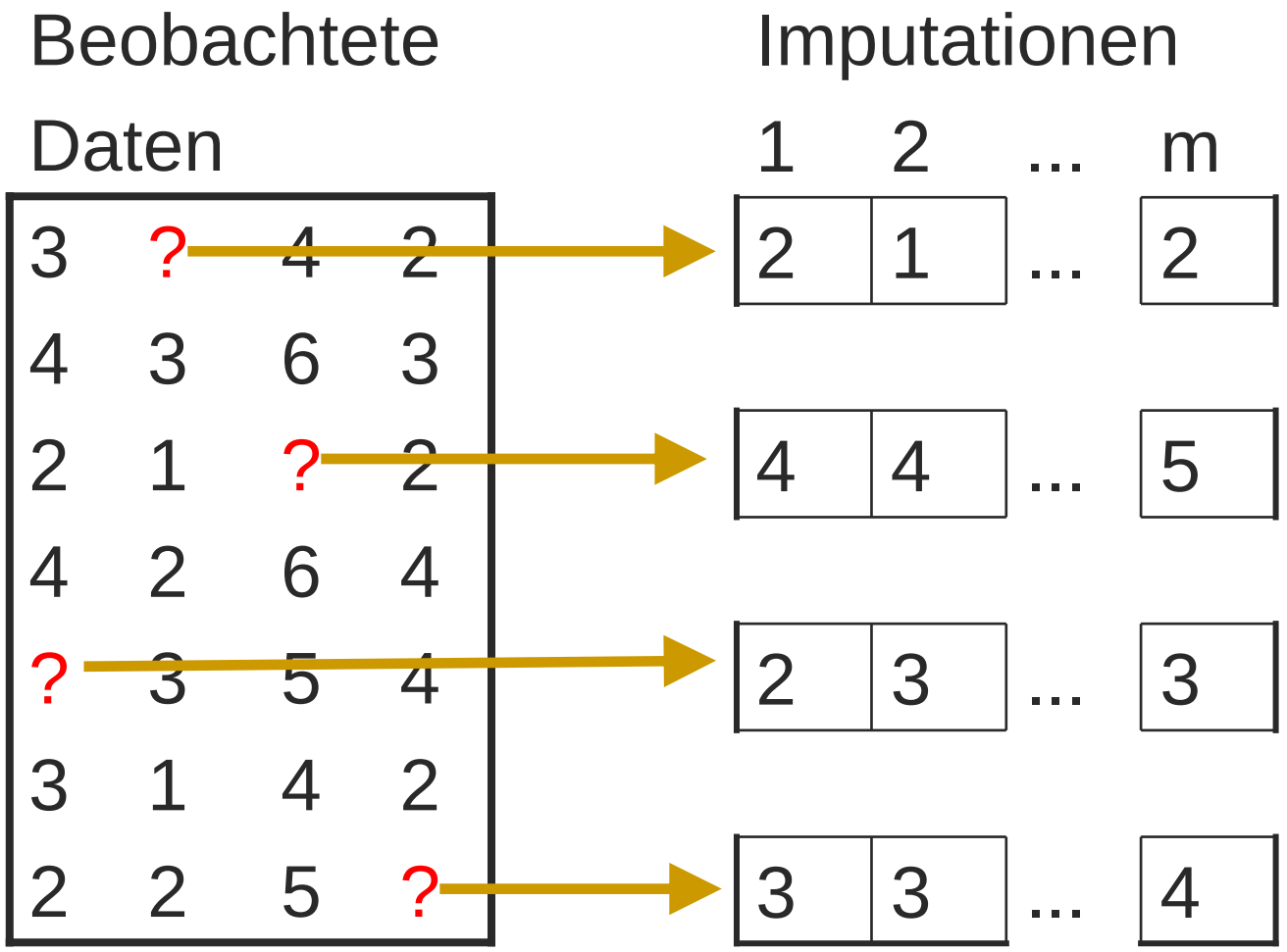
MI: Die Grundidee

1) IMPUTATION ⇨ 2) ANALYSE ⇨ 3) INTEGRATION



*Punktschätzer und Standardfehler

1) Imputation: Datenbeispiel



1) Imputation: DA-Algorithmus

- Iterative „Ziehung“ von Parametern (z.B. NV(μ, σ)) und Variablenwerten aus einer prädiktiven Bayes-Verteilung

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

- 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)

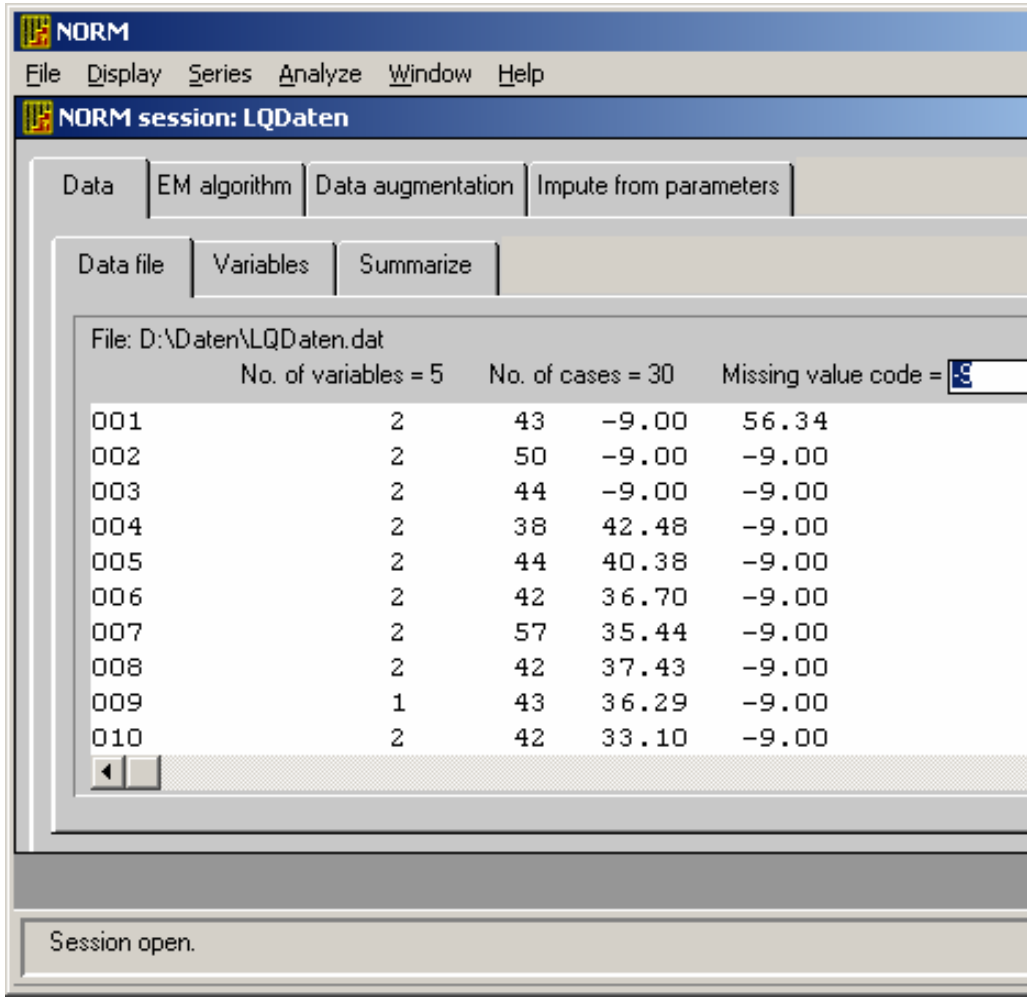
- Berechnung von weiteren Statistiken, z. B. Freiheitsgrade, t-Werte, p-Werte, Konfidenzintervalle (95%)

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



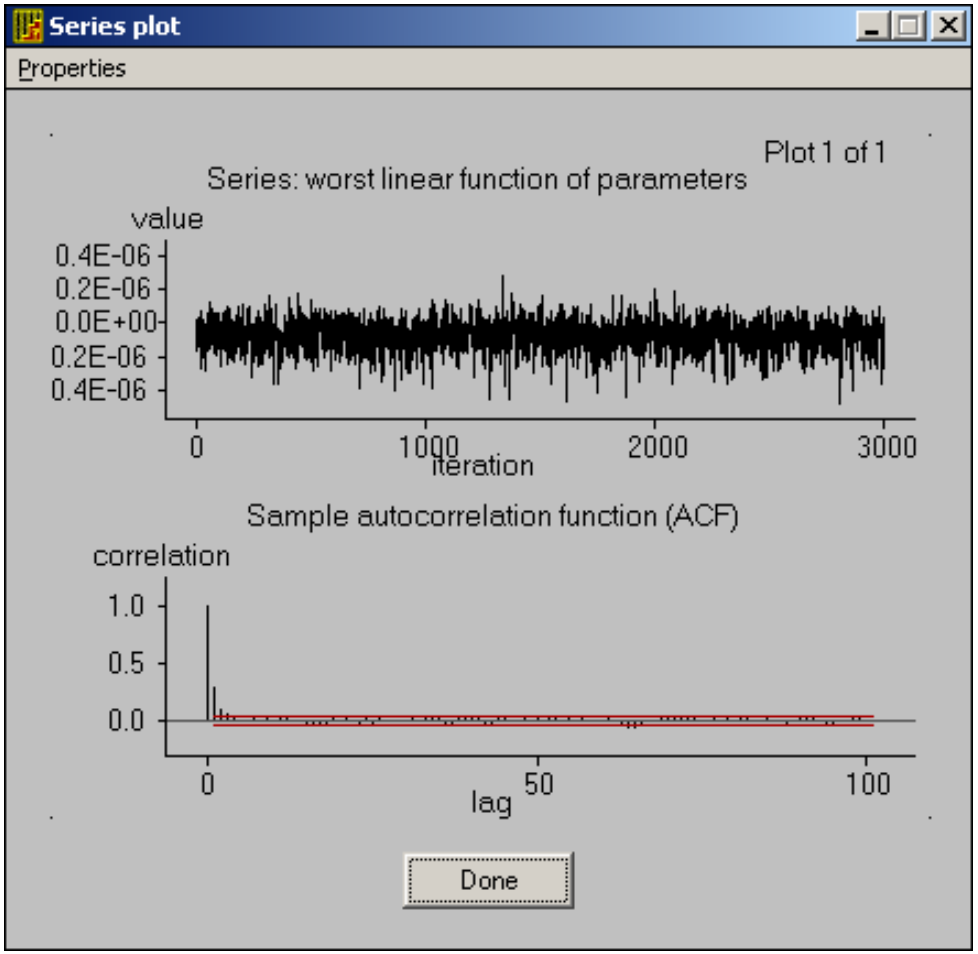
Anwendung

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

Beurteilung der Konvergenz

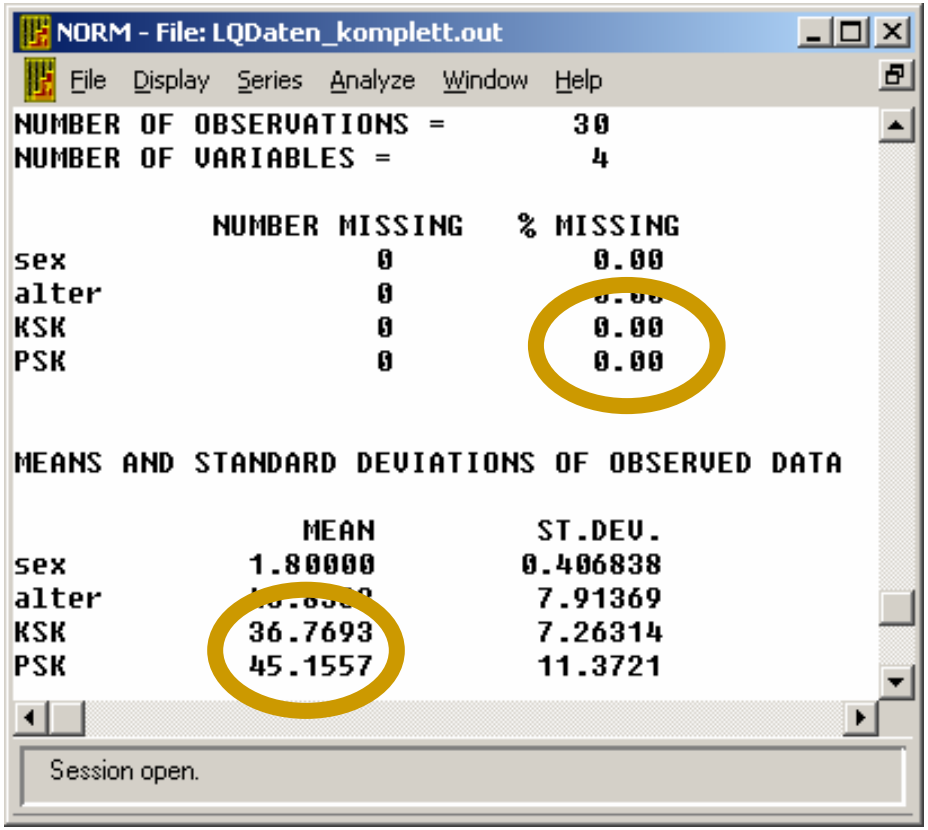
- Worst Linear Function:
  keine Trends

- AutoCorrelation Function (ACF):
  keine bedeutsame Korrelation
  (p<.05)

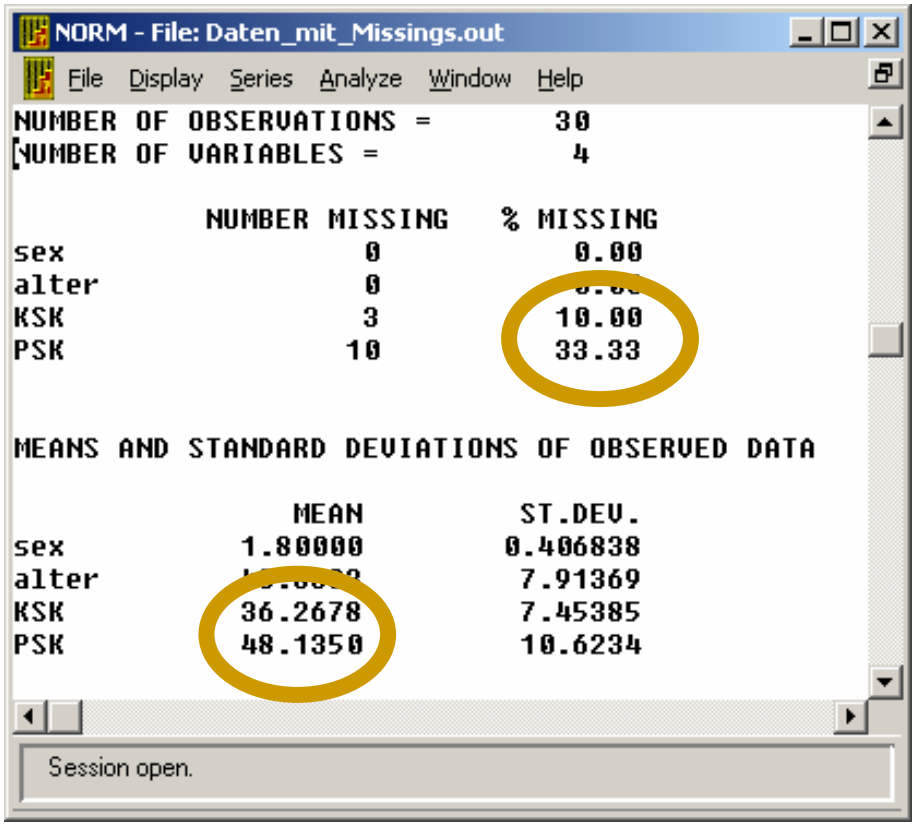


Datenbeispiel

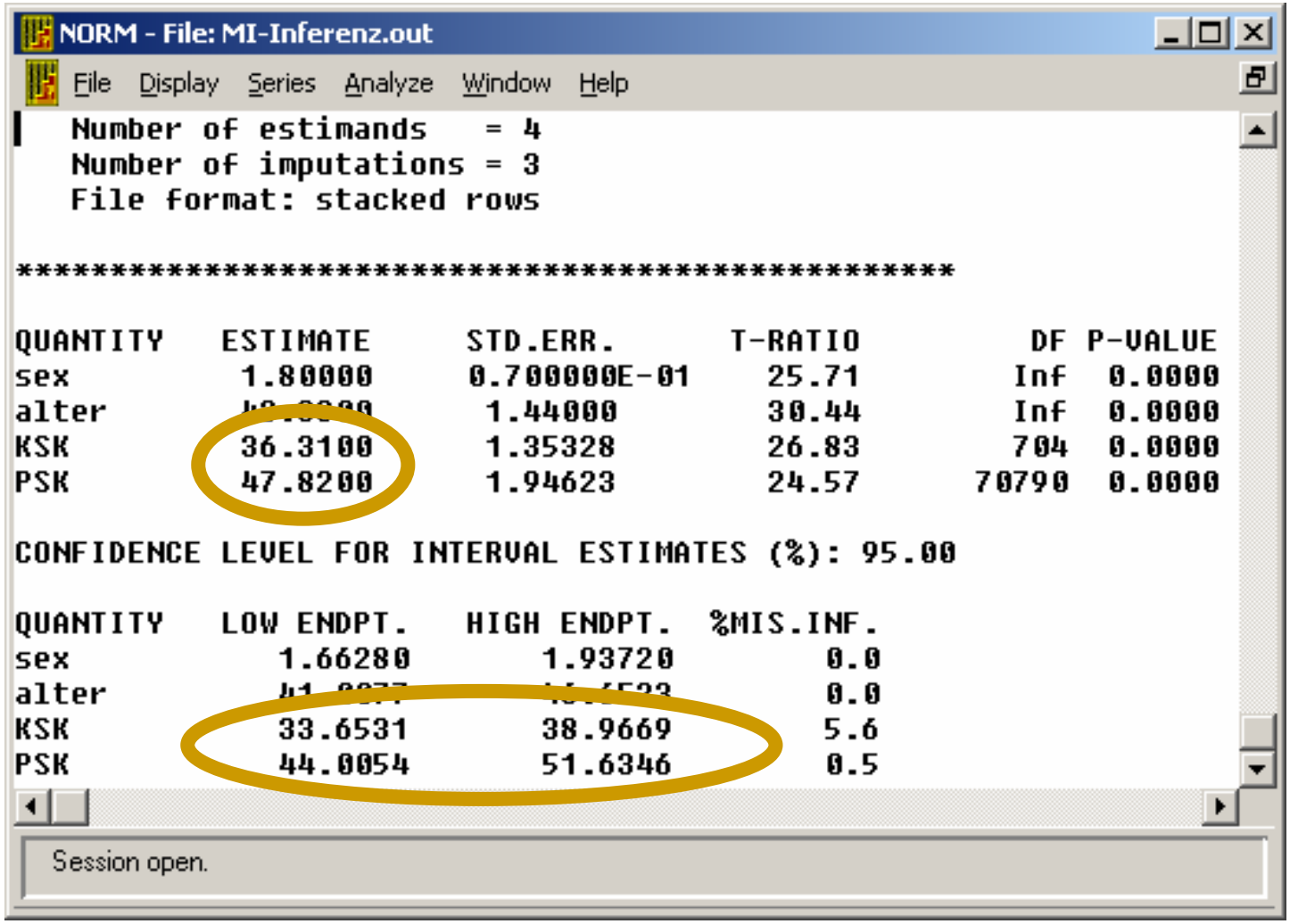
Vollständiger Datensatz



Datensatz mit Missings



Datenbeispiel: MI-Inferenz



III.
Anwendungsempfehlungen für Multiple Imputation

Anwendungsempfehlungen

- Multivariate Normalverteilung der Daten
- Missing-Data-Mechanimus: 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	Anteil fehlender Information (γ)				
m	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-Mechanimus: 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

Vielen Dank

für Ihre Aufmerksamkeit!

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

Anhang

1) Imputation: DA-Algorithmus (II)

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

3) Integration (nach Rubin, 1987)

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

3) Integration (II)

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

$$r = (T - \overline{U})/\overline{U}$$