Analyse von Large-Scale-Assessment-Daten und Resampling-Methoden mit BIFIEsurvey

Konrad Oberwimmer

BIFIE - Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens

2. Salzburger Methoden-Workshop des BIFIE

Gliederung

1. Motivation
2. Resampling $\rightarrow$*Übung*
3. *PAUSE*
4. Multiple Imputation, Plausible Values $\rightarrow$*Übung*
5. BIFIEsurvey advanced $\rightarrow$*Übung*

Allg. Merkmale von LSA-Daten

- *large* (engl.) = groß
  - PISA bzw. TIMSS/PIRLS: mehr als 500.000 Fälle, mehr als 4.000 pro Land
  - BIST-Überprüfungen: Kohorten von 80.000 Schüler/innen
- "komplexe" Stichproben (Stratifizierung, Cluster, hierarchisches Vorgehen)
  - unterschiedliche Ziehungswahrscheinlichkeit für Einheiten → mangelnde Repräsentativität ohne **Gewichtung**
  - unterschätzte Populationsvarianz → Notwendigkeit zum **Resampling**
- Leistungsschätzung durch reflexive Messmodelle bei begrenzter Anzahl von Aufgaben
  → Notwendigkeit zur **multiplen Imputation (PV-Schätzung)**

Allg. Merkmale von LSA-Daten

- *large* (engl.) = groß
  - PISA bzw. TIMSS/PIRLS: mehr als 500.000 Fälle, mehr als 4.000 pro Land
  - BIST-Überprüfungen: Kohorten von 80.000 Schüler/innen
- "komplexe" Stichproben (Stratifizierung, Cluster, hierarchisches Vorgehen)
  - unterschiedliche Ziehungswahrscheinlichkeit für Einheiten → mangelnde Repräsentativität ohne **Gewichtung**
  - unterschätzte Populationsvarianz → Notwendigkeit zum **Resampling**
- Leistungsschätzung durch reflexive Messmodelle bei begrenzter Anzahl von Aufgaben
  → Notwendigkeit zur **multiplen Imputation (PV-Schätzung)**

Allg. Merkmale von LSA-Daten

- *large* (engl.) = groß
  - PISA bzw. TIMSS/PIRLS: mehr als 500.000 Fälle, mehr als 4.000 pro Land
  - BIST-Überprüfungen: Kohorten von 80.000 Schüler/innen
- "komplexe" Stichproben (Stratifizierung, Cluster, hierarchisches Vorgehen)
  - unterschiedliche Ziehungswahrscheinlichkeit für Einheiten → mangelnde Repräsentativität ohne **Gewichtung**
  - unterschätzte Populationsvarianz → Notwendigkeit zum **Resampling**
- Leistungsschätzung durch reflexive Messmodelle bei begrenzter Anzahl von Aufgaben
  → Notwendigkeit zur **multiplen Imputation (PV-Schätzung)**

Gewichtung

Jeder Fall i (meist innerhalb einer übergeordneten Einheit j) hat bei komplexen Stichproben eine Ziehungwahrscheinlichkeit p_{ij}. Sein Gewicht im Rahmen von Analysen ist dann bestimmt als

$$w_{ij} = \frac{1}{p_{ij}}$$

Bspw. in der Berechnung eines Mittelwerts:

$$\bar{x} = \frac{\sum_{ij} x_{ij} \cdot w_{ij}}{\sum_{ij} w_{ij}}$$

Dadurch wird erhöhte Repräsentativität erreicht.

Gewichtung

Kommen mehrere Gewichtungsvorgänge zum Tragen (bspw. zuerst Schulen gezogen, dann *non-response-adjustment* bei Schüler/innen), spricht man im Endeffekt vom *Final Weight*.

Gewichte können linear skaliert werden:

- *Total Weights:* Summe der Gewichte entspricht Populationsgröße
- *Senate Weights:* Summe der Gewichte innerhalb von Subgruppen, bspw. Länder, wird konstant gehalten
- *House Weights*: Summe der Gewichte entspricht Stichprobengröße (Umgewichtung)

LSA mit SPSS

Grundproblem: SPSS-Outputs langsam dargestellt und nicht programmatisch weiter verarbeitbare Objekte. Daher SPSS-Makros:

- OECD PISA Makros (http://www.oecd-ilibrary.org)
  - begrenzter Funktionsumfang für PISA-Berichterstattung
  - Verwendung durch SPSS-Syntax
- IDB-Analyzer (http://www.iea.nl/data.html)
  - etwas größerer Funktionsumfang (logistische Regression)
  - Verwendung durch GUI, die letztlich SPSS-Syntax erzeugt
- BIFIE SPSS Makros:
  - sehr begrenzter Funktionsumfang
  - Verwendung durch praktische SPSS-Syntax
  - Einfach anwendbar auf OECD-, IEA- und BIFIE-Datensätze

LSA mit diversen Statistikprogrammen

- MPlus: TOP in Hinblick auf Möglichkeiten, aber geringe *usability*
- SAS: imperative Programmiersprache, somit praktisch alles möglich; wenigstens für PISA fertige Makros durch OECD
- Stata: beherrscht Resampling und Multiple Imputation mit Verwendung aus GUI heraus

Ein Wort zu SPSS

SPSS ist ein brauchbares Statistik-Programm, bei:

- relativ wenigen, klar definierten Deskriptoren/Hypothesentests
- einmaliger Berechnung
- auf einfachen, unveränderlichen Datentabellen

Auch wenn SPSS langsam ist, R läuft eher in Speicherprobleme.

R kann's, halt ohne GUI

Mit den Paketen *survey*, *lavaan.survey* und *mitools* können LSA-Daten korrekt behandelt werden. Ebenso für Spezialfälle *intsvy* und *svyPVpack*.

Wieso ein neues R-Paket?

- uneinheitliche Syntax $\rightarrow$ schwieriger Einstieg
- tlw. beschränktes Set statistischer Verfahren
- ineffizient in der Berechnung (aber besser als SPSS)

Zielsetzung

BIFIEsurvey errechnet nach den im LSA etablierten Konventionen korrekte Kennwerte und Standardfehler.

Dabei ist es effizient (durch Einbindung von kompiliertem C-Code), die Geschwindigkeitsvorteile gegenüber SPSS-Makros liegen bis 1:100, aber auch deutlich schneller als *survey*.

Und auf beliebige LSA-Daten, aber auch sonstige Daten mit Resampling oder multipler Imputation anwendbar. Nur ein Befehl zum Erfassen der Datenstruktur!

Bundesinstitut bifie

Umfang (v2.0-7)

- Univariate Deskription: Häufigkeiten, Perzentile, Mittelwert/Standardabweichung
- Bivariate Deskription: Kreuztabulierung, Korrelation
- Multivariate Deskription: Missing Value Analysis, Auswertung nach (mehreren) gruppierenden Variablen
- Wald-Tests $\rightarrow$ analog: t-Test, ANOVA, Kontraste
- Lineare und logistische Regression
- Pfadmodelle
- Zwei-Ebenen-Regression

Durch eine einfache Funktion ansatzweise benutzer-definiert erweiterbar.

Bundesinstitut bifie

Klassische frequentistische Inferenz

Grundannahme: Eine unbekannte Wahrscheinlichkeitsverteilung F führt zu beobachtbaren Werten $\mathscr{X}$.

$$F \rightarrow \mathscr{X}$$

Gesucht wird eine Eigenschaft θ der zugrunde liegenden Verteilung F, die sich nach einem bestimmen Algorithmus $t(\cdot)$ aus den beobachtbaren Werten berechnen lässt.

$$\theta = t(\mathscr{X})$$

Praktisch gibt es aber nur eine Realisation x aus möglichen konkreten X, sodass nur der empirsche Kennwert $\hat{\theta} = t(x)$ zur Verfügung steht.

Klassische frequentistische Inferenz

Frage: Wie gut ist $\hat{\theta}$ zur Vorhersage von θ geeignet?

Idee: $\hat{\theta}$ ist eine Realisation von $\hat{\Theta} = t(X)$, wobei X ein (wiederholt) unabhängig gezogenes, theoretisches Sample aus F ist.

Antwort: Die Genauigkeit von $\hat{\theta}$ entspricht der theoretischen Genauigkeit der $\hat{\Theta}$ als Schätzer von θ, gemessen als deren Streuung und bezeichnet als Standardfehler $se_{\hat{\Theta}}(\theta)$.

Alltagssprachlich

"Wie sehr würde ein Kennwert streuen, wenn immer wieder Stichproben (gleicher Größe) gezogen würden?"

Klassische frequentistische Inferenz

Scheinbar wurde das Problem bloß verschoben: Wie streut $\hat{\Theta}$? F ist nach wie vor unbekannt ...

- Plug-in-Prinzip: Bestimmte Eigenschaften von F werden aus den beobachteten Daten x geschätzt, womit $se_{\hat{\Theta}}(\theta)$ selbst zu einer Schätzung wird. Bspw.

$$se_{\hat{\Theta}}(\bar{\mathscr{X}}) = \sqrt{\frac{var(\mathscr{X})}{n}} \text{ wobei } var(\mathscr{X}) \sim \hat{var} = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$

- Pivot-Größen: Umrechnung von interessierenden Merkmalen auf - unter bestimmten Annahmen - bekannte Verteilungen von Teststatistiken (t, χ^2 etc.), die von störenden Parametern unabhängig sind $\rightarrow$ Nullhypothesentests (t-Test etc.)

Klassische frequentistische Inferenz

Scheinbar wurde das Problem bloß verschoben: Wie streut $\hat{\Theta}$? F ist nach wie vor unbekannt ...

- Plug-in-Prinzip: Bestimmte Eigenschaften von F werden aus den beobachteten Daten x geschätzt, womit $se_{\hat{\Theta}}(\theta)$ selbst zu einer Schätzung wird. Bspw.

$$se_{\hat{\Theta}}(\bar{\mathscr{X}}) = \sqrt{\frac{var(\mathscr{X})}{n}} \text{ wobei } var(\mathscr{X}) \sim \hat{var} = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$

- Pivot-Größen: Umrechnung von interessierenden Merkmalen auf - unter bestimmten Annahmen - bekannte Verteilungen von Teststatistiken (t, χ^2 etc.), die von störenden Parametern unabhängig sind $\rightarrow$ Nullhypothesentests (t-Test etc.)

Inferenz durch Resampling

Resampling-Methoden setzen die frequentistische Idee ($\hat{\theta}$ ist eine Realisation von $\hat{\Theta} = t(X)$) algorithmisch statt probabilistisch um.

Die unbekannte Wahrscheinlichkeitsverteilung F wird aus der empirischen Wahrscheinlichkeitsverteilung $\hat{F}$ geschätzt, die sich ergibt, wenn mehrfach Sub-Samples aus dem vorhandenen Datenmaterial gezogen werden. Für jede Replikation gilt:

$$\hat{F} \xrightarrow{iid} x^* \xrightarrow{t} \hat{\theta}^*$$

Es kann gezeigt werden, dass $\hat{F}$ die ML-Schätzung für F gegeben x ist. Damit nähert sich $sd(\hat{\theta}^*)$ asymptotisch dem $se_{\hat{\Theta}}(\theta)$.

Inferenz durch Resampling

Praktisch werden N Kennwerte $\hat{\theta}^1 \ldots \hat{\theta}^N$ berechnet, die jeweils aus einem Sub-Sample bzw. einem umgewichteten Sample stammen. Deren Abweichungsquadrate vom Punktschätzer $\hat{\theta}$ (oder vom Mittelwert aus $\hat{\theta}^1 \ldots \hat{\theta}^N$) sind die Basis für die Berechnung der Streuung, welche als Schätzung für den Standardfehler dient:

$$sd(\hat{\theta}^*) = \sqrt{A \cdot \sum_j (\hat{\theta}^j - \hat{\theta})^2} = \hat{se}(\theta)$$

Die Konstante A ist ein skalierender Faktor, um die Streuung in den numerischen Bereich zu transformieren, in dem sich auch der probabilistische $se_{\hat{\Theta}}(\theta)$ bewegt.

Die konkrete Umsetzung unterscheidet sich in Hinblick auf den Algorithmus zur Generierung von Sub-Samples.

Resampling vs. klassisch

- Resampling ist non-parametrisch, es gibt keine Verteilungsannahme zu F.
- Resampling ist im Prinzip generisch: Ein grundlegender Algorithmus gilt für beliebige Kennwerte θ. Verschiedene Varianten haben allerdings unterschiedliche Performanz bei bestimmten Kennwerten.
- Resampling tendiert zu konservativeren Abschätzungen des Standardfehlers.
- Im Resampling kann durch Umgewichtung Verletzungen der Annahme $F \overset{iid}{\rightarrow} x$ direkt begegnet werden. $\Rightarrow$ Bedeutung im LSA
- Resampling stellt wesentlich höhere Anforderungen an die Rechenleistung. $\Rightarrow$ Bedeutung des heutigen Workshops

Resampling vs. klassisch

- Resampling ist non-parametrisch, es gibt keine Verteilungsannahme zu F.
- Resampling ist im Prinzip generisch: Ein grundlegender Algorithmus gilt für beliebige Kennwerte θ. Verschiedene Varianten haben allerdings unterschiedliche Performanz bei bestimmten Kennwerten.
- Resampling tendiert zu konservativeren Abschätzungen des Standardfehlers.
- Im Resampling kann durch Umgewichtung Verletzungen der Annahme $F \overset{iid}{\rightarrow} x$ direkt begegnet werden. $\Rightarrow$ Bedeutung im LSA
- Resampling stellt wesentlich höhere Anforderungen an die Rechenleistung. $\Rightarrow$ Bedeutung des heutigen Workshops

Resampling vs. klassisch

- Resampling ist non-parametrisch, es gibt keine Verteilungsannahme zu F.
- Resampling ist im Prinzip generisch: Ein grundlegender Algorithmus gilt für beliebige Kennwerte θ. Verschiedene Varianten haben allerdings unterschiedliche Performanz bei bestimmten Kennwerten.
- Resampling tendiert zu konservativeren Abschätzungen des Standardfehlers.
- Im Resampling kann durch Umgewichtung Verletzungen der Annahme $F \overset{iid}{\rightarrow} x$ direkt begegnet werden. $\Rightarrow$ Bedeutung im LSA
- Resampling stellt wesentlich höhere Anforderungen an die Rechenleistung. $\Rightarrow$ Bedeutung des heutigen Workshops

Resampling vs. klassisch

- Resampling ist non-parametrisch, es gibt keine Verteilungsannahme zu F.
- Resampling ist im Prinzip generisch: Ein grundlegender Algorithmus gilt für beliebige Kennwerte θ. Verschiedene Varianten haben allerdings unterschiedliche Performanz bei bestimmten Kennwerten.
- Resampling tendiert zu konservativeren Abschätzungen des Standardfehlers.
- Im Resampling kann durch Umgewichtung Verletzungen der Annahme $F \overset{iid}{\rightarrow} x$ direkt begegnet werden. $\Rightarrow$ Bedeutung im LSA
- Resampling stellt wesentlich höhere Anforderungen an die Rechenleistung. $\Rightarrow$ Bedeutung des heutigen Workshops

Resampling vs. klassisch

- Resampling ist non-parametrisch, es gibt keine Verteilungsannahme zu F.
- Resampling ist im Prinzip generisch: Ein grundlegender Algorithmus gilt für beliebige Kennwerte θ. Verschiedene Varianten haben allerdings unterschiedliche Performanz bei bestimmten Kennwerten.
- Resampling tendiert zu konservativeren Abschätzungen des Standardfehlers.
- Im Resampling kann durch Umgewichtung Verletzungen der Annahme $F \overset{iid}{\rightarrow} x$ direkt begegnet werden. $\Rightarrow$ Bedeutung im LSA
- Resampling stellt wesentlich höhere Anforderungen an die Rechenleistung. $\Rightarrow$ Bedeutung des heutigen Workshops

Jackknife

The Jackknife replicates for unstratified two-stage sample designs

Replicate	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
School 1	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
School 2	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
School 3	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11
School 4	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11
School 5	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11
School 6	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11
School 7	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11
School 8	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11
School 9	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11
School 10	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00

- Jede der *n primary sampling units (PSUs)* wird einmal ausgeschlossen ($\hat{w}_i = 0$, die anderen $\hat{w}_j = w_j \cdot \frac{n}{n-1}$).
- Skalierender Faktor $A = 1$.
- Nachteile: Anzahl der Replikationen von n abhängig. Inkonsistent bei nicht-stetig definierten Kennwerten (bspw. Median).

Bootstrap

Schule	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	2	1	0	0	1	0	0	2	1	0
2	0	0	0	1	2	0	0	1	0	0
3	0	2	1	1	1	1	0	0	2	1
4	0	1	3	4	0	2	2	1	1	2
5	2	0	2	0	1	2	1	1	2	1
6	0	2	2	1	1	1	0	2	0	1
7	1	1	0	0	0	0	1	1	0	1
8	4	2	1	0	2	1	0	0	0	4
9	2	0	1	1	2	2	0	1	1	1
10	1	1	0	1	1	2	2	2	1	0
11	1	1	1	0	0	2	0	1	1	0
12	0	2	2	1	2	0	0	1	2	1
13	0	0	3	1	2	0	5	0	3	1
14	2	0	0	2	2	1	2	1	1	2
15	1	2	1	1	0	3	2	0	1	0
16	1	1	0	2	1	0	0	2	0	1
17	1	1	1	1	1	1	0	1	0	1
18	1	1	1	0	0	1	0	0	2	1
19	0	1	1	2	1	1	3	0	2	2
20	1	1	0	1	0	0	2	3	0	0

- Mehrfache (B mal) Ziehung von PSUs mit Zurücklegen.
- Skalierender Faktor $A = \frac{1}{B-1}$.
- Gute Performance ab $B > 200$, für Bootstrap-Konfidenzintervalle $B > 2.000$.

Balanced Repeated Replicates (BRR)

The Fay replicates

Pseudo-stratum	School	R1	R2	R3	R4	R5	R6	R7	R8	R9	R 10	R 11	R 12
1	1	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5
1	2	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5
2	3	1.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5
2	4	0.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5
3	5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5
3	6	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5
4	7	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5
4	8	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5
5	9	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5
5	10	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5
6	11	1.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5
6	12	0.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5
7	13	1.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5
7	14	0.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5
8	15	1.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5
8	16	0.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5
9	17	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5
9	18	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5
10	19	1.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	1.5
10	20	0.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	0.5

- Unterteilung der PSUs in Pseudo-Strata, Gewichtung nach einem Schema (Hadamard-Matrix).
- PISA: 80 Pseudo-Strata, Fay-Faktor von 0,5 $\rightarrow$ $A = \frac{1}{80 \cdot 0{,}5^2} = 0{,}05$
- Gute Performance bei relativ wenigen Replikationen, aber nicht universell einsetzbar.

Jackknife Repeated Replicates (JRR)

The Jackknife replicates for stratified two-stage sample designs

Pseudo-stratum	School	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	1	2	1	1	1	1	1	1	1	1	1
1	2	0	1	1	1	1	1	1	1	1	1
2	3	1	0	1	1	1	1	1	1	1	1
2	4	1	2	1	1	1	1	1	1	1	1
3	5	1	1	2	1	1	1	1	1	1	1
3	6	1	1	0	1	1	1	1	1	1	1
4	7	1	1	1	0	1	1	1	1	1	1
4	8	1	1	1	2	1	1	1	1	1	1
5	9	1	1	1	1	2	1	1	1	1	1
5	10	1	1	1	1	0	1	1	1	1	1
6	11	1	1	1	1	1	2	1	1	1	1
6	12	1	1	1	1	1	0	1	1	1	1
7	13	1	1	1	1	1	1	0	1	1	1
7	14	1	1	1	1	1	1	2	1	1	1
8	15	1	1	1	1	1	1	1	0	1	1
8	16	1	1	1	1	1	1	1	2	1	1
9	17	1	1	1	1	1	1	1	1	0	1
9	18	1	1	1	1	1	1	1	1	2	1
10	19	1	1	1	1	1	1	1	1	1	0
10	20	1	1	1	1	1	1	1	1	1	2

- Pseudo-Strata, hier "Jackknife-Zonen", wie in BRR, aber mit Ein-/Ausschluss von PSUs durch Gewichte 0/2.
- TIMSS/PIRLS/BIST: Zuordnung zu Jackknife-Zonen unter Beachtung expliziter und impliziter Strata. Skalierender Faktor $A = 1$.
- Vorteil: einfache Umsetzbarkeit, weniger Replikationen als bei Jackknife/Boostrap. Nachteil: Inkonsistent bei nicht-stetig definierten Kennwerten (bspw. Median).

Kurzanleitung zum Übungsteil

1. Bitte Übungs-Daten in einem "gut erreichbaren" Verzeichnis bereithalten.
2. "Aufgabenstellungen" direkt in *BIFIEsurvey_Uebungen.R*, bitte in R-Studio öffnen.
3. "Lösungen" stehen jederzeit in *BIFIEsurvey_Uebungen_Musterloesung.R* bereit, kann jederzeit konsultiert werden.
4. Danke an Lisa Mayrhofer für die Unterstützung!

Fehlende Werte

Fehlende Werte begegnen uns in zwei Formen:

1. als ungeplanter Ausfall: Item- oder Unit-Non-Response
   Dieser Ausfall ist selten *missing-at-random (MAR)*, sondern meist systematisch. Nichts zu tun heißt in dem Fall, die empirischen Gegebenheiten naiv fortzuschreiben.

1. als theoretisch erwarteter Ausfall:
   Werden latente Konstrukte mittels reflexiven Messmodellen erhoben, können meist nicht alle Items vorgegeben werden. Es mangelt an:
   1. Inhaltsvalidität
   2. Reliabilität
   3. stetigen Schätzern für stetige Konstrukte

Fehlende Werte

Fehlende Werte begegnen uns in zwei Formen:

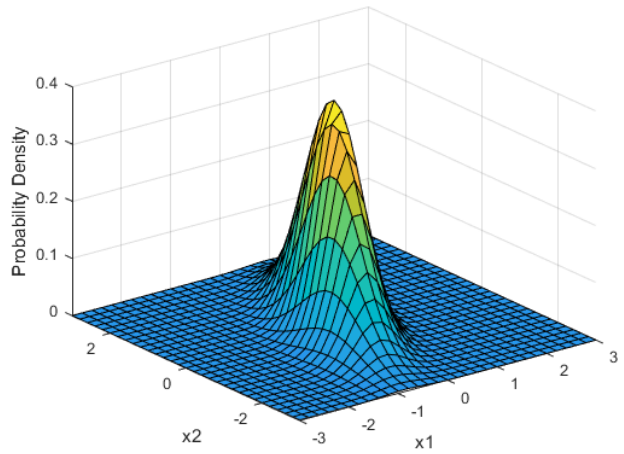
1. als ungeplanter Ausfall: Item- oder Unit-Non-Response
   Dieser Ausfall ist selten *missing-at-random (MAR)*, sondern meist systematisch. Nichts zu tun heißt in dem Fall, die empirischen Gegebenheiten naiv fortzuschreiben.

1. als theoretisch erwarteter Ausfall:
   Werden latente Konstrukte mittels reflexiven Messmodellen erhoben, können meist nicht alle Items vorgegeben werden. Es mangelt an:
   1. Inhaltsvalidität
   2. Reliabilität
   3. stetigen Schätzern für stetige Konstrukte

Multiple Imputation - *Joint Modeling*

Als Multiple Imputation werden diejenigen Verfahren bezeichnet, die fehlende Werte modellbasiert mehrfach ersetzen. Es kommt an einem bestimmten Punkt ein "kontrollierter" Zufall ins Spiel.

V Variablen wird eine multivariate (Normal-)Verteilung unterstellt (R-Paket *norm*).
Nach Schätzung der Parameter dieser Verteilung werden fehlende Werte - unter Beachtung vorhandener Werte - zufällig daraus gezogen.

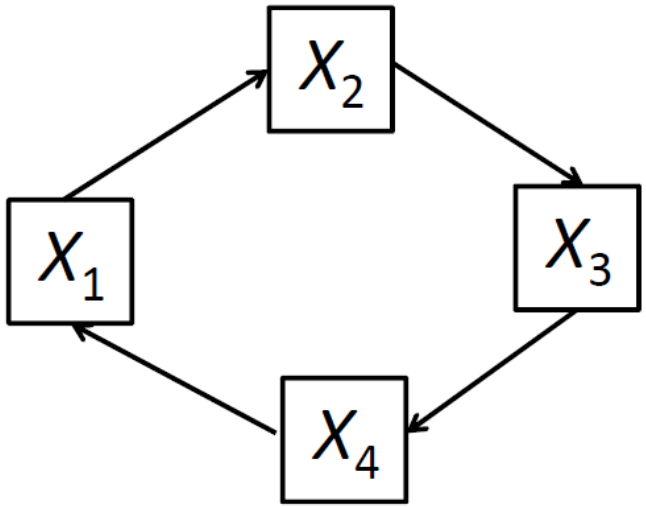


Multiple Imputation - *Fully Conditional Modeling*

Abfolge von bedingten Modellen mit höherer Flexibilität (R-Paket *mice*).

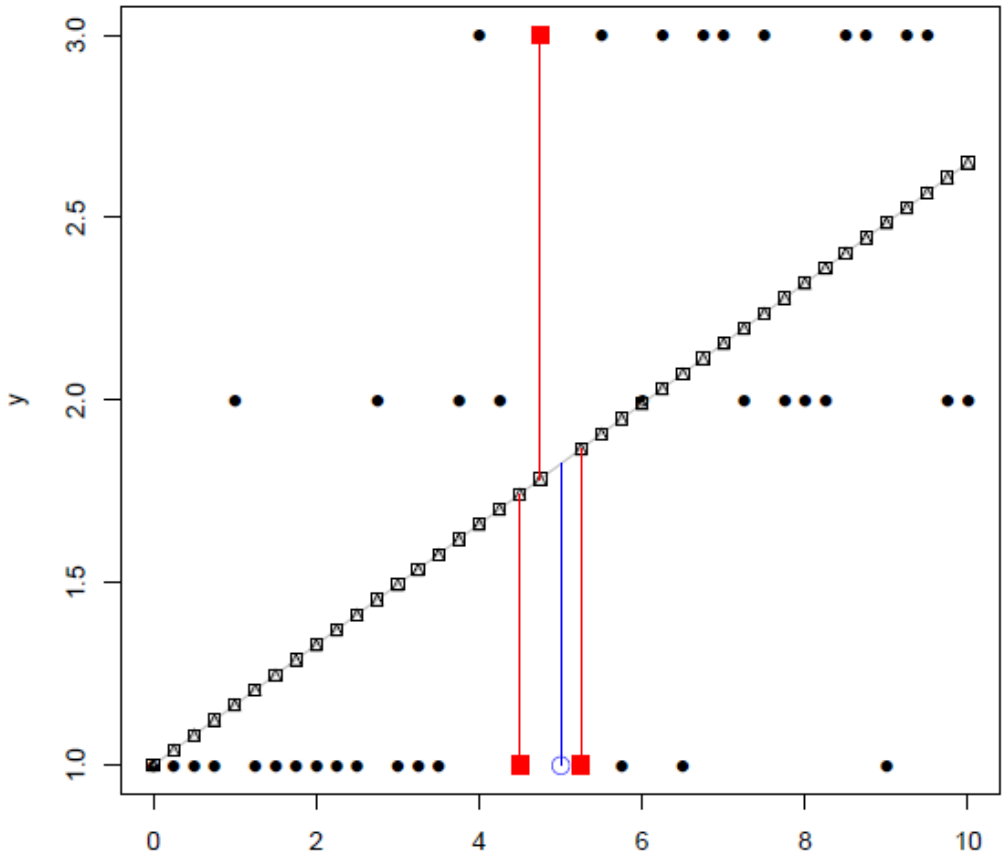
In jedem (zyklischen) Schritt wird eine Variable mit Werten aufgefüllt, die aus der Vorhersage durch kovariierende Variablen kommen. Der Standardfehler des Prädiktionsmodells wird zur zufälligen Abweichung verwendet.

Multiple **I**mputation By **C**hained **E**quations (ein EM-Algorithmus):

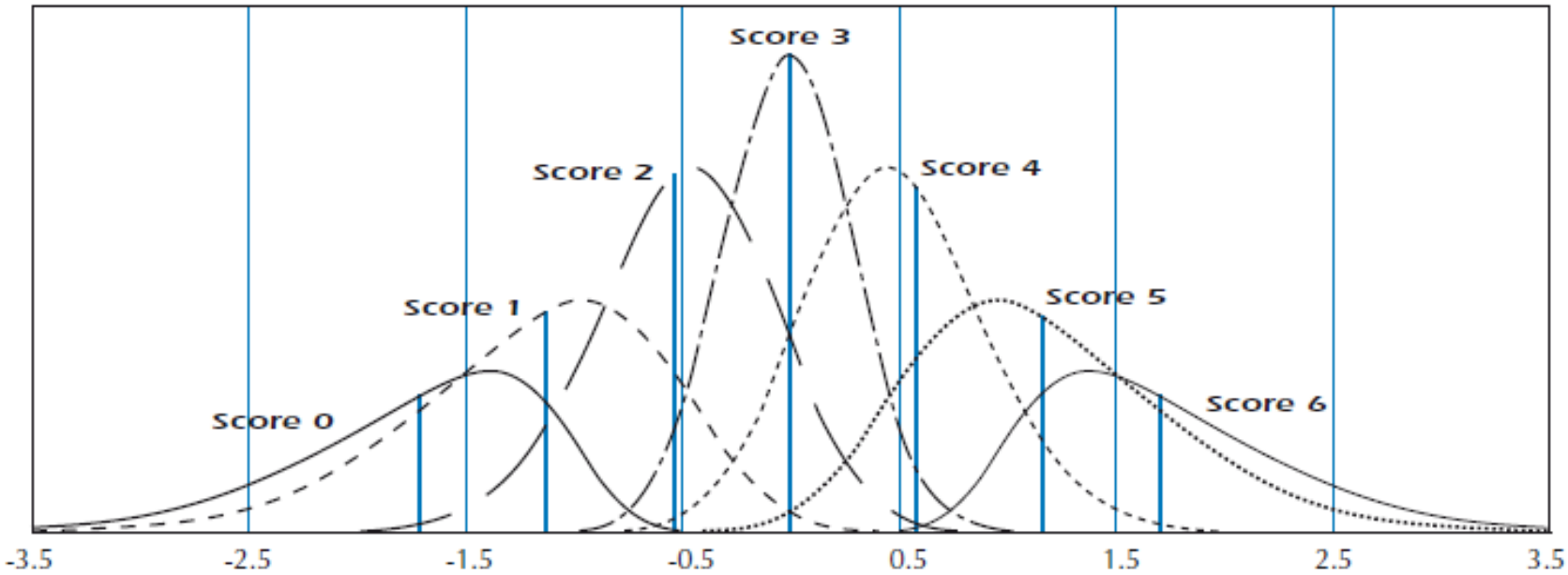


Multiple Imputation - Helferlein

- nur Prädiktoren mit substantiellem Erklärungsbeitrag
- kollineare Prädiktoren werden zu Faktoren zusammengefasst
- *Predictive Mean Matching* für kategoriale Variablen:



Diskrete Scores vs. Stetige Fähigkeiten



- IRT-Modelle übertragen *Item-Response-Patterns* auf eine beschränkte Menge möglicher Personenparameter (WLEs).
- Die latente, stetige Fähigkeit wird über eine *posteriori*-Verteilung (meist $\mathcal{N}(\theta_{EAP}, \sigma^2_{PV})$) geschätzt. Grundlage:
  - *Item-Response-Pattern*
  - Hintergrundvariablen mit Einfluss auf die latente Fähigkeit

Plausible Values

Plausible Values sind M zufällige Ziehungen aus dieser Verteilung, um Präzision sichtbar zu machen. Präzision steigt, wenn:

- Messfehler geringer sind
- Hintergrundvariablen deterministisch sind

Es kann gezeigt werden, dass $M \geq 5$ ausreicht, um Präzision ausreichend gut zu bestimmen.

Praktisch liegen so geschätzte latente Merkmale mehrfach vor.

- In separaten Spalten: PISA, TIMSS, PIRLS etc.
  $\rightarrow$nur Leistungswerte, zugrunde liegende Imputation der Hintergrundvariablen wieder entfernt
- In separaten Datenfiles: aktuelle Versionen der BIST-Datensätze
  $\rightarrow$komplette multiple Imputation auch von Hintergrundvariablen

Plausible Values - Unfair?

Ein Gedankenspiel

Ein/e Fußballtrainer/in lässt alle Spieler/innen einmal aufs bewachte Tor schießen. Wer versenkt, bleibt in der Mannschaft, wer daneben schießt oder abgewehrt wird, fliegt.

Ein/e Statistiker/in erhebt zusätzlich die Erfahrungsjahre im Fußballspiel und schätzt - zusammen mit dem aktuellen Test - die latente Wahrscheinlichkeit, zu versenken. (Je erfahrener, desto wahrscheinlicher.) Wessen *posteriori*-Verteilung ein 95%-Perzentil unter einer gewissen Trefferwahrscheinlichkeit hat, fliegt.

Kennwerte durch Pooling

Eine Schätzung der interessiendern statistischen Kenngröße(n) $\hat{\theta}$ erfolgt korrekt, indem die Ergebnisse der M Datenmatrizen mit imputierten Werten gemittelt werden:

$$\hat{\theta} = \frac{\sum_m \hat{\theta}_m}{M}$$

Es ist nicht korrekt, zuerst die zugrunde liegenden Variablen zu mitteln und dann nur einmal zu berechnen:

- Verschleierung der Imputationsvarianz
- Anfällig für Verzerrungen in den Randbereichen von PV-Verteilungen

Imputationsvarianz

Multiple Imputation erlaubt es, eine Imputationsvarianz zu berichten: Streuung von Kennwerten, die auf die Unsicherheit in den Daten zurückzuführen ist.

$$var_{imp}(\hat{\theta}) = \frac{\sum_m (\hat{\theta}_m - \hat{\theta})^2}{M-1}$$

Sie wird im allgemeinen geringer, wenn:

- weniger Fälle imputiert wurden
- die Imputation aufgrund einer starken Kovarianzstruktur präziser ist

Kombination Sampling- und Imputationsvarianz

Um einen Standardfehler für eine statistische Kenngröße $\hat{\theta}$ angeben zu können, müssen Sampling- und Imputationsvarianz noch kombiniert werden (Rubin, 1987).

$$var(\hat{\theta}) = var_{samp}(\hat{\theta}) + \frac{M+1}{M} \cdot var_{imp}(\hat{\theta}) \sim se(\theta)^2$$

Dabei wird die Samplingvarianz $var_{samp}(\hat{\theta})$ selbst als gepoolter Kennwert den M imputierten Datenmatrizen verstanden:

$$var_{samp}(\hat{\theta}) = \frac{\sum_m var_{samp}(\hat{\theta}_m)}{M}$$

$\rightarrow$ replizierte Berechnung von $N \cdot (M+1)$ Kennwerten

- TIMSS/PIRLS 2011: 75 x 5 + 5 = 450 Replikationen
- PISA 2015: 80 x 10 + 10 = 900 Replikationen
- BIST-Baseline 2010: 132 x 10 + 10 = 1.330 Replikationen

Datentransformation

Datentransformation ist eine häufige Notwendigkeit:

- Kategorisierung, Dichotomisierung, Rekodierung
- Skalenbildung (Mittelwerte, Summen, Faktoren)
- Zusammenführen von Daten über verschiedene Ebenen

In Zusammenspiel mit BIFIEsurvey:

1. Aufbereitung VOR Erzeugung des BIFIEdata-Objekts: gewohnte Transformationsbefehle, Gewichtung u.U. "händisch" zu beachten
2. Transformationen NACH Erzeugung des BIFIEdata-Objekts: mittels Formel-Framework, Gewichtung automatisch berücksichtigt

Bundesinstitut bifie

Generische Prozeduren

Sowohl Pooling von multiplen Kennwerten als auch Berechnung von Sampling- und Imputationsvarianz sind generisch.
D.h. es gibt keine grundlegenden Anforderungen an θ und der Algorithmus ist stets gleich.

$\rightarrow$ BIFIEsurvey kann mittels *BIFIE.by(...)* durch den/die Anwender/in erweitert werden.

Wald-Tests

Mittels Wald-Tests können beliebige Parameter-Vektoren auf Nullhypothesen geprüft werden. Bspw. Gleichheit von zwei Parametern in zwei Subgruppen.

Voraussetzung in iid-Datensätzen: Normalverteilung. Enders (2010) schlägt D1-D3-Statistiken vor, die diesen generischen Tests ähneln.

$\rightarrow$ BIFIEsurvey kann generische Hypothesentests mittels *BIFIE.waldtest(...)* durchführen. Spezifiziert werden muss die mathematische Form:

$$C \times \theta = r$$

Dabei ist C eine Design-Matrix und r der erwartete Ergebnisvektor.

Zwei-Ebenen-Regression

Siehe Übungsfile!

Nested Multiple Imputation (NMI)

... bezeichnet die korrekte Berechnung von Kennwerten und Standardfehlern bei geschachtelten Imputationen. Bspw. k vorgegebene PVs kombiniert mit eigener Imputation von Kontextvariablen (I mal).

Notwendig ist die Kombination der $k \cdot I$ Imputationen.

$$\hat{\theta} = \frac{\sum_k \sum_I \hat{\theta}_{kI}}{k \cdot I}$$

Komplexer noch ist die Berechnung der Imputationsvarianz.

KEINE NMI ist notwendig, wenn mehrere Analysevariablen aus der selben Imputation stammen, also bspw. bei Korrelation von PVs!

Weiterführende Literatur I

- Efron, B. & Hastie, T. (2016).
  Computer Age Statistical Inference.
  Cambridge: University Press.
- Enders, C.K. (2010).
  Applied Missing Data Analysis.
  New York: Guilford Press.
- OECD (2012).
  PISA Data Analysis Manual.
  Paris: OECD.
- Rubin, D.E. (1987).
  Multiple imputation for nonresponse in surveys.
  New York: Wiley.

Bundesinstitut bifie