Sozio-oekonomisches Panel (SOEP) II

Überblick

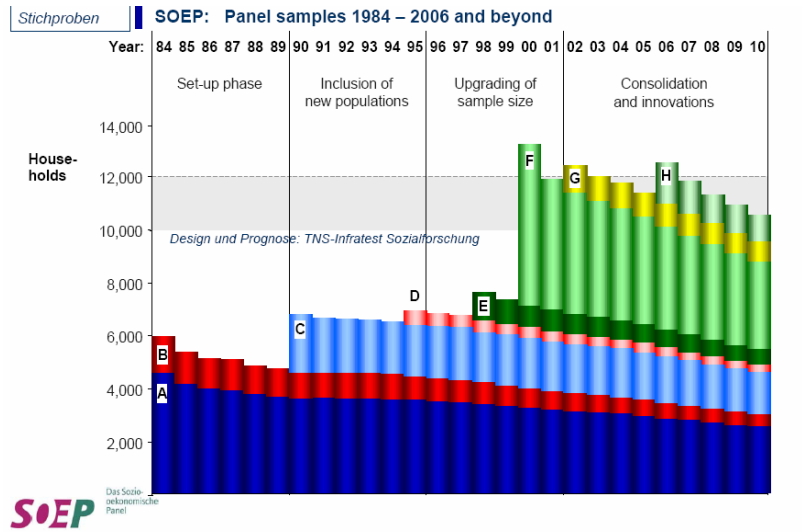
- Stichproben
- Erhebungsinstrumente
- Gewichtung
- Generierung eigener Datensätze

1. Stichproben und Datenerhebung

Stichprobendesign und Datenerhebung

- mehrstufige Zufallsauswahl
- genaues Vorgehen unterschiedlich, Beispiel Sample A („Westdeutsche“ 1984)
  - Stimmbezirke aus ADM-Master-Sample
  - Haushalte aus Stimmbezirken durch random-route-Verfahren
- Erhebung der Daten zu Beginn in Form von face-to-face Interviews, inzwischen auch andere Methoden der Datenerhebung
  - → siehe Abbildung

Stichprobenentwicklung



Schupp 2006

Stichproben

- 1984:
  - Sample A („Westdeutsche“): Personen in Haushalten mit deutschem Haushaltsvorstand, n=4.528, Auswahlwahrscheinlichkeit 0,0002
  - Sample B („Ausländer“): Personen in Haushalten mit türkischem, griechischem, jugoslawischen, spanischen oder italienischen Haushaltsvorstand, n=1.393, Auswahlwahrscheinlichkeit 0,0008

Stichproben

- 1990:
  - Sample C („Ostdeutsche“): Personen in Haushalten mit DDR-Bürger als Haushaltsvorstand, n=2.179, Auswahlwahrscheinlichkeit 0,0004
- 1994/95:
  - Samples D1 und D2 („Zuwanderer“): Personen in Haushalten, in denen mind. eine Person nach 1984 nach Deutschland zugewandert ist, n=236+295=522, Auswahlwahrscheinlichkeit 0,0002

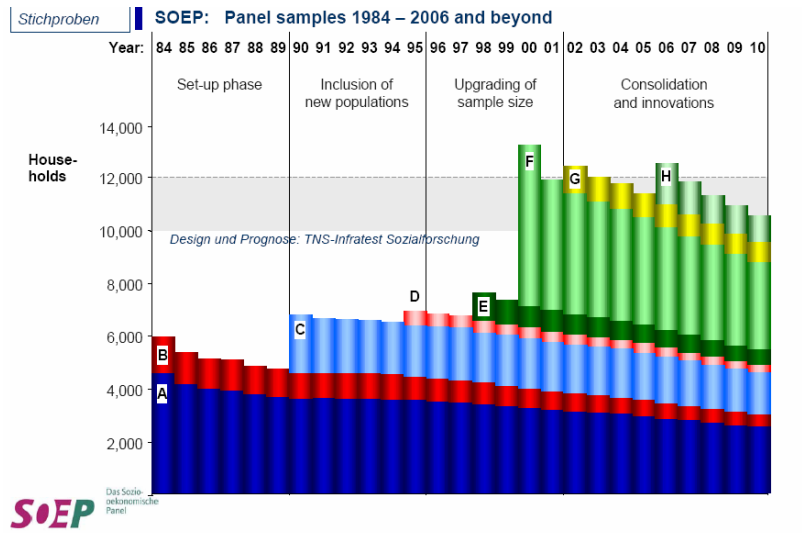
Stichproben

- 1998:
  - Sample E („Auffrischung“): Auswahl ähnlich wie in Sample A, Ziehung unabhängig von Sample A-D, n=1.067, Auswahlwahrscheinlichkeit 0,00003
- 2000:
  - Sample F („Innovation“): Auswahl ähnlich wie in Sample A und E, Überrepräsentierung nicht-deutscher Haushalte Ziehung unabhängig von Sample A-E, n=6.052, Auswahlwahrscheinlichkeiten: 0,00028 („deutsche“ HH), 0,0005 („nicht-deutsche“ HH)

Stichproben

- 2002:
  - Sample G („hohe Einkommen“): Haushalte mit einem Einkommen über DM 7.500, Ziehung unabhängig von Sample A-F, n=1.224
- 2006:
  - Sample H („Auffrischung“): Population wie A-D bzw. E und F, Ziehung unabhängig von bisherigen Stichproben, n=1.505

Stichprobenentwicklung



Schupp 2006

2. Erhebungsinstrumente

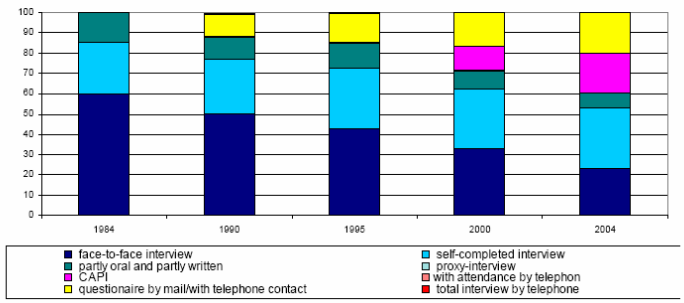
Erhebungsinstrumente

- Address Log: allgemeine Angaben zu HH-(Mitgliedern) und Interview (ausgefüllt durch Interviewer)
- Standardfragebögen:
  - Haushalt
  - Person (ab 16 Jahren) und „Lücke“ (zeitweiliger Unit-Nonresponse)
  - bis 1995 noch zusätzliche Fragebögen für Subgruppen (z.B. Immigranten, Ostdeutsche)
- Lebenslauf Erwachsene: Erstbefragte im Alter von 16+/18+ Jahren (Altersgrenze 2001 geändert)
- Lebenslauf Jugend: Erstbefragte im Alter von 16/17 Jahren (ab 2001)
- Mutter & Kind: Angaben zu Neugeborenen bis 15 Monaten (ab 2003)
- Mutter & Kleinkind: Angaben zu Kleinkindern bis 3 Jahren (ab 2005)

Datenerhebung 1984-2004

Erhebungsmethoden

Entwicklung der Befragungsmethode für Stichprobe A des SOEP



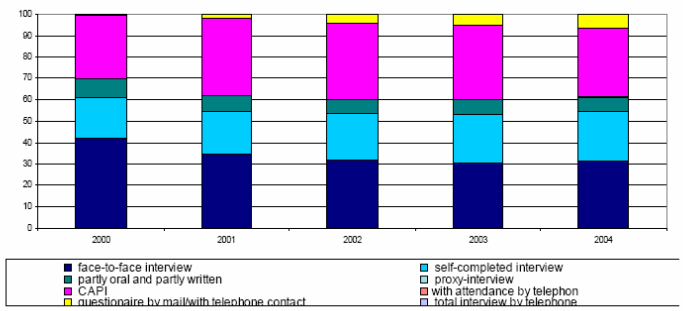
Schupp 2006

Forschungspraktikum: Wohlfahrtsstaat und soziale Ungleichheit SS 2007

Datenerhebung 1984-2004

Erhebungsmethoden

Entwicklung der Befragungsmethode für Stichprobe F des SOEP



Schupp 2006

3. Gewichtung

Gewichtung: Warum sollte gewichtet werden?

- Querschnitt

  → aufgrund unterschiedlicher Teilnahmewahrscheinlichkeiten und Auswahlwahrscheinlichkeiten
- Längsschnitt

  → aufgrund unterschiedlicher Wahrscheinlichkeit an weiteren Befragungen teilzunehmen (Verbleib im oder Ausstieg aus dem Panel, „Panelmortalität“)

Gewichtung: Warum sollte gewichtet werden?

Ausgangspunkt: Ausländeranteil in D etwa 10%

- Ziehung einer proportional geschichteten Stichprobe mit n=1000 ergibt eine Verteilung von etwa 900 Deutschen und 100 Ausländern
  - Problem: detaillierte Betrachtung von Ausländern aufgrund geringer Fallzahl nicht möglich
- Ziehung disproportional geschichteten Stichprobe, 3-fache Auswahlwahrscheinlichkeit von Ausländern ergibt bei n=1000 eine Verteilung von 700 Deutschen und 300 Ausländern
  - Problem: zwar getrennte Analysen für beiden Gruppen möglich, jedoch keine Analysen beider Gruppen zusammen → Überbetonung der Merkmale der Ausländer
  - Lösung: Gewichtung (zur Korrektur des Designeffekts)

Praktische Umsetzung

- Querschnitt
  - Personenebene: Gewichtungsvariable $PHRF
  - Haushaltsebene: Gewichtungsvariable $HHRF

    → oder nach Gruppen getrennte Betrachtung (als vorläufige Lösung)

    - z.B. summarize vp0101 [aweight=aphrf]
- Längsschnitt
  - Personenebene: Gewichtungsvariable $PBLEIB
  - Haushaltsebene: Gewichtungsvariable $HBLEIB

4. Generierung eigener Datensätze

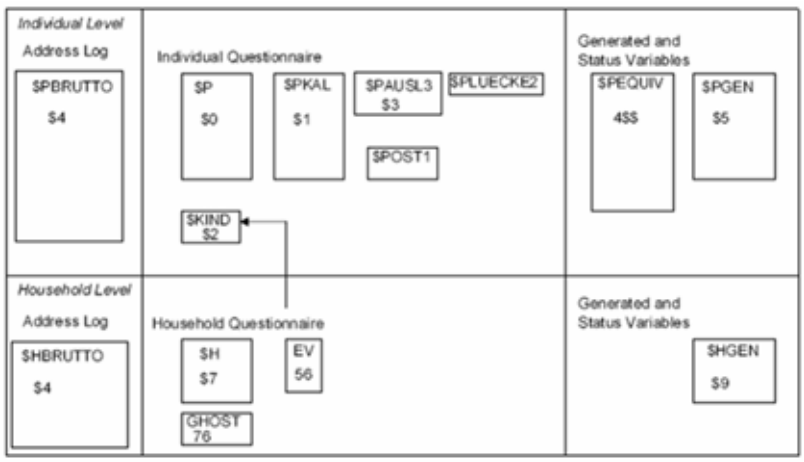
Problem

- Interessierenden Variablen sind in unterschiedlichen Datensätzen enthalten
  - Personen- und Haushaltsebene
  - Original- und generierte Variablen
  - Variablen aus Standard- und weiteren Fragebögen
  - Variablen aus unterschiedlichen Jahren
  - Querschnittsdaten und Spelldaten
  - Angaben zur Befragung und Stichprobe, Gewichtungsfaktoren

Problem

- Interessierenden Variablen sind in unterschiedlichen Datensätzen enthalten
  - Personen- und Haushaltsebene
  - Original- und generierte Variablen
  - Variablen aus Standard- und weiteren Fragebögen
  - Variablen aus unterschiedlichen Jahren
  - Querschnittsdaten und Spelldaten
  - Angaben zur Befragung und Stichprobe, Gewichtungsfaktoren

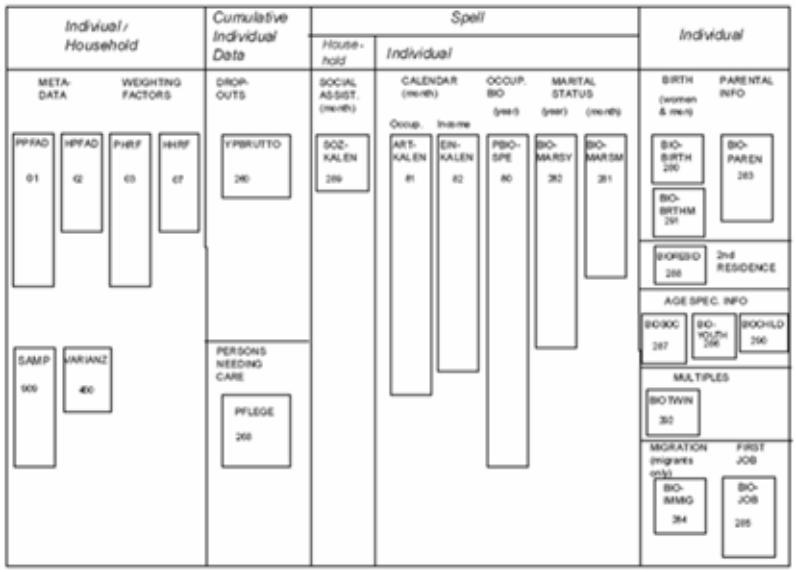
Datenstruktur: Querschnitt



$: Wave specification: A, B, C... U for file names; 1, 2, 3 ... 21 for file numbers.
1 Waves G and H only; 2 Waves B through Q only; 3 Waves A through L only

Haisken-DeNew/Frick 2005 : 30

Datenstruktur: Längsschnitt



Haisken-DeNew/Frick 2005 : 32

Personen- und Haushaltsdaten

hhnr	persnr	welle	sex	alter
1	1	2005	f	32
1	2	2005	m	34
1	3	2005	f	9
2	4	2005	m	62
2	5	2005	f	60
3	6	2005	f	81

hhnr	welle	hheink
1	2005	3000
2	2005	2000
3	2005	1000

Personen- und Haushaltsdaten

hhnr	persnr	welle	sex	alter	hheink
1	1	2005	f	32	3000
1	2	2005	m	34	3000
1	3	2005	f	9	3000
2	4	2005	m	62	2000
2	5	2005	f	60	2000
3	6	2005	f	81	1000

```
use hhdata, clear
sort hhnr
save hhdata, replace
use persdata, clear
sort hhnr
merge hhnr using hhdata
```

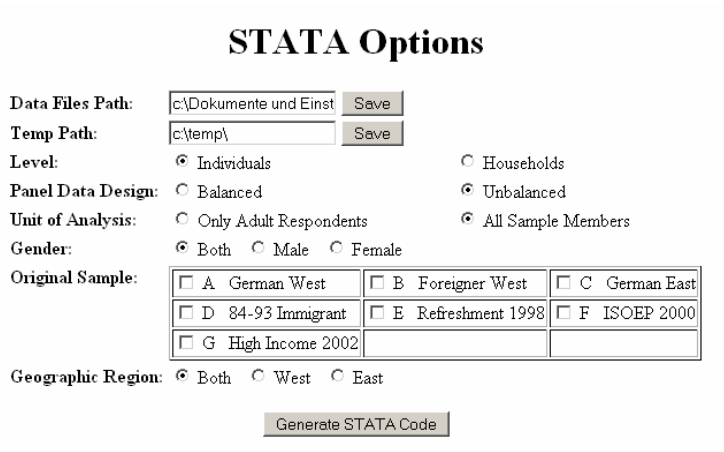

Angaben zur Befragung und Stichprobe, Gewichtungsfaktoren

- ppfad: enthält Angaben zur Befragung und Stichprobe für alle Personen und alle Wellen
- phrf: enthält Gewichtungsfaktoren für alle Personen und Wellen
- analog gibt es dazu Datensätze auf Haushaltsebene (hpfad, hhrf)

Generierung von Datensätzen mit SOEP-Info

- Auswahl der interessierenden Variablen (→ add to basket)
- Select all
- Aktion: Generators/Stata
- Optionen eingeben
- Generate Stata Code
- Stata Code kopieren und als do-file abspeichern
- do-file durchlaufen lassen
- Datensatz wird im angegebenen Verzeichnis gespeichert

SOEP-Info Generator: Optionen



5. Literatur

Literatur

- Haisken-DeNew, John P./ Frick, Joachim R (Hg.) (2005): DTC. Desktop companion to the German Socio-Economic Panel Study (SOEP). Berlin: Kapitel 4
- Kohler, Ulrich/Kreuter, Frauke (2006): Datenanalyse mit Stata, München/Wien: Abschnitt 10.4