Frick, Joachim R.; Rendtel, Ulrich; Wagner, Gert G.

Eine Strategie zur Kontrolle von Längsschnittgewichtungen am Beispiel des Sozio-oekonomischen Panels (SOEP)

DIW Discussion Papers, No. 80

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Frick, Joachim R.; Rendtel, Ulrich; Wagner, Gert G. (1993) : Eine Strategie zur Kontrolle von Längsschnittgewichtungen am Beispiel des Sozio-oekonomischen Panels (SOEP), DIW Discussion Papers, No. 80, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/95735

DIW Diskussionspapiere
Discussion Papers

Diskussionspapier Nr. 80

Eine Strategie zur Kontrolle von Längsschnittgewichtungen am Beispiel des Sozio-oekonomischen Panels (SOEP)

von
Joachim Frick, Ulrich Rendtel und Gerd Wagner

Deutsches Institut für Wirtschaftsforschung, Berlin
German Institute for Economic Research, Berlin

Deutsches Institut für Wirtschaftsforschung

Diskussionspapier Nr. 80

Eine Strategie zur Kontrolle von Längsschnittgewichtungen am Beispiel des Sozio-oekonomischen Panels (SOEP)

von
Joachim Frick, Ulrich Rendtel und Gerd Wagner

Berlin, im November 1993

Deutsches Institut für Wirtschaftsforschung, Berlin
Königin-Luise-Str. 5, 14191 Berlin
Telefon: 49-30 - 82 991-0
Telefax: 49-30 - 82 991-200

Eine Strategie zur Kontrolle von Längsschnittgewichtungen am Beispiel des Sozio-oekonomischen Panels (SOEP)

Joachim Frick, Ulrich Rendtel und Gert Wagner

1 Einleitung

Ein wesentliches Problem für die Längsschnittgewichtung von Paneldaten ist die korrekte Spezifikation eines Modells für die Panelmortalität. Hierbei versteht man unter Panelmortalität erhebungsbedingte Ausfälle nach der ersten Erhebungswelle eines Panels. Diese Ausfälle treten dadurch auf, daß kein Kontakt mit Haushalten aus der vorhergehenden Panelwelle hergestellt werden kann bzw. daß kein Interwiev mit wiedererreichten Haushalten zustande kommt. Hierbei faktorisiert sich die Wahrscheinlichkeit, im Panel zu verbleiben, in eine Sequenz von bedingten Wahrscheinlichkeiten, die jeweils das erfolgreiche Verbleiben auf den einzelnen Erhebungsstufen des Panels bestimmen. Diese Erhebungsstufen sind durch das Wiedererreichen eines Haushaltes in der folgenden Panelwelle und die Antwortgewährung von wiedererreichten Haushalte charakterisiert; vgl. hierzu Rendtel (1991,1993) sowie Pischner/Rendtel (1993).

Man kann das Produkt dieser Wahrscheinlichkeiten im Panel zu verbleiben zu einer Gewichtung der Längsschnittprobe benutzen, die aus denjenigen Personen besteht, die an allen Panelwellen teilgenommen haben. Unter der korrekten Spezifikation eines Modells für die feldbedingten Verluste während des Panels stimmt die so gewichtete Merkmalsverteilung der Längsschnittstichprobe im Erwartungswert mit der Merkmalsverteilung der "Grundgesamtheit" von Welle 1 überein.

Eine derartige Gewichtung bezieht sich auf eine "Grundgesamtheit", die aus den Teilnehmern der ersten Panelwelle abzüglich der demographischen Verluste bis Ende der Längsschnittperiode besteht, wobei demographische Verluste durch Tod oder durch Fortzug ins Ausland auftreten.

Diese Überlegung ist die Basis für den in dieser Arbeit vorgeschlagenen "Spezifikationstest", der zur Kontrolle der Längsschnittgewichtung eines Panels dient. Die Grundidee besteht darin, hinsichtlich bestimmter Merkmale aus Welle 1 die Übereinstimmung der mit den Verbleibewahrscheinlichkeiten gewichteten Längsschnittstichprobe und der "Grundgesamtheit" von Welle 1 zu prüfen. Sind diese

Differenzen größer als es die Stochastik des Ausfallprozeßes erwarten läßt, so ist dies ein Hinweis darauf, daß das betreffende Merkmal bei der Spezifikation der Panelmortalität anders berücksichtigt werden muß[1].

Weiterhin kann man derartige "Soll-Ist"-Vergleiche zu einer Korrektur der Längsschnitt-Hochrechnungsfaktoren (Längsschnittgewichte) benutzen. Dies entspricht dem klassischen Randanpassungsverfahren zwischen Grundgesamtheit und Stichprobe[2]. Das Verfahren der Randanpassung berücksichtigt dabei die in den Längsschnittgewichten berücksichtigten Information derart, daß der Informationsabstand zwischen den bisherigen und den korrigierten Gewichten minimal bleibt; vgl. Ireland/Kullbach (1968), Deville/Särndal (1992) sowie Little/Wu (1991).

Dieses Verfahren wird am Beispiel des Sozio-oekonomischen Panels (SOEP)[3] demonstriert. Die Längsschnittgewichtung des SOEP basiert auf der Erkenntnis, daß im wesentlichen Ereignisse, die mit der unmittelbaren Feldarbeit im Zusammenhang stehen, für das Ausfallrisiko relevant sind. Zudem wurden bei der Ausfallanalyse auch andere Variablen wie Haushaltseinkommen, Alter, Geschlecht, Familienstand und Schulbildung des Haushaltsvorstandes kontrolliert[4]. Trotz des beträchtlichen Aufwandes, der in die Analyse der Ausfälle und die Berechnung der Längsschnittgewichte gesteckt wurde, ist eine "korrekte" Spezifikation des Ausfallprozeßes sicherlich eine Fiktion.

Es sei an dieser Stelle auf einige potentielle Quellen der Fehlspezifikation der SOEP-Längsschnittgewichtung hingewiesen. Die der Gewichtung zugrundeliegende Ausfallanalyse basiert auf der sequentiellen Schätzung der $(9-1) \times 2 = 16$ Teilnahmestufen des SOEP in Abhängigkeit von Haushaltscharakteristika der vorhergehenden Welle und einigen feldrelevanten Charakteristika der aktuellen Panelwelle (z.B. ob der Haushalt umgezogen ist oder sich von einem bestehenden Panelhaushalt abgespalten hat). Das benutzte Ausfallmodell ist fehlspezifiziert, falls:

---

[1] Eine derartige Vorgehensweise hat viele Ähnlichkeiten mit klassischen Repräsentativitätsstudien, die Werte aus der Grundgesamtheit mit Schätzwerten aus der Stichprobe vergleichen, vgl. hierzu Rendtel/Pötter 1993 sowie Pötter/Rendtel 1993. Es gibt allerdings auch wesentliche Unterschiede. Insbesondere kann es keine Zweifel daran geben, daß das für "Grundgesamtheit" und die Stichprobe gemessene Merkmal wirklich denselben Sachverhalt ermittelt.

[2] Im Unterschied zu diesem Vorgehen, bei dem die untersuchten Merkmale für die Nonrespondenten unbekannt sind, besteht jedoch bei der Längsschnittgewichtung die alternative Möglichkeit, das untersuchte Merkmal zusätzlich zu der bisherigen Modellspezifikation in einer multiplen Ausfallanalyse zu berücksichtigen. Allerdings ist eine derartige Ausfallanalyse über zur Zeit 9 Panelwellen des SOEP mit $(9-1) \times 2 = 16$ Teilnahmestufen (Pro Folgewelle 2 Übergänge: Kontaktherstellung und Antwortgewährung) sehr aufwendig, so daß das hier vorgestellte Randanpassungsverfahren einen gewissen Kompromiß zwischen der Nichtberücksichtigung weiterer Merkmale für den Ausfallprozeß und einer aufwendigen Ausfallanalyse darstellt.

[3] vgl. zum SOEP Projektgruppe Panel (1993).

[4] Eine Dokumentation der für die Längsschnittgewichtung des SOEP benutzten Modelle findet man in Pischner/Rendtel (1993).

- die Ausfallwahrscheinlichkeit von Merkmalsveränderungen seit der letzten Panelwelle abhängt;
- die Ausfallwahrscheinlichkeit von Merkmalsveränderungen vor der letzten Panelwelle abhängt.
- Schließlich kann noch der Fall eintreten, daß bestimmte Einflußfaktoren auf jeder Einzelstufe als nicht signifikant beurteilt werden und daher auch nicht in die Gewichtung eingehen[5]. Faßt man jedoch alle Einzelstufen zusammen, so kann sich dieser Einflußfaktor dennoch als bedeutsam erweisen. Dieser Fall tritt ein, wenn die nicht signifikanten Einzelbeiträge gleichgerichtet sind.

Der Artikel ist wie folgt gegliedert. Im folgenden Abschnitt werden die statistischen Grundlagen der hier vorgeschlagenen Kontrollstrategie behandelt. Diese Strategie kann prinzipiell für beliebige Längsschnitte angewandt werden. Nach einer kurzen Beschreibung der Vorgehensweise zur empirischen Umsetzung der Prüfung der SOEP-Längsschnittgewichtung werden im dritten Abschnitt Längsschnitte über 2 Wellen behandelt. Diese prüfen die Zuverlässigkeit der Ausfallanalyse zwischen je 2 Panelwellen. Anschließend wird die Ausfallanalyse auf das Zeitintervall von Welle 1 bis Welle 5 ausgedehnt. Zur Prüfung werden die Merkmale "Haushaltseinkommen" und "Bezug von Sozialhilfe" gewählt.

Die Wahl dieser Merkmale und des Analysezeitraumes beruht auf der Auseinandersetzung darüber, ob aus dem SOEP zum Themenkomplex "Haushalte im Niedrigeinkommensbereich" zuverlässige Längsschnittanalysen gewonnen werden können; vgl. Andreß et al. (1993) und Lipsmeier (1993). Abschließend enthält Abschnitt 4 eine exemplarische Korrektur für den Fall einer Fehlspezifikation des Ausfallmodells.

2 Statistische Grundlagen

Es sei t_A die Anfangswelle und t_E die Endwelle des interessierenden Längsschnittintervalls. Die Menge G bestehe aus denjenigen Personen der Stichprobe der Panelwelle t_A, die bis Welle t_E noch im Erhebungsgebiet leben. Aus dieser "Grundgesamtheit" G wird eine "Stichprobe" S gezogen, die aus denjenigen Personen von G besteht, die bis Welle t_E ununterbrochen am Panel teilgenommen haben. Eine Person i gelangt von G in die "Stichprobe" S mit einer Wahrscheinlichkeit π_i. Diese Wahrscheinlichkeit ergibt sich als das Produkt der Wahrscheinlichkeiten, die zwischen den Panelwellen t_A und t_E liegende Teilnahmestufen erfolgreich zu überwinden. Das Ziehungsverfahren entspricht damit "Ziehen ohne Zurücklegen" mit ungleichen Ziehungswahrscheinlichkeiten π_i.

[5]Dieser Verzicht liegt darin begründet, daß durch die Berücksichtigung von (beliebig) vielen Merkmalen die Varianz der Schätzergebnisse vergrößert wird

Es sei Y_i ein Merkmalsindikator für Charakteristika von Person i in Welle t_A. Der "Populationswert"

$$P_Y = \sum_{i \in G} Y_i \tag{1}$$

kann über den Horvitz-Thompson Schätzer

$$\hat{P}_Y = \sum_{i \in S} \frac{1}{\pi_i} Y_i \tag{2}$$

erwartungstreu geschätzt werden. Für Merkmale aus Welle t_A sind sowohl P_Y als auch $\hat{P}_Y$ bekannt. Allerdings gilt die Erwartungstreue von $\hat{P}_Y$ nur unter der korrekten Spezifikation des Modells für die π_i. Unter dieser Voraussetzung muß das Verhältnis

$$R_Y = \frac{\hat{P}_Y}{P_Y} \tag{3}$$

den Erwartungswert 1 haben[6].

Um zu bestimmen, welche Abweichungen vom Erwartungswert signifikant sind, braucht man ein Maß für die statistische Streuung von $\hat{P}_Y$ bzw. R_Y unter der Nullhypothese, daß die π_i korrekt spezifiziert sind. Da P_Y im Rahmen des hier benutzten Randomisierungsansatzes als Konstante behandelt wird, erhält man die Varianz von R_Y sofort als $V(\hat{P}_Y)/P_Y^2$.

Es bezeichne im folgenden π_{ij} die Wahrscheinlichkeit, daß sowohl Person i als auch Person j in die Stichprobe S gelangen. Man erhält dann für die Varianz von $\hat{P}_Y$ den Ausdruck:

$$V(\hat{P}_Y) = \sum_{i \in G} \frac{\pi_i(1-\pi_i)}{\pi_i^2} Y_i^2 + \sum_{i \neq j \in G} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) Y_i Y_j \tag{4}$$

Um $V(\hat{P}_Y)$ zu berechnen, müssen also die gemeinsamen Auswahlwahrscheinlichkeiten bekannt sein. Hierbei ist eine Eigenschaft im Teilnahmeverhalten der Befragungspersonen des SOEP wesentlich: Das Teilnahmeverhalten bestimmt sich im wesentlichen auf Haushaltsebene (vgl. hierzu Rendtel 1993). D.h. für Personen, die während der gesamten Zeitperiode in einem Haushalt leben, gilt:

[6]Die Berechnung von R_y läßt sich sehr leicht mit den in der SOEP-Datenbank gespeicherten "Bleibewahrscheinlichkeiten" durchführen. Beispielsweise erhält man für $t_A = 1$ (Welle A) und $t_E = 5$ (Welle E) $1/\pi_i = W_i$ =BPBLEIB * CPBLEIB * DPBLEIB * EPBLEIB. Für Personen, die bis Welle 5 aus dem Panel ausgeschieden sind, hat die so gebildete Variable W_i den Wert 0. Man erhält dann R_y indem man W_i über alle Personen aus G mit $Y_i = 1$ mittelt.

$$\begin{aligned} \pi_{ij} &= \pi_{j|i}\pi_i \\ &= 1 \cdot \pi_i \end{aligned} \tag{5}$$

Ein Problem ergibt sich dadurch, daß sich für einige Personen der Haushaltszusammenhang während der betrachteten Zeitperiode auflöst. Zwar gibt es Hinweise, daß auch nach der Auflösung des Haushaltskontextes noch weiter (positiv korrelierte) Abhängigkeiten im Teilnahmeverhalten bestehen (vgl. hierzu Rendtel 1993), jedoch ist dieser Zusammenhang nicht mehr perfekt, so daß für derartige Fälle

$$\pi_i\pi_j \leq \pi_{ij} \leq \pi_i \tag{6}$$

gilt.

Damit ergibt sich eine Möglichkeit, untere und obere Schranken für $V(\hat{P}_Y)$ anzugeben. Es sei H_i die Menge aller derjenigen Personen, die mit Person i über die gesamte Zeitperiode in einem Haushalt leben und H_i^* die Menge derjenigen Personen, die mit Person i zu Beginn der Zeitperiode in einem Haushalt zusammen leben. Dann folgt aus (4)-(6):

$$\begin{aligned} V_U(\hat{P}_Y) &= \sum_{i\in G} \frac{1-\pi_i}{\pi_i} Y_i^2 + \sum_{i\in G}\sum_{j\in H_i} \left(\frac{1}{\pi_j} - 1\right) Y_iY_j \\ &\leq V(\hat{P}_Y) \\ &\leq \sum_{i\in G} \frac{1-\pi_i}{\pi_i} Y_i^2 + \sum_{i\in G}\sum_{i\in H_i^*} \left(\frac{1}{\pi_j} - 1\right) Y_iY_j \\ &= V_O(\hat{P}) \end{aligned} \tag{7}$$

Da die Merkmale für die Mitglieder der Grundgesamtheit bekannt sind, können $V_U(\hat{P}_Y)$ und $V_O(\hat{P}_Y)$ bequem bestimmt werden, wenn die π_i-Werte auch für die ausgefallenen Personen in der Datenbank bereit gestellt sind.

Dies ist prinzipiell möglich, da π_i auch für die ausgefallenen Personen geschätzt wird. Allerdings sind diese Werte nicht in der SOEP-Datenbank gespeichert. Da die nachträgliche Berechnung der π_i für die ausgefallenen Personen relativ aufwendig ist, seien hier zwei alternative Vorgehensweisen beschrieben.

Die erste Vorgehensweise benutzt die Eigenschaft, daß

$$\hat{V}(\hat{P}_Y) = \sum_{i\in S} \frac{1-\pi_i}{\pi_i^2} Y_i^2 + \sum_{i\in S}\sum_{j\in S} \left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right) \frac{Y_iY_j}{\pi_{ij}} \tag{8}$$

ein erwartungstreuer Schätzer für $V(\hat{P}_Y)$ ist, falls $\pi_{ij} > 0$ für alle $i, j \in G$. Diese Bedingung ist wegen $\pi_{ij} \geq \pi_i\pi_j > 0$ erfüllt.

Wiederum ergibt sich das Problem, daß die π_{ij} nicht für alle $i-j$ Kombinationen bekannt sind. Unter Verwendung von (6) erhält man wiederum Unter- und Obergrenzen von $\hat{V}(\hat{P}_Y)$:

$$\begin{aligned}
\hat{V}_U(\hat{P}_Y) &= \sum_{i\in S}\frac{1-\pi_i}{\pi_i^2}Y_i^2+\sum_{i\in S}\sum_{j\in H_i}\left(\frac{1}{\pi_j}-1\right)\frac{Y_iY_j}{\pi_i} \qquad (9)\\
&\leq \hat{V}(\hat{P}_Y)\\
&\leq \sum_{i\in S}\frac{1-\pi_i}{\pi_i^2}Y_i^2+\sum_{i\in S}\sum_{j\in H_i^*}\left(\frac{1}{\pi_j}-1\right)\frac{Y_iY_j}{\pi_i}\\
&= \hat{V}_O(\hat{P}_Y)
\end{aligned}$$

Bei der Berechnung von $\hat{V}_U(\hat{P}_Y)$ sind die Mengen H_i und H_i^* zu bestimmen. Ignoriert man die Möglichkeit, daß Personen im Verlaufe des Panels nach einer Trennung des Haushalts wieder zusammenziehen, so wird die Menge H_i recht genau durch diejenigen Personen aus S beschrieben, die auch noch in Welle t_E in einem Haushalt zusammen wohnen. Die Menge H_i^* besteht aus allen Personen aus S, die mit Person i in Welle t_A in einem Haushalt gelebt hat[7]. Je weiter t_A und t_E auseinander liegen, desto weiter werden die Varianzschätzungen $\hat{V}_U(\hat{P}_Y)$ und $\hat{V}_O(\hat{P}_Y)$ voneinander abweichen.

Alternativ zu dieser analytischen Varianzschätzung sei hier noch eine zweite Methode der Varianzschätzung vorgeschlagen. Diese Methode benutzt den Random Group Ansatz, der beispielsweise bei Wolter (1985, S. 30 ff) beschrieben ist. Diese Methode wird auch für die Schätzung der Varianz von Populationsschätzwerten im SOEP benutzt (vgl. Rendtel 1991).

Die Grundidee des Random Group Konzepts besteht darin, eine Situation herzustellen, in der R unabhängig gezogene Stichproben nach demselben Erhebungsdesign existieren. Zu diesem Zweck wird die Ausgangsstichprobe in Teilstichproben (Random Groups) aufgeteilt. Diese Aufteilung geschieht derart, daß die Random Groups wie Realisationen des ursprünglichen Erhebungsdesign mit verringerter Stichprobengröße betrachtet werden kann. In diesem Fall ist das Erhebungsdesign "Ziehen ohne Zurücklegen". Dieses Erhebungsdesign wird repliziert, in dem man unabhängig für jede gezogene Einheit eine Zufallszahl $r \in \{1,\ldots,R\}$ zieht und alle Personen, die zu dieser Einheit gehören, der Teilstichprobe r zuordnet (vgl. Wolter 1985, S. 31 ff). Als Einheiten werden hierbei Haushalte

[7] $\hat{V}_U(\hat{P}_Y)$ läßt sich zu $\sum_{i\in S}1/\pi_iY_i(\sum_{j\in K_i}(1/\pi_j-1)Y_j)$ vereinfachen. Hierbei ist K_i die Menge H_i erweitert um die Person i. Einen analogen Ausdruck erhält man für $\hat{V}_O(\hat{P}_Y)$. Bei der praktischen Berechnung wird man zunächst die Aggregate $\sum_{j\in K_i}(1/\pi_j-1)Y_j$ berechnen. Dieses Aggregat wird für alle Haushalte in Welle t_E (Berechnung von $\hat{V}_U(\hat{P}_Y)$ bzw. für alle Haushalte in Welle t_A (Berechnung von $\hat{V}_O(\hat{P}_Y)$ bestimmt. In einen zweiten Rechenschritt wird diese Haushaltsinformation mit der Personeninformation Y_i/π_i zusammengeführt.

gewählt. Hierdurch wird die Abhängigkeit in Teilnahmeverhalten zwischen Personen reproduziert. Hierbei können wieder die Starthaushalte in Welle t_A oder die Endhaushalte in Welle t_E gewählt werden. Die Wahl der Starthaushalte führt zu größeren Personenclustern und damit zu höheren Varianzschätzungen. Sie bezieht sich auf $V_O(\hat{P}_Y)$. Die Wahl der Endhaushalte bezieht sich auf $V_U(\hat{P})$. Varianzschätzungen mit Hilfe der Randomisierung der Endhaushalte sind also kleiner, d.h. Abweichungen einer gewichteten Randverteilung von der Verteilung in der "Grundgesamtheit" (1. Welle) werden eher als signifikant eingestuft[8].

Bei der praktischen Umsetzung dieses Konzeptes wird zunächst die Zufallsgröße r für alle Personen in S generiert (zur empirischen Umsetzung siehe Anhang A). Dies liefert eine disjunkte Einteilung von S in R disjunkte Teilstichproben $S_r (r = 1, \ldots, R)$. Diese Einteilung liefert ihrerseits r Schätzungen von $R = R_Y$. Hierbei erhält man die r-te Schätzung, R_r, durch[9]:

$$R_r = \frac{R \sum_{i \in S_r} \frac{1}{\pi_i} Y_i}{\sum_{i \in G} Y_i} \qquad r = 1, \ldots, R \tag{10}$$

Es gilt:

$$\begin{aligned} \bar{R} &= \frac{1}{R} \sum_{r=1}^{R} R_r \\ &= R_Y \end{aligned} \tag{11}$$

Dann ist:

$$\hat{V} = \frac{1}{R(R-1)} \sum_{r=1}^{R} (R_r - R_Y)^2 \tag{12}$$

der Random-Group Schätzer von $V(R_Y)$.

Eine einfache Intervallschätzung für R_Y erhält man auf Basis der geordneten Werte $R_{(r)}$ der R_r. Hierbei gilt:

$$R_{(1)} \leq R_{(2)} < \ldots < R_{(R)} \tag{13}$$

[8]Das SOEP enthält zur Abschätzung der Varianz von Populationsschätzungen standardmäßig eine "Randomisierungsvariable" RGROUP. Theoretisch ist jedoch die Wahl der RGROUP-Variablen für die folgenden Analysen abzulehnen, da pro Sample-Point nur eine Zufallszahl gezogen wurde. Daher wird bei der empirischen Umsetzung eine neue Zufallsgruppeneinteilung vorgenommen, die unabhängig von der Sample-Point Zugehörigkeit ist (vgl. Anhang A).

[9]Rechentechnisch erhält man R_r, indem man die Variable r_i berechnet, die den Wert 1 hat, falls Person i der r-ten Teilstichprobe angehört und 0 sonst. Dann wird die Variable Rr_i/π_i über die Merkmalsträger aus G mit $Y_i = 1$ gemittelt.

Für R=8 hat das Intervall $(R_{(2)}, R_{(7)})$ eine Überdeckungswahrscheinlchkeit von 0.93 (vgl. Rendtel 1991).

Die in der Einleitung bereits skizzierte Kontrolle der Längsschnittgewichtung prüft nun, ob R_Y mehr als zufällig von seinen Erwartungswert abweicht. Eine derartige Entscheidung kann darüber gefällt werden, ob die Ungleichung

$$\mid R_Y - 1 \mid \geq C\sqrt{V(R_Y)} \tag{14}$$

für ein vorgegebenes C erfüllt wird. Hierbei ist der Wert C= 2 üblich (2-Sigma-Regel). Unter Normalverteilung von $R_Y/\sqrt{V(R_Y)}$ wird diese Ungleichung unter der Nullhypothese nur mit der Wahrscheinlichkeit 0.05 auftreten. Alternativ kann man für R= 8 die Nullhypothese verwerfen, falls

$$1 \notin (R_{(2)}, R_{(7)}) \tag{15}$$

d.h. das Konfidenzintervall überdeckt nicht den erwarteten Wert 1. Im Gegensatz zu dem Varianzvergleich benötigt das Verfahren über die Ordnungsstatistik keine Normalverteilungsannahme zur Absicherung des Fehlerniveaus von 0.07.

In der Regel wird man eine derartige Überprüfung der Nullhypothese ("Korrekte Spezifizierung der Ausfallwahrscheinlichkeiten") für mehrere Merkmale durchführen. Diese Strategie berücksichtigt jedoch nicht Effekte des multiplen Testens. Dieser Effekt besteht darin, daß die Wahrscheinlichkeit, bei irgendeinem der geprüften Merkmale die Nullhypothese abzulehnen, größer ist als das Fehlerniveau der Einzeltests. Umgekehrt kann der Fehler 2. Art, d.h. das Festhalten an H_O, obwohl H_O falsch ist, beträchtlich sein. Dies sind prinzipielle Schwächen von Soll/Ist-Vergleichen der Stichproben mit der Grundgesamtheit; vgl. hierzu Pötter/Rendtel (1993).

Allerdings sind die Konsequenzen einer Ablehnung von H_O nicht so gravierend, wie bei der formal sehr ähnlichen "Repräsentativitätsprüfung" von Stichproben, da hier eher die Möglichkeit besteht, über eine Re-Spezifikation des Ausfallmodells zu einer eventuell verbesserten Längsschnittgewichtung zu gelangen.

Der einfachste Weg, ein neues Merkmal in der Längsschnittgewichtung zu berücksichtigen besteht darin, die π_i mit dem Kehrwert von R_Y zu multiplizieren. Dies entspricht einer klassischen Soll/Ist-Anpassung:

$$\pi_i^* = \frac{\sum_{j \in G} Y_j}{\sum_{j \in S} \frac{1}{\pi_j} Y_j} \pi_i \qquad i \in S \tag{16}$$

Die allgemeinen Eigenschaften einer Randanpassung wurden von Gokhale/Kullbach (1978) unter Verwendung von Informations- und Entropiemaßen beschrieben. Unter modelltheoretischen Gesichtspunkten wird hier ein zusätzlicher Haupteffekt beschrieben, vgl. Little/Wu (919). Allerdings bleiben Interaktionen mit bereits im Ausfallmodell aufgenommenen Merkmalen unberücksichtigt.

3 Zur Prüfung der Längsschnittgewichtung des SOEP

Die aus der Literatur bekannten Zustandsvariablen, die die Ausfallwahrscheinlichkeiten beeinflussen, sind bei der SOEP-Gewichtung berücksichtigt worden (z.B. der Bildungsabschluß), nicht jedoch Ereignisse wie Arbeitslosigkeit, Ehescheidung oder Bezug von Transferzahlungen. Deren Einfluß kann allerdings häufig indirekt modelliert werden (z.B. durch die Berücksichtigung der Einkommensklasse des Haushaltes und eines Interviewerwechsels)[10].

Hängt die Teilnahmewahrscheinlichkeit von einem nicht durch die Modellierung berücksichtigten Ereignis ab, dann wird die davon betroffene Gruppe im Laufe der Zeit in der Stichprobe unterrepräsentiert sein. Dies ist im SOEP z.B. für Geschiedene der Fall, da - wie jüngste Analysen gezeigt haben - die Abhängigkeit des Ausfallrisikos zwischen sich trennenden Haushalten bislang nicht zutreffend modelliert wird (vgl. Rendtel 1993).

Ein ähnliches Problem ist auch für mobile Haushalte zu zeigen: Zieht ein Haushalt um, dann ist in der Längsschnittgewichtung des SOEP zwar im Jahr des Umzuges die Modellierung des Teilnahmeverhaltens nahezu perfekt und in den Jahren $t+2$ fortfolgend ist kein Umzugseffekt mehr beobachtbar, der die Teilnahmewahrscheinlichkeit senkt. Jedoch ist im Jahr $t+1$ eine überdurchschnittliche Ausfallwahrscheinlichkeit gegeben, die bislang nicht bei der Modellierung berücksichtigt wurde (vgl. Frick 1993, Anhang C) [11].

Merkmale wie "Arbeitslose im Haushalt", oder "Sozialhilfebezug" gehen nicht in die Längsschnittsgewichtung des SOEP ein. Deswegen vermuten Andreß et al. (1993) sowie Lipsmeier (1993) eine Unterschätzung dieser Population im SOEP. Zudem wird befürchtet, "daß sich ein solcher (wenn auch nicht hoher) Selektionseffekt über mehrere Wellen kumuliert, so daß der Anteil von Haushalten mit geringem Einkommen an der Panelpopulation im Laufe der Jahre deutlich abnimmt" (Lipsmeier 1993, S. 10) Dann wäre eine spezifische Form eines "Mit-

[10]Gemessen an den Anforderungen eines theoriegeleiteten Gewichtungsverfahren ist die SOEP- Gewichtung, die auch feldbezogene Merkmale und Ereignisse berücksichtigt, weit besser als die bei Querschnitten übliche Gewichtung mit Hilfe von sozio-ökonomischen Merkmalen, die in die inhaltlichen Analysen eingehen.

[11]Rendtel (1993) weist bei der Berechnung der Inklusionswahrscheinlichkeit darauf hin, daß eine querschnittsbezogene Randanpassung anhand externer statistischer Informationen über die Zahl der Geschiedenen in der Wohnbevölkerung zu Verzerrungen führt, da dadurch nicht nur Personen "hochgewichtet" würden, die während der SOEP-Laufzeit geschieden wurden, sondern auch bereits zuvor geschiedene Personen, für die kein erhöhtes Ausfallrisiko im Längsschnitt existiert. Für "umgezogene Haushalte" existiert eine solche Vergleichsstatistik nicht (da Umzüge in der amtlichen Statistik nur auf der Personenebene gezählt werden). Eine querschnittsbezogene Randanpassung zur Korrektur von ereignisbedingten Gewichtungsproblemen scheidet also aus. Eine längsschnittsbezogene Randanpassung ist schwierig, da es kaum Verlaufsstatistiken gibt, die als besserer Populationsschätzer anzusehen wären als das SOEP selbst.

telstandsbias" im SOEP gegeben[12].

Die folgende Analyse prüft daher, ob die Variablen, die Haushalte mit Niedrigeinkommen charakterisieren, durch die Längsschnittgewichtung des SOEP adäquat berücksichtigt werden.

Eine entsprechende Analyse kann für jede andere Fragestellung gezielt von jedem Nutzer der SOEP-Daten selbst durchgeführt werden (vgl. den SPSS Programm Code in Anhang A).

3.1 Zwei-Wellen-Analysen

Die von Lipsmeier (1993) vorgenommene Beschränkung der Analyse auf Ausfälle aufgrund von Teilnahmeverweigerungen durch die Befragten selbst ist methodisch sicherlich interessant (vgl. z.B. auch Pischner und Rendtel 1993), allerdings kommt es für Analysezwecke auch darauf an, ob ein Haushalt nicht teilnimmt, weil er vom Umfrageinstitut nicht wieder erreicht bzw. gefunden wurde. Bei der folgenden Replikation des Vorgehens von Lipsmeier beziehen wir deswegen in einem zweiten Schritt alle feldbedingten Ausfälle ein (Kontaktverlust und Teilnahmeverweigerung). Es sei vorweggenommen, daß diese "Ausdehnung" der Analyse die Resultate nur unwesentlich beeinflußt; freilich kommt es bei einer Längschnittsanalyse auch auf kleine Unterschiede an, da sich diese über mehrere Wellen hinweg kumulieren können.

Das Ausfallverhalten kann nur auf der Grundlage der letzten (oder weiter zurückliegender) Beobachtungen analysiert werden: So wird im Rahmen der folgenden Analysen - wie auch bei Andreß und Lipsmeier - das Haushaltsnettoeinkommen des Jahres $t-1$ als erklärende Variable gewählt; desweiteren der Bezug von laufender Hilfe zum Lebensunterhalt (HLU) im Vorjahr und die Betroffenheit

[12]Eine weitere Form des "Mittelstandsbias" glauben Leibfried und Voges (1992, S. 11) gefunden zu haben. Die Autoren werfen dem SOEP vor, daß es untere Einkommensschichten nicht überrepräsentiert (oversampled). Dadurch wäre im Vergleich zum US amerikanischen Panel PSID ein "Mittelschicht-Bias" gegeben. Leibfried und Voges übersehen freilich (wahrscheinlich geblendet von ihrer eigenwilligen Begrifflichkeit), daß die Stichprobe B des SOEP die "Gastarbeiter"-Population des Jahres 1984 mit einem überproportionalen Auswahlsatz enthält. Damit ist eine der wichtigsten Armutspopulationen im SOEP überrepräsentiert (vgl. z.B. Krause 1993). Weiterhin übersehen Leibfried und Voges, daß die PSID keineswegs ein zutreffendes Bild der Armen in den USA zeichnet. Zum einen ist es problematisch, das Niedrigeinkommens- Subsample in multivariate Analysen über Einkommen einzubeziehen, weil die endogene Selektion dieses Subsamples zu verfälschten Schätzergebnissen führen kann. Zum zweiten ist bei der PSID über zwanzig Wellen (=Jahre) lang keine Ergänzung der Stichprobe um Zuwanderer erfolgt. Immigranten stellen in den USA jedoch eine der wichtigsten "Armutspopulationen". Erst im Jahre 1990 erfolgte für die PSID eine arbiträre Ergänzung um die Zuwanderergruppe der "Hispanics"; Diese Gruppe ist allerdings keineswegs für alle Zuwanderer repräsentativ (vgl. Hill 1992, S. 10 und S. 60). Für das SOEP hingegen wird für das Jahr 1994 eine erste Zuwanderergänzung vorbereitet, die später regelmäßig wiederholt werden soll (vgl. Schulz et al. 1993).

von Arbeitslosigkeit[13]. Bei der Variablen "Alter des Haushaltsvorstandes" wird der für das aktuelle Jahr errechnete Wert genutzt.

Tabelle 1 zeigt die Ausfallwahrscheinlichkeiten von Haushalten für 2 Wellen-Längsschnitte nach ausgewählten Merkmalen (vgl. auch Anhang A für das benutzte SPSS-Programm). Es zeigt sich sowohl für die Analyse der Verweigerungen als auch für die Summe der Kontaktverluste und der Verweigerungen - wie von Lipsmeier (1993) dargestellt - eine höhere Ausfallrate für Haushalte im Niedrigeinkommensbereich (monatliches Haushaltsnettoeinkommen bis 1.000 DM) sowie für Haushalte, die diese Angabe verweigerten.

Der Bezug von laufender Hilfe zum Lebensunterhalt (HLU) und die Betroffenheit durch Arbeitslosigkeit sind bei dieser ungewichteten Betrachtung ebenfalls fast durchgängig über alle Wellen hinweg mit einer höheren Ausfallwahrscheinlichkeit verbunden. Junge und alte Haushaltsvorstände weisen - wie dies zu erwarten ist - ebenfalls überdurchschnittliche Ausfallraten auf.

Die Unterschiede in der Struktur der Ausfälle sind für die Unterscheidung "Verweigerung" sowie "Kontaktverlust plus Verweigerung" unbedeutend[14]. Für das spätere Analyseergebnis ist freilich allein "Kontaktverlust plus Verweigerung" relevant, d.h. daß alle Haushalte, die in der Vorwelle befragt wurden und nicht demographisch bedingt ausfielen, die Basis der Ausfallanalysen bilden.

In einem zweiten Schritt wird nun das Verhältnis R_Y (vgl. Gleichung (3)) für die unterschiedlichen Merkmalsausprägungen angegeben. Formal ergibt sich R_Y als merkmalsspezifischer Mittelwert der in der SOEP Datenbank bereitgestellten inversen Bleibewahrscheinlichkeit (s. Tabelle 2). Ein Wert unterhalb von 1 zeigt eine Untererfassung des entsprechenden Merkmalsträger an[15]. So nehmen z.B. 81,1% der Haushalte, die in Welle 2 die Einkommensangabe verweigerten, auch in Welle 3 wieder teil. Das Verhältnis R beträgt für dieses Merkmal 1.0034, es erfolgte also eine nahezu "perfekte" Modellierung des Ausfallprozesses.

Eine der wenigen Ausnahmen, bei denen der Erwartungswert einer Population deutlich über- bzw. unterschätzt wurde, ist die Gruppe der Haushalte mit Bezug von "Laufender Hilfe zum Lebensunterhalt" (HLU) in Welle 1. Der Wert für R_Y beträgt für Welle 2 nur 95,48%.

Mit Hilfe der in Abschnitt 2 diskutierten Methoden ist es möglich, eine Varianzschätzung dieses Wertes vorzunehmen (analytisch und/oder durch eine Randomisierung der Stichprobe). Hierbei zeigte sich, daß diese Abweichung um 4,5%-

[13]Theoretisch kann damit jedoch nicht ausgeschlossen werden, daß eine Einkommensveränderung seit der letzten Befragung auslösendes Moment für eine eventuelle Verweigerung war, und nicht das Einkommensniveau im Vorjahr. Gleiches gilt für HLU und Arbeitslosigkeit.

[14]Nur am Rande sei angemerkt, daß der quantitative Umfang der Kontaktverluste, bedingt durch die Lerneffekte des Umfrageinstituts, auf einen vernachläßigbaren Anteil von weniger als 1 Prozent zurückgegangen ist.

[15]Dies sind die inversen Teilnahmewahrscheinlichkeiten, die für Haushalte in den Variablen xHBLEIB abgelegt sind, wobei das Prefix x für die Wellenkennung steht. Diese Variable xHBLEIB hat für alle erfolgreich interviewten Respondenten, die auch im Vorjahr befragt werden konnten, einen Wert größer Null; sonst gleich Null.

TABELLE 1

AUSFALLWAHRSCHEINLICHKEITEN FÜR FOLGEWELLEN (2-WELLEN-LÄNGSSCHNITTE) IM SOEP

	Welle 2				Welle 3				Welle 4				Welle 5			
	Verweigerung		Verweigerung + Kontaktverlust		Verweigerung		Verweigerung + Kontaktverlust		Verweigerung		Verweigerung + Kontaktverlust		Verweigerung		Verweigerung + Kontaktverlust	
	N_1	vH	N_2	vH	N_1	vH	N_2	vH	N_1	vH	N_2	vH	N_1	vH	N_2	vH
Alle Haushalte	5937	10,4	6051	12,1	5459	8,8	5506	9,6	5197	5,1	5235	5,8	5125	7,0	5156	7,5
Haushalts-Netto-Einkommen																
Angabe verweigert	335	17,9	337	18,4	309	18,1	312	18,9	272	12,9	274	13,5	237	16,9	237	16,9
< 1000 DM	457	13,8	469	16,2	370	10,0	380	12,4	296	5,4	305	8,2	274	8,8	276	9,4
1000-2000 DM	1816	11,2	1857	13,2	1520	9,6	1533	10,4	1383	5,3	1392	5,9	1241	6,5	1249	7,1
2000-3000 DM	1713	8,3	1741	9,8	1569	7,5	1578	8,1	1466	4,2	1474	4,7	1402	6,5	1416	7,4
3000-4000 DM	992	9,6	1005	10,8	993	6,3	1001	7,1	1007	4,2	1011	4,6	1085	6,1	1089	6,4
> 4000 DM	624	8,2	641	10,6	698	8,6	702	9,1	773	5,1	779	5,8	886	6,4	889	6,8
Bezug von HLU	121	14,1	127	18,1	126	9,5	130	12,3	120	8,3	121	9,1	117	11,1	117	11,1
Alter des Haushaltsvorstands																
< 25 Jahre	234	17,1	248	21,8	257	14,0	268	17,5	266	13,5	276	16,7	259	13,1	269	16,4
25-35 Jahre	1110	8,9	1150	12,1	1031	7,7	1043	8,7	980	6,0	993	7,3	959	9,7	971	10,8
35-55 Jahre	2626	9,2	2666	10,5	2409	7,6	2424	8,2	2280	4,2	2293	4,8	2238	6,3	2245	6,6
55-65 Jahre	957	9,9	968	11,0	861	8,9	867	9,6	814	4,2	814	4,2	810	4,9	810	4,9
> 65 Jahre	1002	13,8	1011	14,5	894	11,4	896	11,6	848	4,4	850	4,6	854	5,7	855	5,9
Haushalte mit arbeitlos gemeldeten Personen	411	9,3	427	12,7	540	9,1	545	9,9	475	7,6	487	9,9	490	9,0	497	10,3

N_1 = Anzahl der Haushalte mit Kontaktaufnahme; N_2 = Anzahl der Haushalte mit Kontaktaufnahme sowie Feldverlust; vH = von Hundertsatz der Haushalte ohne realisiertes Interview (Basis N_1 bzw. N_2);

Basis: alle Teilnehmerhaushalte des Vorjahres sowie "neue" Haushalte; ohne demographische Abgänge wegen Tod, Auflösung und Umzug in das Ausland.

TABELLE 2

BLEIBEWAHRSCHEINLICHKEIT FÜR FOLGEWELLEN (2-WELLEN-LÄNGSSCHNITTE) IM SOEP

BASIS: HAUSHALTE MIT KONTAKTAUFNAHME PLUS FELDVERLUSTE;
OHNE DEMOGRAPHISCHE ABGÄNGE (TOD, AUFLÖSUNG UND UMZUG IN DAS AUSLAND)

	Welle 2			Welle 3			Welle 4			Welle 5		
	Teilnahme-Wahrscheinlichkeit	Bleibe-Wahrscheinlichkeit		Teilnahme-Wahrscheinlichkeit	Bleibe-Wahrscheinlichkeit		Teilnahme-Wahrscheinlichkeit	Bleibe-Wahrscheinlichkeit		Teilnahme-Wahrscheinlichkeit	Bleibe-Wahrscheinlichkeit	
		alle Haushalte	realisierte Haushalte		alle Haushalte	realisierte Haushalte		alle Hauhalte	realisierte Haushalte		alle Haushalte	realisierte Haushalte
Alle Haushalte	88,0	1,0007	1,1377	90,4	1,0084	1,1151	94,2	1,0061	1,0681	92,5	1,0059	1,088
Haushalts-Netto-Einkommen												
Angabe verweigert	81,6	1,0097	1,2374	81,1	1,0034	1,2374	86,5	1,0063	1,1634	83,1	1,0119	1,2174
< 1000 DM	83,8	0,9912	1,1824	87,6	0,9857	1,1248	91,8	0,9981	1,0872	90,6	1,0143	1,1198
1000-2000 DM	86,8	1,0027	1,1550	89,6	0,9962	1,1114	94,1	0,9998	1,0624	93,0	1,0024	1,0784
2000-3000 DM	90,2	1,0006	1,1089	91,2	1,0148	1,1036	95,3	1,0082	1,0577	92,6	0,9989	1,0789
3000-4000 DM	89,3	1,0004	1,1208	92,9	1,0268	1,1052	95,5	1,0125	1,0608	93,6	1,0050	1,0741
> 4000 DM	89,4	0,9976	1,1159	90,9	1,0086	1,1098	94,2	1,0079	1,0696	93,3	1,0190	1,0928
Bezug von HLU	81,9	0,9548	1,1660	87,7	0,9970	1,1369	90,9	0,9930	1,0923	88,9	1,0184	1,1457
Alter des Haushaltsvorstands												
< 25 Jahre	78,2	1,0218	1,3062	82,5	1,0218	1,2391	83,3	1,0247	1,2297	83,6	1,0370	1,2398
25-35 Jahre	87,9	0,9948	1,1316	91,3	1,0070	1,1032	92,8	1,0155	1,0949	89,2	0,9913	1,1114
35-55 Jahre	89,5	1,0036	1,1218	91,8	1,0174	1,1084	95,3	1,0017	1,0517	93,4	1,0024	1,0736
55-65 Jahre	89,1	1,0001	1,1231	90,4	0,9912	1,0962	95,8	1,0076	1,0515	95,1	1,0137	1,0664
> 65 Jahre	85,5	0,9961	1,1656	88,4	1,0009	1,1323	95,4	1,0036	1,0518	94,2	1,0148	1,0779
Haushalte mit arbeitslos gemeldeten Personen	87,4	1,0104	1,1567	90,1	1,0311	1,1445	90,1	0,9757	1,0823	89,7	0,9886	1,1016
N	6051	6051	5322	5506	5506	4979	5235	5235	4931	5156	5156	4767

Punkte vom Erwartungswert (100%) nicht signifikant ist (bei einer Irrtumswahrscheinlichkeit von 7%). Die für den SOEP-Nutzer einfach nachvollziehbare Varianzschätzung mit Hilfe der Randomisierung der Stichprobe wird im Anhang B dargestellt.

Die genaue Inspektion der Abweichungen der Längschnittsgewichtung vom Erwartungswert zeigte, daß keine einzige der in Tabelle 2 dargestellten Abweichungen signifikant ist. Dies ist auch insofern erwartbar, da einige der Tabellen-Merkmale explizit im Gewichtungsprozeß enthalten sind. Dies trifft für das Alter des Haushaltsvorstandes und die Einkommensklasse zu; nicht jedoch für den Bezug von HLU und für die Betroffenheit von Arbeitslosigkeit. Gleichwohl gelingt auch für diese Merkmale die Längschnittsgewichtung, da zum ersten die Merkmale HLU und Arbeitslosigkeit mit dem Merkmal "Niedrigeinkommen", das in die Gewichtung eingeht, hoch korreliert sind und zum zweiten die hohe räumliche Mobilität junger Sozialhilfeempfänger und Arbeitsloser ebenfalls in den Gewichtungsprozeß eingeht.

Abbildung 1a bis 1c zeigen die Ergebnisse der Gewichtung anschaulich. Die ungewichteten Teilnahmewahrscheinlichekeiten für Niedrigeinkommens-Haushalte, HLU-Bezieher und von Arbeitslosigkeit Betroffene liegen geringfügig unter dem Durchschnitt der Stichprobe. Nach Gewichtung mit den bereitgestellten Bleibewahrscheinlichkeiten liegen sie durchweg auf der 100-Prozent-Linie.

3.2 Fünf-Wellen-Analysen

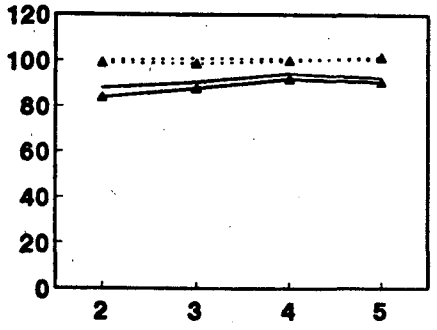
Über das von Andreß et al. (1993) aufgeworfene Analyseziel, das jeweils zwei Wellen des SOEP umfaßt, kann man jedoch auch Fragen behandeln, die größere Zeitintervalle analysieren und daher höhere Anforderungen an die Längsschnittgewichtung stellen. So mag die Frage von Interesse sein, welche ökonomische Position Niedrigeinkommenshaushalte der Startperiode 1 nach fünf Jahren haben. Geringfügige Fehler in der lediglich auf Folgewellen basierenden Längschnitts-Gewichtung können sich dann zu einer nennenswerten Abweichung kumulieren. Neben der in Abschnitt 3.1 gezeigten deskriptiven Prüfung wird für die 5-Wellen-Analysen zusätzlich eine Varianzabschätzung (vgl. Abschnitt 2) vorgenommen.

Tabelle 3 weist für die Haushalte der 1. Welle, die bis Welle 5 nicht demographisch bedingt (also durch Tod, Umzug in das Ausland oder Auflösung) ausfallen[16], den Verlauf der Teilnahmewahrscheinlichkeit (linker Block) in Abhängigkeit der bekannten Haushaltsmerkmale (aus Welle 1) aus. Es zeigt sich bei dieser ungewichteten Analyse ein ähnliches Bild wie in Tabelle 1 (2- Wellen-Längsschnitt): Haushalte, die in der ersten Welle die Angabe des Haushaltseinkommens verweigerten oder im Niedrigeinkommensbereich (< 1000 DM) lagen, haben mit

[16]Allerdings schließt diese Analyse auf Haushaltsebene neugegründete Haushalte bzw. Haushaltsabspaltungen von der Analyse aus. Dies bedeutet auf Personenebene, daß neben demographischen Ausfällen auch noch weitere Ausfälle existieren, die bei der Längsschnittgewichtung allerdings nicht als Ausfälle berücksichtigt wurden.

Abbildung 1: Abbildung 1a bis 1c

Bleibewahrscheinlichkeiten im SOEP vor und nach Längsschnittgewichtung Merkmal: HH mit Niedrigeinkommen in der Vorwelle



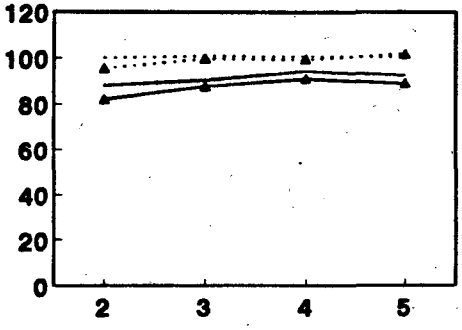
Alle HH (vor)	88	90,4	94,2	92,5
Alle HH (nach)	100,07	100,84	100,61	100,59
Niedrig-Eink. (vor)	83,8	87,6	91,8	90,6
Niedrig-Eink. (nach)	99,12	98,57	99,81	101,43

Welle

Hinweis: Um demographische Ausfälle (Tod, Umzug in das Ausland, Auflösung des HH) bereinigt.

Quelle: SOEP (West); Wellen 1 bis 5. DIW '93

Bleibewahrscheinlichkeiten im SOEP vor und nach Längsschnittgewichtung Merkmal: HH mit Bezug von HLU in der Vorwelle



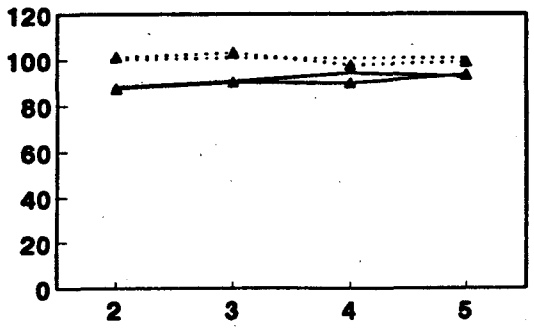
Alle HH (vor)	88	90,4	94,2	92,5
Alle HH (nach)	100,07	100,84	100,61	100,59
HLU-Bezug (vor)	81,9	87,7	90,9	88,9
HLU-Bezug (nach)	95,48	99,7	99,3	101,8

Welle

Hinweis: Um demographische Ausfälle (Tod, Umzug in das Ausland, Auflösung des HH) bereinigt.

Quelle: SOEP (West); Wellen 1 bis 5. DIW '93

Bleibewahrscheinlichkeiten im SOEP vor und nach Längsschnittgewichtung Merkmal: HH mit Arbeitslosigkeit in der Vorwelle



Alle HH (vor)	88	90,4	94,2	92,5
Alle HH (nach)	100,07	100,84	100,61	100,59
mit ALO (vor)	87,4	90,1	89,7	93,3
mit ALO (nach)	101,04	103,11	97,57	96,86

Welle

Hinweis: Um demographische Ausfälle (Tod, Umzug in das Ausland, Auflösung des HH) bereinigt.

Quelle: SOEP (West); Wellen 1 bis 5. DIW '93

Zunahme des Beobachtungszeitraumes deutlich niedrigere Teilnahmequoten als Haushalte mit höherem Einkommen. Bis zur 5. Welle beträgt die Teilnahmewahrscheinlichkeit der Niedrigeinkommens- Haushalte aus Welle 1 nur noch 60% und liegt damit deutlich unter dem Gesamtdurchschnitt von knapp 73%. Haushalte mit Bezug von HLU in Welle 1 sind 1988 (also in Welle 5) nur noch zu zwei Drittel enthalten.

Der rechte Block in Tabelle 3 zeigt, inwieweit die Gewichtung des SOEP diese unterschiedlichen Ausfallprozesse ausgleicht.

Während sich für die Gesamtpopulation Werte ergeben, die nur gering vom Erwartungswert 100 abweichen, wird besonders die Gruppe der Niedrigeinkommenshaushalte aus Welle 1 in der Längsschnittgewichtung unterschätzt (vgl. auch Abbildung 2a). Kumulativ bis Welle 5 liegt die durchschnittliche Bleibewahrscheinlichkeit dieser Population nach SOEP-Längsschnittgewichtung nur bei 92.6%[17].

Dieses Ergebnis verblüfft, da von Welle-zu-Welle die Gewichtung für Niedrigeinkommenshaushalte durchaus zufriedenstellend war (vgl. Tabelle 2 und Abbildung 1a bis 1c). Dieses Ergebnis hat 2 mögliche Interpretationen. Da nach fünf Wellen nur noch 20 Prozent der Niedrigeinkommenshaushalte aus Welle 1 in dieser niedrigen Einkommensklasse verblieben sind[18], werden 80% nicht mehr als Niedrigeinkommenshaushalte gewichtet. Diese Haushalte weisen in der ersten Interpretation ein unterdurchschnittliches Teilnahmeverhalten auf, das mit den übrigen Gewichtungsmerkmalen nicht kompensiert wird[19].

Die zweite Interpretation zielt auf die hier benutzte Definition von Längsschnitthaushalten. Diese schließt alle abgespaltenen (=erhebungstechnisch neue Haushalte) von der Analyse aus. Diese "Verluste" werden nicht durch die demographischen Verluste durch Tod und Fortzug ins Ausland abgedeckt. Es kann also sein, daß Haushalte mit Niedrigeinkommen häufiger von Haushaltsabspaltungen betroffen sind und die Person, die über ein geringes Einkommen verfügt, in vielen Fällen in den neuen Haushalt zieht. In diesem Fall ist die Definition eines Längsschnitthaushalts nicht neutral gegenüber dem betrachteten Merkmal.

Diese Überlegung gilt in noch stärkerem Maße für die Überrepräsentierung der Haushalte mit hohem Einkommen im Längsschnitt[20].

Korrespondierend zur niedrigsten Einkommensklasse scheinen die Haushalte mit einem etwas höheren Einkommen überschätzt zu werden. Für die Haus-

[17]Im Rahmen einer Varianzschätzung nach dem Random-Group-Ansatz ist jedoch auch diese Abweichung von 7.4%-Punkten vom Erwartungswert (noch) nicht signifikant.

[18]Die Tatsache, daß diese Gruppe stärker von demographischen Ausfällen bis Welle 5 betroffen ist (10%) als die Gesamtpopulation (4,5%) spielt hier keine Rolle, da die Analysepopulation um diese Abgänge bereinigt wurde.

[19]Sei es aufgrund von echter "unbeobachteter Heterogenität" oder nur aufgrund von prinzipiell im Datensatz vorhandenen, aber nicht für die Gewichtung benutzter Merkmale.

[20]Derartige Schwierigkeiten werden bei einer personenbezogenen Längsschnittanalyse vermieden.

TABELLE 3

LÄNGSSCHNITT-TEILNAHMEWAHRSCHEINLICHKEIT UND BLEIBEWAHRSCHEINLICHKEIT NACH HAUSHALTSMERKMALEN

BASIS: ALLE BEFRAGUNGSHAUSHALTE AUS WELLE 1 OHNE DEMOGRAPHISCHE ABGÄNGE BIS WELLE 5

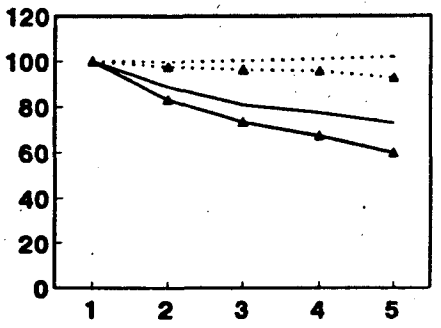
	N	Teilnahmewahrscheinlichkeit (in %)					Bleibewahrscheinlichkeit[1] (in %)				
		Welle 1	Welle 2	Welle 3	Welle 4	Welle 5	Welle 1	Welle 2	Welle 3	Welle 4	Welle 5
Alle Haushalte	5662	100	88,8	81,0	77,2	72,8	100	99,8	100,3	100,9	101,8
Haushalts-Netto-Einkommen											
Angabe verweigert	318	100	81,8	73,6	67,6	63,2	100	99,5	101,5	99,9	102,8
< 1000 DM	434	100	83,2	73,3	67,3	59,7	100	97,6	96,3	95,6	92,6
1000-2000 DM	1744	100	87,3	79,1	76,1	72,3	100	100,2	100,8	102,4*	104,3
2000-3000 DM	1655	100	90,9	83,2	79,5	74,7	100	99,8	100,0	100,6	100,6
3000-4000 DM	929	100	91,1	85,5	81,5	78,3	100	100,1	102,2	102,1	104,4*
> 4000 DM	582	100	91,6	82,8	79,2	74,6	100	99,9	99,0	99,5	100,3
Bezug von HLU	114	100	85,1	79,0	70,2	66,7	100	97,5	99,7	94,3	100,2
Alter des Haushaltsvorstands											
< 25 Jahre	145	100	84,8	75,9	71,7	66,2	100	96,6	96,1	102,0	104,4
25-35 Jahre	1043	100	90,1	82,7	77,7	71,8	100	99,7	99,8	100,2	100,0
35-55 Jahre	2605	100	89,8	82,8	79,2	74,7	100	100,6	101,9	102,4*	102,8
55-65 Jahre	924	100	88,9	80,6	77,4	74,6	100	99,3	99,1	99,6	102,3
> 65 Jahre	938	100	85,2	75,3	71,5	67,7	100	98,8	98,5	98,6	100,7
Haushalte mit arbeitslos gemeldeten Personen	378	100	90,2	82,5	78,0	73,5	100	102,1	103,1	103,2	104,3

* signifikante Abweichung vom Mittelwert auf dem 93%-Niveau

[1] Verhältnis von geschätzten (Welle t) und tatsächlichen (Welle 1) Merkmalsträgern.

Abbildung 2: Abbildung 2a bis 2c

Bleibewahrscheinlichkeiten im SOEP vor und nach Längsschnittgewichtung Merkmal: HH mit Niedrigeinkommen in Welle 1



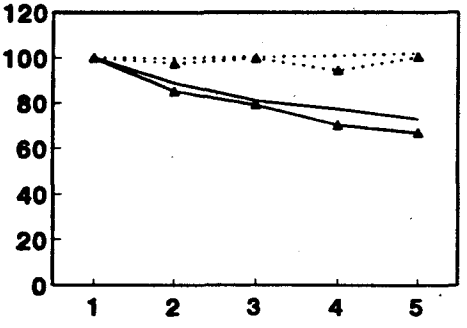
Alle HH (vor)	100	88,8	81	77,2	72,8
Alle HH (nach)	100	99,8	100,3	100,9	101,8
Niedrig-Eink. (vor)	100	83,2	73,3	67,3	59,7
Niedrig-Eink. (nach)	100	97,6	96,3	95,6	92,6

Welle

Hinweis: Um demographische Ausfälle (Tod, Umzug in das Ausland, Auflösung des HH) bereinigt.

Quelle: SOEP (West); Wellen 1 bis 5. DIW '93

Bleibewahrscheinlichkeiten im SOEP vor und nach Längsschnittgewichtung Merkmal: HH mit Bezug von HLU in Welle 1



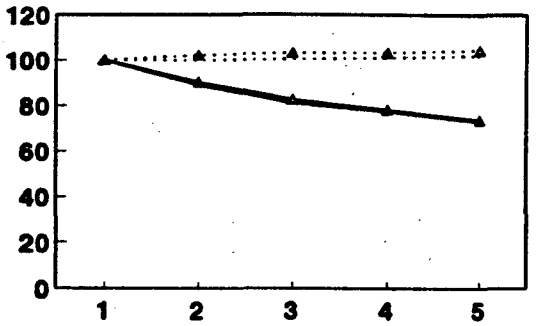
Alle HH (vor)	100	88,8	81	77,2	72,8
Alle HH (nach)	100	99,8	100,3	100,9	101,8
HLU-Bezug (vor)	100	85,1	79	70,2	66,7
HLU-Bezug (nach)	100	97,5	99,7	94,3	100,2

Welle

Hinweis: Um demographische Ausfälle (Tod, Umzug in das Ausland, Auflösung des HH) bereinigt.

Quelle: SOEP (West); Wellen 1 bis 5. DIW '93

Bleibewahrscheinlichkeiten im SOEP vor und nach Längsschnittgewichtung Merkmal: HH mit Arbeitslosigkeit in Welle 1



Alle HH (vor)	100	88,8	81	77,2	72,8
Alle HH (nach)	100	99,8	100,3	100,9	101,8
mit ALO (vor)	100	90,2	82,5	78	73,5
mit ALO (nach)	100	102,1	103,1	103,2	104,3

Welle

Hinweis: Um demographische Ausfälle (Tod, Umzug in das Ausland, Auflösung des HH) bereinigt.

Quelle: SOEP (West); Wellen 1 bis 5. DIW '93

halte mit Einkommen von 1000 bis 2000 DM ergibt sich nach fünf Wellen eine durchschnittliche Bleibewahrscheinlichkeit von 104.3%. Auch diese Abweichung ist nach dem Randomisierungsansatz nicht signifikant. Die Haushalte mit Einkommen zwischen 3000 und 4000 DM hingegen werden mit 104.4% signifikant überschätzt (vgl. Anhang B).

Die in Abschnitt 2 beschriebene analytische Varianzschätzung, vgl. Gleichung (9), vereinfacht sich für die Analyse von Längsschnitthaushalten. In diesem Fall sind alle Ziehungen von Längsschnitthaushalten unabhängig, so daß sich Gleichung (8) zu

$$\hat{V}(\hat{P}_Y) = \sum_{i \in S} \frac{1}{\pi_i} \left(\frac{1}{\pi_i} - 1 \right) Y_i$$

vereinfacht. Hierbei wurde $Y_i^2 = Y_i$ für 0/1 Merkmale benutzt. Die Standardabweichung von R_Y ist damit durch

$$\frac{\sqrt{\hat{V}(\hat{R}_Y)}}{P_Y}$$

gegeben. Im Falle der Niedrigeinkommen erhält man beispielsweise den Wert

$$\frac{\sqrt{264.809}}{434} = 0.0375$$

Bei Verwendung der 2-Sigma-Regel ergibt sich für die gesamte Population eine durchschnittliche Bleibewahrscheinlichkeit von 92.6% im Vertrauensintervall (85.1%; 100.1%). Beim Erwartungswert von 100% erfolgt also noch keine signifikante Unterschätzung. Der empirisch festgestellte Wert für die Gesamtpopulation (s. Tabelle 3 rechts) liegt mit 101.8% jedoch schon über dem Erwartungswert und wird vom genannten Konfidenzintervall nicht überdeckt: Die Niedrigeinkommenshaushalte werden demnach signifikant unterschätzt.

Da eines der beiden verwendeten Verfahren zur Abschätzung der Varianz signifikante Abweichungen aufzeigt, bietet sich eine Korrektur der Längsschnittgewichtung für die 5-Wellen-Analyse an (s. Abschnitt 4).

3.3 Analyse der Randverteilungen

Bevor in Abschnitt 4 auf die Methode der Korrektur der SOEP-Längsschnittgewichte eingegangen wird, sei darauf hingewiesen, daß selbst die signifikanten Abweichungen in R_Y über fünf Wellen hinweg noch nicht zu signifikanten Abweichungen in den eigentlich interessierenden Randverteilungen führen müssen, da diese bereits in der ersten Welle einem Stichprobenfehler unterliegen, während die Bleibewahrscheinlichkeiten auf die erste Welle konditioniert sind.

Tabelle 4 stellt die Auswirkungen der Längsschnittgewichtung mit den im SOEP bereitgestellten Bleibewahrscheinlichkeiten auf Populationsschätzungen dar. Idealtypisch sollte die Verteilung der interessierenden Merkmale aus Welle 1 auch nach 5 Wellen noch erhalten sein (da Gestorbene und ins Ausland Verzogene ausgeschlossen werden und für jede Welle nach den Merkmalen von Welle 1 klassifiziert wird). Für die Gruppe der Niedrigeinkommensbezieher, die - nach Gewichtung - eine signifikante Unterschätzung der Bleibewahrscheinlichkeit von ca. 7%-Punkten bis Welle 5 erfahren (siehe Tabelle 3, rechter Block), zeigt sich entsprechend eine Abnahme des Anteils an allen Haushalten von 7,7% auf 7,0%.

Diese Abnahme des Bestands um 0.7 Prozentpunkte sieht weniger dramatisch aus als es vielleicht die vorhergehenden Resultate suggerieren. Man sollte sich vor Augen führen, daß man letztendlich an den in Tabelle 4 gezeigten Merkmalsverteilungen interessiert ist. Der Bezug auf die wesentlich größere Referenzzahl aller Haushalte glättet viele Unterschiede, die bei Bezug auf die kleineren Fallzahlen innerhalb der Merkmale noch als sehr groß erscheinen.

Dieser relative Bedeutungsverlust über die 5 Wellen hinweg ist zwar stetig, jedoch (noch) nicht signifikant, wenn man berücksichtigt, daß das Ergebnis für die 1. Welle auch einen Zufallsfehler aufweist, wodurch das 93%-Vertrauensintervall auch den Wert 7.0% einschließt [21]. Darüber hinaus zeigt sich für die Gruppe der Haushalte mit Bezug von HLU in Welle 1 eine adäquate Modellierung auch über 5 Wellen: Der Anteil dieser Haushalte an der Gesamtpopulation beträgt konstant 2%.

3.4 Zwischenergebnis

Zusammenfassend läßt sich sagen, daß die interne Validierung der Längsschnittgewichtung des SOEP nur in Ausnahmefällen nicht zu den erwarteten Ergebnissen führt: Für "kurze" Längsschnitte (2-Wellen) zeigen sich nach Gewichtung keine signifikanten Abweichungen, während in einem Längsschnitt über 5 Wellen lediglich für das Merkmal "Niedrigeinkommensbezieher" signifikante Unterschätzungen gefunden wurden. Diese können Ausgangspunkt für eine Korrektur sein.

[21] Tabelle 4 berücksichtigt nur den Einfluß der Längsschnittgewichtung mit Hilfe der im SOEP bereitgestellten Bleibewahrscheinlichkeiten xHBLEIB. Ist man an den Populationsschätzungen interessiert, die sich nach Gewichtung und Hochrechnung ergeben, so muß für Welle 1 mit dem Querschnittshochrechnungsfaktor AHHRF und für die Wellen 2 bis 5 mit dem jeweiligen Längsschnitthochrechnungsfaktor AxHHRF gerechnet werden. Die anschließende Varianzabschätzung mit Hilfe der Variablen HRGROUP (SOEP-Standardrandomisierungsvariable) liefert in diesem Fall ebenfalls keine signifikante Unterschätzung der Niedrig-Einkommenshaushalte der ersten Welle bis Welle 5.

TABELLE 4

GEWICHTETE VERTEILUNG VON HAUSHALTSBEZOGENEN MERKMALEN IM LÄNGSSCHNITT (WELLE 1 BIS 5)

WELLE 1: UNGEWICHTETE VERTEILUNG DER SOEP-STICHPROBE
WELLE 2 BIS 5: LÄNGSSCHNITTGEWICHTETE VERTEILUNG DER NOCH IN DER STICHPROBE VERBLIEBENEN HAUSHALTE

BASIS: ALLE BEFRAGUNGSHAUSHALTE AUS WELLE 1 OHNE DEMOGRAPHISCHE ABGÄNGE BIS WELLE 5

	N	Anteil (in %)				
		Welle 1	Welle 2	Welle 3	Welle 4	Welle 5
Alle Haushalte	5662	100	100	100	100	100
Haushalts-Netto-Einkommen						
Angabe verweigert	318	5,6	5,6	5,7	5,6	5,7
< 1000 DM	434	7,7	7,5	7,4	7,3	7,0
1000-2000 DM	1744	30,8	30,9	30,9	31,3	31,6
2000-3000 DM	1655	29,2	29,2	29,1	29,2	28,9
3000-4000 DM	929	16,4	16,5	16,7	16,6	16,8
> 4000 DM	582	10,3	10,3	10,1	10,1	10,1
Bezug von HLU	114	2,0	2,0	2,0	1,9	2,0
Alter des Haushaltsvorstands						
< 25 Jahre	145	2,6	2,5	2,5	2,6	2,6
25-35 Jahre	1043	18,4	18,4	18,3	18,3	18,1
35-55 Jahre	2605	46,0	46,4	46,7	46,7	46,4
55-65 Jahre	924	16,3	16,2	16,1	16,1	16,4
> 65 Jahre	938	16,6	16,4	16,3	16,2	16,4
Haushalte mit arbeitslos gemeldeten Personen	378	6,7	6,8	6,9	6,8	6,8

4 Eine exemplarische Korrektur der Längsschnittgewichtung

Durch eine Korrektur der Längsschnittgewichtung des SOEP sollte sich die Summe aller betrachteten Einheiten nach Gewichtung nicht ändern[22], d.h., die Längsschnittpopulation bleibt von Welle 1 bis Welle 5 konstant.

Will man die Gruppe der Haushalte, die in Welle 1 weniger als 1000 DM Einkommen angaben, in der korrigierten kumulativen Längschnittbetrachtung bis Welle 5 gewichten, so ist dies auch für alle anderen Haushalten sinnvoll, um die Gesamtfallzahl nicht zu verändern. Je nach Fragestellung ist für jede Welle oder nur für die letzte betrachtete Welle eine Umgewichtung durchführbar. Da es hier nur um die Demonstration des Prinzips geht, nehmen wir nur eine Korrektur für Welle 5 vor.

Realisiert wird diese "Feinanpassung", indem für jeden Haushalt die kumulative Bleibewahrscheinlichkeit bis Welle 5 mit der Inversen der gruppenspezifischen, in diesem Beispiel einkommensklassenspezifischen, Bleibewahrscheinlichkeit als Korrekturfaktor multipliziert wird, vgl. Gleichung (16) in Abschnitt 2. Wie nicht anders zu erwarten ist, ergibt sich dann für die in Tabelle 5 dargestellte Randverteilung eine perfekte Übereinstimmung mit Welle 1 (vgl. Anhang A für den SPSS Code). Bei dieser Vorgehensweise, die sich im genannten Beispiel nur auf die Einkommensklassenzugehörigkeit bezieht, ist zu beachten, daß die Verteilungen bzw. Populationsschätzungen anderer Variablen sich durchaus verschieben können. Eine - auch dieses Problem berücksichtigende - alternative Korrektur der Längsschnittgewichtung setzt daher auf einer neuen Schätzung der Ausfallwahrscheinlichkeiten bezüglich der interessierenden Merkmale auf[23].

[22]Würde man statt einer einfachen Korrektur eine verfeinerte multiple Ausfallanalyse vornehmen (wie sie der SOEP-Gewichtung zugrundeliegt; vgl. Rendtel 1991, Pischner und Rendtel 1993), würde ein zusätzliches erklärendes Merkmal in der anschließenden Gewichtung die Summe der Einheiten ebenfalls nicht ändern.

[23]Auch diese Berechnungen sind mittels einer einfachen Logitschätzung mit SPSS oder SAS durchführbar.

Tabelle 5 Verteilung der Haushaltsnetto-Einkommen im Längsschnitt Welle 1-5 vor und nach Korrektur der Bleibewahrscheinlichkeiten (Anteil in %)

Haushalts-Netto-Einkommen	Welle 1	Anteil (in %) Welle 5	
	Ungewichtet	Vor Korrektur	Nach Korrektur
Alle Haushalte	100	100	100
Angabe verweigert	5,6	5,7	5,6
< 1000 DM	7,7	7,0	7,7
1000-2000 DM	30,8	31,6	30,8
2000-3000 DM	29,2	28,9	29,2
3000-4000 DM	16,4	16,8	16,4
> 4000 DM	10,3	10,1	10,3
N	5662	5766	5662

5 Fazit

Die bestehende Gewichtung des SOEP führt bei einigen Merkmalen zu signifikanten Abweichungen von erwarteten Fallzahlen. So im Falle von Scheidungen, da Abhängigkeiten zwischen befragten Haushalten noch nicht berücksichtigt werden sowie im zweiten Jahr nach einem Umzug.

Für die Analyse von Niedrigeinkommensbeziehern konnte gezeigt werden, daß die im SOEP- Datensatz zur Verfügung gestellten Gewichtungsvariablen für fast alle betrachteten Subgruppen und Analysezeiträume zu unverzerrten Populationsschätzungen führen. Nur bei Längsschnittdatensätzen wurden mit zunehmender Beobachtungsdauer Ergebnisse erzielt, die Zweifel an der korrekten Modellierung der Ausfallprozesse wecken. Allerdings können die hier präsentierten Abweichungen in den Schätzwerten von Haushalten im Niedrig-Einkommensbereich auch das Ergebnis der Selektion auf Längsschnitthaushalte sein.

Mit Hilfe einer relativ einfachen endogenen Prüfung können signifikante Abweichungen erkannt und ausgeglichen werden.

Literatur

Andreß, Hans-Jürgen et al. 1993: Zwischenbericht des DFG- Projektes "Versorgungsstrategien privater Haushalte im unteren Einkommensbereich", Bielefeld.

Deville, Jean-Claude und Carl-Eric Särndal 1992: Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, S. 376-382.

Frick, Joachim 1993: Demographische Ereignisse im Haushaltskontext als Determinante räumlicher Mobilität, DIW Diskussionspapier Nr. 63, Berlin.

Gokhale, und Solomon Kullback 1978: The Information in Contingency Tables, New York.

Hill, Martha S. 1992: The Panel Study of Income Dynamics - A User's Guide, Newbury Park u.a.

Ireland C. und Solomon Kullback 1968: Contingency Tables with given marginals, Biometrica 55, S. 179-188.

Krause, Peter 1992: Einkommensarmut in der Bundesrepublik Deutschland, in: Aus Politik und Zeitgeschichte (Beilage zur Wochenzeitung "Das Parlament") B. 49/92, S. 3-17.

Leibfried, Stephan und Wolfgang Voges 1992: Vom Ende einer Ausgrenzung? - Armut und Soziologie, in dies. (Hrsg.): Armut im modernen Wohlfahrtsstaat, Sonderheft der Kölner Zeitschrift für Soziologie und Sozialpsychologie, S. 9-33.

Lipsmeier, Gero 1993: Zur Repräsentation des unteren Einkommensbereiches im Sozio-ökonomischen Panel (SOEP), Arbeitspapier Nr. 10 des DFG-Projektes "Versorgungsstrategien privater Haushalte im unteren Einkommensbereich", Bielefeld.

Little, Roderick und Mei-Miau Wu 1991: Models for Contingency Tables with Known Margins when Target and Sampled Population Differ. Journal of the American Statistical Association, 86, S. 87–95.

Pischner, Rainer und Ulrich Rendtel 1993: Quer- und Längsschnittgewichtung des Sozio-ökonomischen Panels, DIW Diskussionspapier Nr. 69, Berlin.

Pötter, Ulrich und Ulrich Rendtel 1993: Über Sinn und Unsinn von Repräsentativitätsstudien, in Allgemeines Statistisches Archiv, Bd. 77, S.260-280.

Projektgruppe Panel 1993: Das Sozio-ökonomische Panel (SOEP) nach zehn Jahren, in: Vierteljahreshefte zur Wirtschaftsforschung, Heft 1-2.

Rendtel, Ulrich 1991: Die Schätzung von Populationswerten in Stichprobenerhebungen, in Allgemeines Statistisches Archiv, 75(3), S. 225-244.

Rendtel, Ulrich 1993: Über die Repräsentativität von Panelstichproben - Eine Analyse der feldbedingten Ausfälle im Sozio-ökonomischen Panel (SOEP), DIW Diskussionspapier Nr. 70, Berlin.

Rendtel, Ulrich und Ulrich Pötter 1993: Empirie ohne Daten - Kritische Anmerkungen zu einer Repräsentativitätsstudie über den Allbus, in Kölner Zeitschrift für Soziologie und Sozialpsychologie, 45 Jg., Heft 2, S. 350-358.

Rendtel, Ulrich und Ulrich Pötter 1993b: Über Sinn und Unsinn von Repräsentativitätsstudien, in Allgemeines Statistisches Archiv, Bd. 77, S.260-280.

Schnell, Rainer 1993: Die Homogenität sozialer Kategorien als Voraussetzung für "Repräsentativität" und Gewichtungsverfahren, in Zeitschrift für Soziologie, 22(1), S. 16-32.

Schulz, Erika, Ulrich Rendtel, Jürgen Schupp und Gert Wagner 1993: Zur Problematik von Zuwanderer-Ergänzungen in Wiederholungsbefragungen am Beispiel des Sozio-ökonomischen Panels (SOEP), DIW Diskussionspapier Nr. 71, Berlin.

Wolter, Kirk 1985: Introduction to Variance Estimation, New York.

A SPSS Programm Code zur Erzeugung der Tabellen 1 bis 4

```
fILE HANDLE ALT / NAME='xxx sys a'
GET FILE = ALT

recode ALL (sysmis =-4)
MISSING VALUES ALL ()

* Zufallsgruppen ****
compute zufgr1 = rnd(uniform(8))
if (zufgr1 eq 0) zufgr1 = 8

value lables hlu1 hlu2 hlu3 hlu4 hlu5
             (1) ja (0) nein

compute kontakt2 = 0
compute kontakt3 = 0
compute kontakt4 = 0
compute kontakt5 = 0
var   labels kontakt2 'nur Kontakt; ohne Feldverlust'
           / kontakt3 'nur Kontakt; ohne Feldverlust'
           / kontakt4 'nur Kontakt; ohne Feldverlust'
           / kontakt5 'nur Kontakt; ohne Feldverlust'
value labels kontakt2 kontakt3 kontakt4 kontakt5
 (0) nix (1) teiln. im VJ (2) neuer HH (3) Ausfall im VJ

* ****************************************************************
* *** ACHTUNG: neue HH (_HTYP=5), deren Ur-HH im VJ nicht mitgemacht
               hat, werden auf kontakt=3 (AUSFALL im VJ) gesetzt, da
               fuer diese HH die Vars aus dem Vorjahr nicht besetzt
               sind: ALOZ_ HHY_ HLU_
* ****************************************************************

do if (bhergs>=0 and bhergs<=4)
+ compute        kontakt2 = 1
+ if (bhtyp = 5) kontakt2 = 2
end if
do if (chergs>=0 and chergs<=4)
+ if (chtyp le 4                    ) kontakt3 = 1
+ if (chtyp = 5                     ) kontakt3 = 2
+ if (bhergs gt 1 and chtyp  le 4) kontakt3 = 3
+ if (chtyp = 5 and hhy2=-4 and hlu2=-4) kontakt3 = 3
end if
do if (dhergs>=0 and dhergs<=4)
+ if (dhtyp le 4                    ) kontakt4 = 1
+ if (dhtyp = 5                     ) kontakt4 = 2
+ if (chergs gt 1 and dhtyp  le 4) kontakt4 = 3
+ if (dhtyp = 5 and hhy3=-4 and hlu3=-4) kontakt4 = 3
end if
do if (ehergs>=0 and ehergs<=4)
+ if (ehtyp le 4                    ) kontakt5 = 1
+ if (ehtyp = 5                     ) kontakt5 = 2
+ if (dhergs gt 1 and ehtyp  le 4) kontakt5 = 3
+ if (ehtyp = 5 and hhy4=-4 and hlu4=-4) kontakt5 = 3
end if
```

```
compute kont2 = 0
compute kont3 = 0
compute kont4 = 0
compute kont5 = 0
var   labels kont2 'Kontakt plus Feldverlust'
           / kont3 'Kontakt plus Feldverlust'
           / kont4 'Kontakt plus Feldverlust'
           / kont5 'Kontakt plus Feldverlust'
value labels kont2 to kont5
 (0) nix (1) teiln. im VJ (2) neuer HH (3) Ausfall im VJ

* ****************************************************************
* *** ACHTUNG: neue HH (_HTYP=5), deren Ur-HH im VJ nicht mitgemacht
              hat, werden auf kontakt=3 (AUSFALL im VJ) gesetzt, da
              fuer diese HH die Vars aus dem Vorjahr nicht besetzt
              sind: ALOZ_ HHY_ HLU_
* ****************************************************************

do if ((bhergs>=0 and bhergs<=4) or bhergs ge 8)
+ compute        kont2 = 1
+ if (bhtyp = 5) kont2 = 2
end if
do if ((chergs>=0 and chergs<=4) or chergs ge 8)
+ if (chtyp le 4                  ) kont3 = 1
+ if (chtyp = 5                   ) kont3 = 2
+ if (bhergs gt 1 and chtyp  le 4) kont3 = 3
+ if (chtyp = 5 and hhy2=-4 and hlu2=-4) kont3 = 3
end if
do if ((dhergs>=0 and dhergs<=4) or dhergs ge 8)
+ if (dhtyp le 4                  ) kont4 = 1
+ if (dhtyp = 5                   ) kont4 = 2
+ if (chergs gt 1 and dhtyp  le 4) kont4 = 3
+ if (dhtyp = 5 and hhy3=-4 and hlu3=-4) kont4 = 3
end if
do if ((ehergs>=0 and ehergs<=4) or ehergs ge 8)
+ if (ehtyp le 4                  ) kont5 = 1
+ if (ehtyp = 5                   ) kont5 = 2
+ if (dhergs gt 1 and ehtyp  le 4) kont5 = 3
+ if (ehtyp = 5 and hhy4=-4 and hlu4=-4) kont5 = 3
end if

compute hhy1no =  0
compute hhy2no =  0
compute hhy3no =  0
compute hhy4no =  0
compute hhy5no =  0
if (hhy1 = -1) hhy1no = 1
if (hhy2 = -1) hhy2no = 1
if (hhy3 = -1) hhy3no = 1
if (hhy4 = -1) hhy4no = 1
if (hhy5 = -1) hhy5no = 1
recode hhy1 hhy2 hhy3 hhy4 hhy5
  (0 thru 1000=1)(1001 thru 2000=2)(2001 thru 3000=3)
  (3001 thru 4000=4)(4001 thru hi=5)(-4=-4)(-1=-1)(else=-2)
  into hhy1k hhy2k hhy3k hhy4k hhy5k
recode althv1 althv2 althv3 althv4 althv5
  (15 thru 24=1)(25 thru 34=2)(35 thru 54=3)
  (55 thru 64=4)(65 thru hi =5)(else=-2)
  into althv1k althv2k althv3k althv4k althv5k
```

```
recode
   aloz1 aloz2 aloz3 aloz4 aloz5 (1 thru hi=1)

compute raus2 = 1
compute raus3 = 1
compute raus4 = 1
compute raus5 = 1
if (bhergs=1 or bhergs=0) raus2 = 0
if (chergs=1 or chergs=0) raus3 = 0
if (dhergs=1 or dhergs=0) raus4 = 0
if (ehergs=1 or ehergs=0) raus5 = 0
compute teil2 = 0
compute teil3 = 0
compute teil4 = 0
compute teil5 = 0
if (bhergs=1 or bhergs=0) teil2 = 1
if (chergs=1 or chergs=0) teil3 = 1
if (dhergs=1 or dhergs=0) teil4 = 1
if (ehergs=1 or ehergs=0) teil5 = 1

compute teil12 = -4
compute teil13 = -4
compute teil14 = -4
compute teil15 = -4
var labels teil12 'kumul. Teilnahmewahrscheinlichkeit Welle 1-2'
         / teil13 'kumul. Teilnahmewahrscheinlichkeit Welle 1-3'
         / teil14 'kumul. Teilnahmewahrscheinlichkeit Welle 1-4'
         / teil15 'kumul. Teilnahmewahrscheinlichkeit Welle 1-5'
if (a=1) teil12 = 0
if (a=1) teil13 = 0
if (a=1) teil14 = 0
if (a=1) teil15 = 0
if ((bhergs ge 5 and bhergs le 7)) teil12 = -2
if ((bhergs ge 5 and bhergs le 7) or
    (chergs ge 5 and chergs le 7)) teil13 = -2
if ((bhergs ge 5 and bhergs le 7) or
    (chergs ge 5 and chergs le 7) or
    (dhergs ge 5 and dhergs le 7)) teil14 = -2
if ((bhergs ge 5 and bhergs le 7) or
    (chergs ge 5 and chergs le 7) or
    (dhergs ge 5 and dhergs le 7) or
    (ehergs ge 5 and ehergs le 7)) teil15 = -2
if (a=1 and b=1                          ) teil12 = 1
if (a=1 and b=1 and c=1                  ) teil13 = 1
if (a=1 and b=1 and c=1 and d=1          ) teil14 = 1
if (a=1 and b=1 and c=1 and d=1 and e=1) teil15 = 1

* Variablen fuer logistische Regression *****
compute alt11 = 0
compute alt12 = 0
compute alt13 = 0
compute alt14 = 0
compute alt15 = 0
if (althv1k = 1) alt11 = 1
if (althv1k = 2) alt12 = 1
if (althv1k = 3) alt13 = 1
if (althv1k = 4) alt14 = 1
if (althv1k = 5 or althv1 le 0) alt15 = 1
```

```
compute hhy1low = 0
if (hhy1 gt 0 and hhy1 le 1000) hhy1low = 1

*  Korrekturfaktoren fuer Laengsschnitt Welle 1 bis 5
     auf der Basis gruppenspezifischer Abweichungen
compute korr15 = 1
if (hhy1k =-1) korr15 =  .973
if (hhy1k = 1) korr15 = 1.08
if (hhy1k = 2) korr15 =  .958
if (hhy1k = 3) korr15 =  .994
if (hhy1k = 4) korr15 =  .958
if (hhy1k = 5) korr15 =  .997

compute bleib12 = 0
compute bleib13 = 0
compute bleib14 = 0
compute bleib15 = 0
compute bleib15a= 0
var labels bleib12  'kumul. Bleibewahrscheinlichkeit Welle 1-2'
         / bleib13  'kumul. Bleibewahrscheinlichkeit Welle 1-3'
         / bleib14  'kumul. Bleibewahrscheinlichkeit Welle 1-4'
         / bleib15  'kumul. Bleibewahrscheinlichkeit Welle 1-5'
         / bleib15a 'kumul. Bleibewlk. Welle 1-5 KORRIGIERT '

missing values bleib12 to bleib15a (-2)
do if (a=1        )
+  compute bleib12  = bhbleib
+  compute bleib13  = bhbleib*chbleib
+  compute bleib14  = bhbleib*chbleib*dhbleib
+  compute bleib15  = bhbleib*chbleib*dhbleib*ehbleib
+  compute bleib15a = bhbleib*chbleib*dhbleib*ehbleib * korr15
end if

compute tebl12 = teil12*bleib12
compute tebl13 = teil13*bleib13
compute tebl14 = teil14*bleib14
compute tebl15 = teil15*bleib15
var labels tebl15 'Teilnahmewlk. Welle 1-5 * Bleibewlk. Welle 1-5'

* ****************************************************************
  abh. Variable fuer log. Regr. auf Verbleib im SOEP

compute drin15 = 0
if (bleib15 gt 0) drin15 = 1
```

```
* ***************************************************************
  Variable fuer theoretische Varianzabschaetzung
  Wurzel aus der SUMME von VAR_ dividiert durch Zahl der Merkmalstraeger
  BSP: SQRT(Summe von VARLOWY1) / 433  (im Falle von Welle 1-5)
     = 0.0375  ---> bei 92,6 % Verbleib nach Gewichtung und Anwendung
                    der 2-Sigma Regel ergibt sich immer noch eine
                    signifikante Unterschaetzung im Vergleich zum
                    Populationsmittelwert von 101.84%
                    weil: 92,6 + (2*0.0375) = 100.1  <  101.84

compute varhlu1  = (bleib15-1)*bleib15*hlu1
compute varlowy1 = (bleib15-1)*bleib15*hhy1low
compute varalo1  = (bleib15-1)*bleib15*aloz1
compute varhhy1n = (bleib15-1)*bleib15*hhy1no
compute varalt11 = (bleib15-1)*bleib15*alt11
compute varalt12 = (bleib15-1)*bleib15*alt12
compute varalt13 = (bleib15-1)*bleib15*alt13
compute varalt14 = (bleib15-1)*bleib15*alt14
compute varalt15 = (bleib15-1)*bleib15*alt15
```

```
* *** AUSWERTUNGEN *** ****************************************


* *** Tabellen 1 und 2 ***

* *** Abgrenzung: Verweigerung plus Feldverluste * ***
*  Bleibewahrscheinlichkeiten fuer realisierte HH nach HH-Merkmalen *
                                    ==========
temporary
select if (kont2 ge 1 and kont2 le 2 and b=1)
breakdown bhbleib  by   hhy1no hhy1k althv2k aloz1 hlu1  by  zufgr1

temporary
select if (kont3 ge 1 and kont3 le 2 and c=1)
breakdown chbleib  by   hhy2no hhy2k althv3k aloz2 hlu2  by  zufgr1

temporary
select if (kont4 ge 1 and kont4 le 2 and d=1)
breakdown dhbleib  by   hhy3no hhy3k althv4k aloz3 hlu3  by  zufgr1

temporary
select if (kont5 ge 1 and kont5 le 2 and e=1)
breakdown ehbleib  by   hhy4no hhy4k althv5k aloz4 hlu4  by  zufgr1


* *** Abgrenzung: Verweigerung plus Feldverluste * ***
* Ausfall- und Teilnahmewahrscheinlichkeiten fuer alle HH ***

temporary
select if (kont2 ge 1 and kont2 le 2)
breakdown   raus2 teil2  by kont2  hhy1no hhy1k althv2k aloz1 hlu1
       / bhbleib  by   hhy1no hhy1k althv2k aloz1 hlu1  by  zufgr1

temporary
select if (kont3 ge 1 and kont3 le 2)
breakdown   raus3 teil3  by kont3  hhy2no hhy2k althv3k aloz2 hlu2
       / chbleib  by   hhy2no hhy2k althv3k aloz2 hlu2  by  zufgr1

temporary
select if (kont4 ge 1 and kont4 le 2)
breakdown   raus4 teil4  by kont4  hhy3no hhy3k althv4k aloz3 hlu3
       / dhbleib  by   hhy3no hhy3k althv4k aloz3 hlu3  by  zufgr1

temporary
select if (kont5 ge 1 and kont5 le 2)
breakdown   raus5 teil5  by kont5  hhy4no hhy4k althv5k aloz4 hlu4
       / ehbleib  by   hhy4no hhy4k althv5k aloz4 hlu4  by  zufgr1

* *** Abgrenzung: nur Verweigerungen (d.h. ohne Feldverluste) * ***
*  Bleibewahrscheinlichkeiten fuer realisierte HH nach HH-Merkmalen *
                                    ==========

temporary
select if (kontakt2 ge 1 and kontakt2 le 2 and b=1)
breakdown bhbleib  by   hhy1no hhy1k althv2k aloz1 hlu1  by  zufgr1

temporary
select if (kontakt3 ge 1 and kontakt3 le 2 and c=1)
```

```
breakdown chbleib  by   hhy2no hhy2k althv3k aloz2 hlu2  by  zufgr1

temporary
select if (kontakt4 ge 1 and kontakt4 le 2 and d=1)
breakdown dhbleib  by   hhy3no hhy3k althv4k aloz3 hlu3  by  zufgr1

temporary
select if (kontakt5 ge 1 and kontakt5 le 2 and e=1)
breakdown ehbleib  by   hhy4no hhy4k althv5k aloz4 hlu4  by  zufgr1


* *** Abgrenzung: nur Verweigerungen (d.h. ohne Feldverluste) * ***
* Ausfall- und Teilnahmewahrscheinlichkeiten fuer alle HH ***

temporary
select if (kontakt2 ge 1 and kontakt2 le 2)
breakdown   raus2 teil2  by kontakt2  hhy1no hhy1k althv2k aloz1 hlu1

temporary
select if (kontakt3 ge 1 and kontakt3 le 2)
breakdown   raus3 teil3  by kontakt3  hhy2no hhy2k althv3k aloz2 hlu2

temporary
select if (kontakt4 ge 1 and kontakt4 le 2)
breakdown   raus4 teil4  by kontakt4  hhy3no hhy3k althv4k aloz3 hlu3

temporary
select if (kontakt5 ge 1 and kontakt5 le 2)
breakdown   raus5 teil5  by kontakt5  hhy4no hhy4k althv5k aloz4 hlu4
```

```
* *** Tabelle 3 ***

weight off

select if (a=1 and teil15 ne -2)
breakdown  bleib12 bleib13 bleib14 bleib15
             by   hhy1no hhy1k althv2k aloz1 hlu1
             by   zufgr1


* *** Tabelle 4 ***

weight off

frequencies general = hhy1no hhy1k althv2k aloz1 hlu1
crosstabs hhy1k althv2k aloz1 hlu1 by zufgr1
options 3 4 5

weight by bleib12
frequencies general = hhy1no hhy1k althv2k aloz1 hlu1

weight by bleib13
frequencies general = hhy1no hhy1k althv2k aloz1 hlu1

weight by bleib14
frequencies general = hhy1no hhy1k althv2k aloz1 hlu1

weight by bleib15
frequencies general = hhy1no hhy1k althv2k aloz1 hlu1

crosstabs hhy1k althv2k aloz1 hlu1 by zufgr1
options 3 4 5


* *** Tabelle 5 ***

weight by bleib15a
frequencies general = hhy1no hhy1k althv2k aloz1 hlu1


* *** Beispiel fuer logistische Regression (5-Wellen)***

weight off

logistic regression
  drin15   with   hlu1 hhy1low alt11 alt12 alt13 alt14 aloz1
  / method = enter


finish
```

B Die Varianzschätzung mit Hilfe einer Randomisierung der Stichprobe

Im folgenden wird kurz dargestellt, wie mit Hilfe einer einfachen "Randomisierungsvariablen" eine Varianzabschätzung von Populationswerten, Anteilen, Differenzen, etc. durchgeführt werden kann[24].

Diese mit SPSS-X erzeugte Variable zu ZUFGR1 zerlegt die gesamte SOEP-Stichprobe in acht etwa gleich große Teilstichproben. Sortiert man die Schätzergebnisse für diese acht Teilgruppen nach der Größe, so bietet für praktische Anwendungen das Konfidenzintervall zwischen dem zweitkleinsten und dem zweitgrößten Wert eine - im üblichen Zuverlässigkeitsbereich liegende - Schätzung für die Streuung des Wertes in der Grundgesamtheit. Die Irrtumswahrscheinlichkeit liegt bei 7% (vgl. Abschnitt 2).

Anhand zweier Beispiele wird die Vorgehensweise der Varianzabschätzung illustriert. Dabei soll geprüft werden, ob die im SOEP bereitgestellten Längsschnittgewichte bzw. Bleibewahrscheinlichkeiten einer bestimmten Population für unterschiedliche Längsschnittdatensätze (2 bzw. 5 Wellen) ausreichen, um die - ungewichtet nachweisbaren - unterschiedlichen Ausfallquoten auszugleichen und darauf aufbauend unverzerrte Ergebnisse zu erhalten. Die Berechnungen für die folgenden Tabellen wurden mit Hilfe von SPSS- X durchgeführt (Für den SPSS-Code vgl. Anhang A).

Zuerst wird die Bleibewahrscheinlichkeit der Haushalte mit HLU-Bezug von der ersten bis zur zweiten Welle für die Gesamtstichprobe und alle acht Teilpanels berechnet. Diese Werte werden nun der Größe nach sortiert und dem Erwartungswert von 100% gegenübergestellt (vgl. Tabelle B1).

Angesichts der niedrigen Fallzahl von 127 Haushalten mit HLU-Bezug in der ersten Welle ist es nicht weiter verwunderlich, daß die Bleibewahrscheinlichkeit in den einzelnen Zufallsgruppen schwankt. Da der Erwartungswert (100%) im 93%- Vertrauens-Intervall zwischen dem zweitkleinsten und dem zweitgrößten Zufallsgruppen-Wert liegt, kann davon ausgegangen werden, daß die Längsschnittgewichtung die interessierende Population nicht unterschätzt.

[24]Für den methodischen Hintergrund der "Randomisierung" von SOEP-Ergebnissen vgl. Rendtel (1991) sowie das SOEP-Benutzerhandbuch, Kapitel K.

Tabelle B1 Bleibewahrscheinlichkeit von Haushalten mit Bezug von laufender Hilfe zum Lebensunterhalt (HLU) in der ersten Welle bis Welle 2 (n=127)

Erwartungswert	Gesamtstichprobe	Random Groups							
		1	2	3	4	5	6	7	8
100.0	95.5	75.3	89.3	89.9	95.5	97.4	98.9	105.3	109.9
	(n=127)	(13)	(15)	(18)	(21)	(17)	(13)	(12)	(18)

Als zweites Beispiel dient die Bleibewahrscheinlichkeit der Haushalte mit 3000 bis 4000 DM Einkommen von der ersten bis zur fünften Welle (vgl. Tabelle B2).

Im Gegensatz zu Beispiel 1 ergibt sich hier - bei einer Irrtumswahrscheinlichkeit von 7% - eine signifikante Überschätzung, da lediglich der kleinste Wert einer Random Group unter dem Erwartungswert von 100 liegt und das Konfidenzintervall zwischen dem zweitgrößten und dem zweitkleinsten Zufallsgruppenwert den Wert 100 nicht einschließt. Wählt man jedoch als Vergleichswert nicht den theoretischen Erwartungswert (100%), sondern den empirischen Durchschnittswert der Gesamtpopulation (101.8%), so ist die Abweichung nach der 2-Sigma-Regel nicht mehr als signifikant einzustufen.

Tabelle B2 Bleibewahrscheinlichkeit von Haushalten mit Einkommen zwischen 3000 und 4000 DM in der ersten Welle bis Welle 5 (n=929)

Erwartungswert	Gesamtstichprobe	Random Groups							
		1	2	3	4	5	6	7	8
100.0	104.4	97.2	101.1	103.8	104.8	104.8	105.8	107.2	109.3
	(n=929)	(109)	(109)	(101)	(123)	(126)	(117)	(132)	(112)