Diskussionspapiere
Discussion Papers

Diskussionspapier Nr. 61

Über Sinn und Unsinn von Repräsentativitätsstudien

von
Ulrich Rendtel und Ulrich Pötter

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

Die in diesem Papier vertretenen Auffassungen liegen ausschließlich in der Verantwortung des Verfassers und nicht in der des Instituts.

Opinions expressed in this paper are those of the author and do not necessarily reflect views of the Institute.

Deutsches Institut für Wirtschaftsforschung

Diskussionspapier Nr. 61

Über Sinn und Unsinn von Repräsentativitätsstudien

von
Ulrich Rendtel und Ulrich Pötter

Ulrich Rendtel:
Deutsches Institut für Wirtschaftsforschung, Berlin.
Mitarbeiter im Projekt "Das sozio-oekonomische Panel" (SOEP)"

Ulrich Pötter:
Max-Planck-Institut für Bildungsforschung, Berlin.
Mitarbeiter im Forschungsbereich "Bildung, Arbeit und gesellschaftliche Entwicklung".

Berlin, im November 1992

Deutsches Institut für Wirtschaftsforschung, Berlin
Königin-Luise-Str. 5, 1000 Berlin 33
Telefon: 49-30 - 82 991-0
Telefax: 49-30 - 82 991-200

Über Sinn und Unsinn von Repräsentativtätsstudien [1]

Ulrich Rendtel [2]
Deutsches Institut für Wirtschaftsforschung
Berlin
Ulrich Pötter [3]
Max–Planck–Institut für Bildungsforschung
Berlin

Oktober 1992

[1] Wir danken Gert Wagner, der auch die Anregung zu diesem Aufsatz gab, sowie Johannes Huinink, Marlis Riebschläger und Rainer Schnell für viele hilfreiche und kritsche Kommentare.

[2] Mitarbeiter im Projekt " Das Sozio–oekonomische Panel (SOEP) "

[3] Mitarbeiter im Forschungsbereich "Bildung, Arbeit und Gesellschaftliche Entwicklung"

"Umfrageforschung wie amtliche Statistik verfolgen das Ziel, Eigenschaften der Bevölkerung und von Bevölkerungsteilen korrekt wiederzugeben", so lautet in der KZfSS (1992) die Einleitung einer Repräsentativitätsstudie über den Allbus.[1] Diese Arbeit berichtet über ein ZUMA- Grundlagenforschungsprojekt, ist also nicht irgendeine Repräsentativitätsstudie, mit der versucht wird, die "Güte" einer Stichprobe zu beweisen. Die in dieser Studie benutzten Prämissen, Schlußweisen und Ergebnisse erscheinen uns "repräsentativ" für eine bestimmte Umgangsweise mit Umfragedaten zu sein, die durch eine weitgehende Ausblendung statistischer und forschungslogischer Überlegungen charakterisiert ist.

Das Ziel dieses Aufsatzes ist es, einerseits auf methodische Schwachstellen derartiger Repräsentativitätsstudien hinzuweisen und mißverständliche Ausdrucksweisen zu präzisieren. Dabei geht es uns nicht um inhaltsleere mathematisch statistische Rechthaberei, sondern um die Präzisierung von oberflächlich niederschmetternden Aussagen wie: "Demnach entstammen die Allbus-Daten einer anderen Grundgesamtheit als der Mikrozensus". Derartige Aussagen können nur allzu leicht als generelles Verdikt gegen Umfragedaten mißverstanden werden.

Zudem soll versucht werden, zu einer genaueren Analyse von Ausfall- und Selektionsprozessen in sozialwissenschaftlichen Umfragen anzuregen. Dreh- und Angelpunkt für einen Fortschritt in der lang anhaltenden Debatte über "Mittelstandsbias" und "Repräsentativität" sind unserer Meinung nach eine genauere Auswertung von Feldinformationen bzw. deren Bereitstellung für Analysen sowie eine klarere statistische Modellierung des Erhebungsprozesses.

Viele der von uns geäußerten Gedanken sind keineswegs neu. Wir verweisen u.a. auf die umfangreiche Studie von W. Kruskal und F. Mosteller (1979 a,b,c) über "Representative Sampling" sowie die Dokumentation der Auseinandersetzung um eine soziologische Studie von S. Lang (1981). Dort werden die unterschiedlichen wissenschaftlichen Präzisierungen des Sammelbegriffs "repräsentativ", sowie deren nichtwissenschaftliche ("nonscientific") Zielvorstellungen erläutert.

Der Aufsatz ist folgendermaßen gegliedert: Zunächst setzen wir uns mit

[1] Peter H. Hartmann und Bernhard Schimpl-Neimanns, 1992: Sind Sozialstrukturanalysen mit Umfragedaten möglich? Analysen zur Repräsentativität einer Sozialforschungsumfrage, Kölner Zeitschrift für Soziologie und Sozialpsychologie, 44, Seite 315-340.

den Implikationen des Bildes von der Stichprobe als Miniatur der Grundgesamtheit auseinander. Dann präzisieren wir den Begriff der "Verzerrung" für Abweichungen von Stichprobe und Grundgesamtheit. Im dritten Abschnitt behandeln wir die Varianzen von Schätzungen der Populationswerte. Nach diesen Vorbereitungen kommentieren wir die Vergleiche univariater Verteilungen von Mikrozensus und Allbus.

Bei derartigen Vergleichen werden die Ergebnisse aus dem Mikrozensus anstelle der Werte der Grundgesamtheit benutzt. Der fünfte Abschnitt diskutiert am Beispiel der Haushaltsgröße, daß derartige Vergleichsgrößen ihrerseits nicht nur zufällige, sondern auch erhebungsbedingte Abweichungen von der Grundgesamtheit zeigen.

Am Beispiel der Variablen Schulabschluß und Erwerbstätigkeit zeigen wir im sechsten Abschnitt, daß die Abgrenzung von Restkategorien geradezu konstitutiv für den später reklamierten "Bildungsbias" ist.

Eine weitere Schlußweise der hier behandelten Repräsentativitätsstudie besteht darin zu testen, ob der Allbus und der Mikrozensus Realisationen derselben (Super-)population sind. Im siebten Abschnitt diskutieren wir die dabei implizit benutzten Homogenitätsannahmen und zeigen, daß man bei weniger restriktiven, jedoch realistischeren Annahmen zu einer anderen Interpretation der Ergebnisse kommt.

Die multivariate Ausfallanalyse von Hartmann/Schimpl-Neimanns läßt sich auch als Schätzung von Teilnahmewahrscheinlichkeiten des Allbus interpretieren. Dieser Ansatz ist eng verwandt mit dem Verfahren der Randanpassung ("Redressment"), das Teilnahmwahrscheinlichkeiten ebenfalls ohne die Verwendung von Informationen über Nonrespondenten schätzt. In den Abschnitten 8 und 9 werden die beiden Verfahren zueinander in Beziehung gesetzt und ihre impliziten Modellannahmen diskutiert.

Schließlich setzen wir uns im zehnten Abschnitt mit dem von Hartmann/Schimpl-Neimanns für den Allbus reklamierten Bildungsbias auseinander. Daran anschließend geben wir einige Anregungen, wie man nicht ignorierbare Einflüsse der Stichprobenauswahl bei Sozialstrukturanalysen aufdecken kann. Abschließend setzen wir uns mit dem Begriff "Repräsentativitätsstudie" auseinander.

1 Repräsentative Stichprobe = Miniatur der Grundgesamtheit ?

Hartmann und Schimpl-Neimanns (im folgenden HS), die Autoren der Studie, gehen von der gängigen Vorstellung aus, daß eine repräsentative Stichprobe eine Miniatur der Grundgesamtheit darstellt: "Unter Repräsentativität wird hier die Übereinstimmung der multivariaten Randverteilungen der Merkmale einer Stichprobe mit den Verteilungen der Grundgesamtheit verstanden." Dahinter verbirgt sich die weit verbreitete[2] Idee, daß unter solchen Umständen die Stichprobe ein Miniaturbild der Grundgesamtheit ist, ihr also in allen Aspekten gleicht.

Aber ist die Forderung, daß Verteilungen aus Stichproben mit denen der Grundgesamtheit übereinstimmen müssen, sinnvoll? Wir zeigen im folgenden: Sie ist es nicht.

Erstens führt die naive Idee einer Stichprobe als Miniaturbild der Grundgesamtheit auf einen infiniten Regress: Wenn die Stichprobe in allen Aspekten der Grundgesamtheit gleicht, dann müßte sie wie die Grundgesamtheit eine Teilstichprobe enthalten, die ebenfalls "repräsentativ" wäre. Diese Teilstichprobe müßte ihrerseits eine weitere Sub-Teilstichprobe enthalten, und so weiter bis zu einer Stichprobe vom Umfang eins, dem "repräsentativen Menschen" (oder gar dessen Teilen, "repräsentative" Homunkuli). Der infinite Regress kann nur durch eine willkürliche Entscheidung durchbrochen werden, etwa: Eine "repräsentative" Stichprobe enthält mindestens 100 verschiedene Beobachtungen.

Weiterhin bricht die Idee der Übereinstimmung multivariater Verteilungen in Grundgesamtheit und Stichprobe allein wegen der hohen Zahl der möglichen Merkmalskombinationen zusammen. Die Autoren verwenden die Merkmale Geschlecht, Alter, Stellung im Beruf und Bildung, was zu einer Tabelle mit $2 \cdot 12 \cdot 5 \cdot 5 = 600$ Zellen führt. Würde zudem das Merkmal Haushaltsgröße (5 Ausprägungen) — immerhin eine wichtige Designvariable — benutzt, ergäben sich 3000 Zellen. Die Verteilung von 3000 Beobachtungen

[2] In einer Einführung für Sozialwissenschaftler (Böltken 1976, Seite 128) liest man beispielsweise: "Unter Repräsentativität verstehen wir, daß in einer Auswahl alle für die Grundgesamtheit typischen und charakteristischen Merkmale und Merkmalskombinationen getreu ihrer relativen Häufigkeit vertreten sein müssen und somit die Auswahl ein verkleinertes Abbild der Grundgesamtheit selbst für solche Merkmale ist, von deren Vorhandensein wir (noch) gar nichts wissen."

in 3000 Zellen kann aber nicht mehr mit der Verteilung in der Grundgesamtheit übereinstimmen. Schließen HS daher, daß Umfragedaten mit 3000 Beobachtungen zu klein für Sozialstrukturforschungen sind? Vielleicht würden HS an dieser Stelle darauf verweisen, daß nur eine näherungsweise Übereinstimmung der "wichtigen" Verteilungen gemeint war. Dann muß aber zunächst darüber entschieden werden, welche Variablen "wichtig" sind. Sodann muß ein Abstandsmaß zwischen den Verteilungen definiert werden und jemand hat zu entscheiden, wann der so definierte Abstand klein genug ist, um einer Stichprobe "Repräsentativität" zu bescheinigen.

Nicht zuletzt widerspricht die Idee der Übereinstimmung der Verteilungen dem tatsächlichen Ziehungsverfahren sozialwissenschaftlicher Stichproben. Diese beruhen fast ausnahmslos auf zufälligen Auswahlverfahren. Abweichungen der Verteilungen voneinander sind allein aufgrund der Zufälligkeit der Auswahl zu erwarten.[3] Eine exakte Übereinstimmung mit gegebenen Verteilungen könnte dagegen einfach durch ein entsprechendes Quotenverfahren erreicht werden. Solche Verfahren führen aber bekanntermaßen zu in der Praxis nicht kontrollierbaren Selektionsprozessen. Zufällige Auswahlverfahren haben das Ziel, die absichtliche Selektion von Beobachtungen (und damit von Ergebnissen) durch die Planer einer Untersuchung oder die Interviewer auszuschließen.[4] Dieser Aspekt der Glaubwürdigkeit auf Grund des Auswahlverfahrens kann kaum überschätzt werden.

Zunächst erlaubt eine Analyse der Verteilungen in der Stichprobe lediglich eine Beurteilung des Informationsgehalts der Stichprobe. Enthält die Stichprobe nur wenige Fälle mit interessierenden Merkmalskombinationen, so wird man die Umfrage bezüglich dieser Merkmale als wenig informativ ansehen.

[3]Hier hilft auch nicht die Anrufung des Gesetzes der großen Zahlen, das die Konvergenz der Verteilungen gegen den Populationswert "beweisen" soll. Die Bedingungen für die Anwendung dieses Theorems — unabhängige und identisch verteilte Merkmalsausprägungen — sind schlicht nicht gegeben, vgl. Abschnitt 7.

[4]Die Zufälligkeit der Auswahl als ein Kriterium der Güte einer Stichprobe bezieht sich auf das Verfahren der Stichprobenziehung. Sie kann post hoc an Hand eines vorliegenden Datensatzes nicht mehr überprüft werden. Bei einem Würfel, der eine 6 zeigt, kann nachträglich nicht festgestellt werden, ob er geworfen oder entsprechend gelegt wurde.

2 Wann sind "verzerrte" Schätzungen verzerrt ?

Abweichungen der Stichprobenverteilung von der Grundgesamtheit werden bei HS als "Verzerrungen"[5] gewertet, wobei wir hinzufügen möchten, daß dieses Etikett a) einen pejorativen Charakter besitzt, b) genauso unpräzise ist wie das Etikett "repräsentativ" und c) folgerichtig von HS in unterschiedlicher Bedeutung verwandt wird.

Um die Aussage zu begründen, daß nicht jede "verzerrte" Schätzung im statistischen Sinne verzerrt ist, bedarf es der Präzisierung: Es bezeichne $i = 1, ..., N$ die i-te Einheit der Gundgesamtheit vom Umfang N. Aus dieser Grundgesamtheit werden im Rahmen einer zufälligen Stichprobenziehung n Einheiten ausgewählt ($C_i = 1$ für ausgewählte Einheiten, $C_i = 0$ sonst). Für die Stichprobenmitglieder ist die interessierende Merkmalsausprägung ($Y_i = 1$, falls Merkmalsausprägung vorhanden, $Y_i = 0$ sonst) bekannt.

Von Interesse ist der unbekannte Populationswert $P = \frac{1}{N}\sum_{i=1}^{N} Y_i$. P soll auf Basis der Stichprobe geschätzt werden. Verwendet man die Stichprobenverteilung $\hat{P} = \frac{1}{n}\sum_{i=1}^{N} C_i Y_i$ als Schätzung für P, so werden P und $\hat{P}$ im allgemeinen nicht für jede Merkmalskombination übereinstimmen. Denn bei der Vielzahl der erhobenen Variablen lassen sich sofort seltene Merkmalskombinationen angeben, die zwar in der Grundgesamtheit auftreten, nicht jedoch in der Stichprobe. Trotzdem möchte man eine Güteforderung an die Erhebung von Umfragedaten stellen, die Aussagen über die Relation von P und $\hat{P}$ ermöglichen und diese Differenz in einem zu präzisierenden Sinn klein halten.

Der Randomisierungsansatz der Stichprobentheorie (vgl. z.B. Cochran 1977, Kapitel 1) behandelt im Gegensatz zu dem weiter unten dargestellten Superpopulationsansatz lediglich die Auswahlindikatoren C_i als zufällige Größen, während die Y_i als Konstanten behandelt werden. Diese asymmetrische Betrachtungsweise beruht auf der Überlegung, daß die Auswahlwahrscheinlichkeiten, d.h. $P(C_i = 1)$, ansatzweise für jede Einheit i bekannt sind, während die Y_i entweder nicht als Resultat eines Zufallsexperiments interpretiert werden können oder über die Form des zugrundeliegenden Experiments nichts bekannt ist. Alle Aussagen nach dem Randomisierungsansatz sind daher konditional bezüglich der Merkmalsausprägungen in der Grund-

[5] z.B. HS auf Seite 320: "Dagegen fehlen in der Sozialforschungsumfrage besonders Einpersonenhaushalte. Die Verzerrung ist überraschend stark."

gesamtheit. Zufällig ist allein die Ziehung der Stichprobe.

Innerhalb dieses Ansatzes läßt sich nun ein über alle im Rahmen des Erhebungsdesigns möglichen Stichprobenziehungen gemittelter Wert für $\hat{P}$ berechnen. Hierbei ist der Stichprobenumfang n fest. Besonders einfach ist dieser erwartete Schätzwert für Schätzer der Form $\hat{P} = \frac{1}{n}\sum_{i=1}^{N} \alpha_i C_i Y_i$ bestimmbar, wobei die α_i noch zu bestimmende Konstanten sind. vgl. Horvitz/Thompson (1952). Eine tatsächlich erfüllbare Qualitätsanforderung an eine Erhebung ist die Übereinstimmung von $E(\hat{P})$ mit P für jedes Merkmal. Diese Qualitätsanforderung wird wegen

$$E(\hat{P}) = \frac{1}{n}\sum_{i=1}^{N} \alpha_i E(C_i) Y_i \tag{1}$$

für

$$\alpha_i = \frac{n}{N}\frac{1}{E(C_i)} = \frac{n}{N}\frac{1}{P(C_i = 1)} \tag{2}$$

erfüllt.

Damit lassen sich zwei Qualitätsanforderungen an eine Stichprobenerhebung formulieren:

(R1) $P(C_i = 1)$ muß für alle Einheiten der Grundgesamtheit positiv sein.
(R2) $P(C_i = 1)$ muß für alle Einheiten der Stichprobe bekannt sein.

Bedingung (R1) sichert, daß α_i für alle $i = 1, ..., N$ definiert ist, während Bedingung (R2) fordert, daß α_i für alle Elemente der Stichprobe bekannt ist, so daß $\hat{P}$ berechnet werden kann. $\hat{P}$ in der Form (1) und (2) ist unter diesen Voraussetzungen ein erwartungstreuer, also unverzerrter Schätzer für P.

Die Bedingung (R1) läßt sich auch so interpretieren: Mit Umfragedaten können Aussagen nur über den Teil der Grundgesamtheit gemacht werden, der eine positive Auswahlchance hat. Schnell (1991) liefert eine eindrucksvolle Aufstellung, wer bei "allgemeinen Bevölkerungsumfragen" die Auswahlchance 0 hat.

Die Erhebung über ein Quotenverfahren erfüllt die Bedingung (R2) nicht. Die mit diesem Erhebungsverfahren gewonnenen Populationsschätzungen besitzen keine statistisch fixierbaren Eigenschaften.

Es sollte aber klar sein. daß die Erfüllung der Forderungen R1 und R2 keineswegs unvernünftige Erhebungsdesigns ausschließt. So liefert eine Stichprobe, bei der mit Wahrscheinlichkeit 1 entweder nur Männer oder nur Frauen befragt werden, einen unverzerrten Schätzer für den fixen Wert 0.5, der dem Männeranteil in der Bevölkerung entspricht (wenn diese aus 50% Männern und 50% Frauen besteht). In diesem Stichprobendesign sind die Forderungen R1 und R2 offenbar erfüllt und die $P(C_i = 1)$ sind sogar für alle Personen gleich. Jede einzelne Realisation der Stichprobe ergibt aber einen Schätzwert von 0 oder 1, d.h. einen Schätzwert, der maximal von dem "wahren" Wert 0.5 abweicht. Keine der möglichen Stichproben enthält Information über den Männeranteil in der Population !

Im Rahmen des Randomisierungsansatzes sind Abweichungen vom Populationswert zunächst einmal zufällige Abweichungen. wobei Zufälligkeit durch die Zufälligkeit der Stichprobenziehung definiert ist. Damit ist das böse Wort von der "Verzerrung" für jede Abweichung nicht mehr zulässig und eine präzise Definition von Unverzerrtheit als Erwartungstreue unter dem Ziehungsexperiment wurde präsentiert. Diese Präzisierung bezieht sich auf das Verhalten von Schätzern bezüglich aller realisierbaren Stichproben. Sie ist kein Qualitätskriterium für die Beurteilung der gegebenen Stichprobe.

3 Wie genau sind Schätzungen der Populationswerte ?

Zudem ist Erwartungstreue eine relativ schwache Forderung, die die Streuung um den Erwartungswert unberücksichtigt läßt. Die erwartete Streuung um den Populationswert sollte in "vernünftigen" Größenordnungen liegen. Hierzu muß erstens definiert werden, was als Beurteilungsmaß für die Streuung der Stichprobenschätzung angesehen wird und zweitens muß man in der Lage sein, dieses Streuungsmaß — zumindest annähernd — zu schätzen. Bei erwartungstreuen Schätzern wird häufig die Varianz der Schätzergebnisse verwandt.

Macht man Repräsentativität an der Varianz von $\hat{P}$ fest, so stellt man fest, daß diese Eigenschaft nicht an die Erhebung als solche, sondern an einzelne Variablen gekoppelt ist. In der Tat haben Schätzungen mit einem relativen Fehler von 0.30 kaum einen Aussagewert.

Es ist leider bei Sozialstrukturanalysen gängige Praxis, auf Genauig-

keitsangaben der präsentierten Populationsschätzungen zu verzichten. Diese Praxis mag daher rühren, daß Konfidenzintervalle für P — bzw. Standard-Abweichungen für $\hat{P}$ — bei den verwendeten komplexen Erhebungsdesigns schwer zu bestimmen sind. Allerdings sind die Effekte des Stichprobendesigns in diesem Zusammenhang essentiell.

Es sei darauf hingewiesen, daß $Var(\hat{P})$ konzeptionell verschieden ist von der Varianz der Y_i in der Stichprobe. Lediglich für einfache Stichproben (Urnenexperiment mit Zurücklegen) fallen beide Schätzungen zusammen, vgl. Cochran (1977, Seite 23). Für komplexe Auswahldesigns, die wie beim Allbus systematisches Ziehen mit Intervall nach Anordnung der Stichprobeneinheiten verwenden, ist die analytische Bestimmung - d.h. die Berechnung von $Var(\hat{P})$ über eine Formel - nicht möglich, vergleiche z.B. Kirschner (1984). Konsequenterweise verwenden HS auch nicht die durch das Erhebungsdesign implizierten Varianzen bei ihren Vergleichen von Allbus und Mikrozensus. Die von ihnen verwendeten χ^2- Statistiken als Teststatistik für signifikante Merkmalsunterschiede können nicht im Rahmen des Randomisierungsansatzes gerechtfertigt werden.

Es ist aber keinesfall so, daß man im Randomisierungs-Ansatz bei komplexen Erhebungsdesigns auf die Berechnung von $Var(\hat{P})$ verzichten muß oder sich mit der "Approximation" dieser Varianz durch eine einfache Stichprobe plus irgendwelcher willkürlichen Designfaktoren begnügen muß. Die Statistik bietet mittlerweile eine Vielzahl von Resampling - Verfahren an, mit denen $Var(\hat{P})$ geschätzt werden kann, vgl. Wolter (1985). Diese Verfahren basieren darauf, über eine geeignete Zerlegung der Stichprobe mehrere Versionen der Populationschätzung zu konstruieren und aus deren Streuung auf die Varianz von $\hat{P}$ zu schließen.

Im Vergleich zu einer einfachen Stichprobe verringert die netzartige Überdeckung des Erhebungsgebiets durch das systematische Auswahlverfahren des Allbus die Varianz von $\hat{P}$ qualitativ, während der Klumpungseffekt durch die vollständige Befragung der Bewohner eines Zählbezirks beim Mikrozensus zu einer Vergrößerung der Varianz von $\hat{P}$ führt. Die quantitativen Größenordnungen dieser gegenläufigen Effekte, die von Merkmal zu Merkmal variieren, lassen sich über die genannten Resampling-Methoden abschätzen.[6]

[6]Eine Anwendung derartiger Methoden auf das Sozio-oekonomische Panel (SOEP), dessen erste Welle bis auf die dritte Auswahlstufe — Auswahl einer wahlberechtigten Person je Haushalt — dem Allbus sehr ähnlich ist, findet man bei Rendtel (1991).

4 Was kann man aus dem Vergleich univariater Verteilungen lernen ?

HS analysieren im ersten Teil ihres Beitrags insgesamt 36 univariate Merkmalsverteilungen auf ihre Übereinstimmung mit dem Mikrozensus. Aus Abschnitt 1 wissen wir bereits, daß eine perfekte Übereinstimmung der Verteilungen nicht erwartet werden kann. Die in Abschnitt 2 angegebenen Präzisierung des Begriffs "Verzerrrung" beschreibt die Eigenschaften des Ziehungsverfahrens und nicht der realisierten Stichprobe. Prinzipiell lassen sich mit den oben dargestellten Varianzschätzungen Signifikanztests auf Gleichheit der zu schätzenden Populationswerte mit vorgebebenen Werten formulieren.

Nun enthalten aber Studien wie der Allbus hunderte von Variablen. Da die Auswahl der Variablen zum Vergleich mit der Grundgesamtheit nicht begründet wird und wohl auch nicht unabhängig von spezifischen Fragestellungen begründet werden kann, ist es immer möglich, Variablenkonstellationen zu finden, deren Verteilungen "signifikant" von denen in der Grundgesamtheit abweichen.[7] Auf diese Weise kann jeder Umfragestudie "Nichtrepräsentativität" bescheinigt werden, wenn Interessenten an einer solchen Aussage nur lange genug suchen. Ein Vergleich von Verteilungen in der Stichprobe mit denen der Grundgesamtheit mit dem Ziel, Aussagen über "Repräsentativität" abzuleiten, erscheint unter diesem Gesichtswinkel als akademische Spielerei.

Läßt man sich auf Vergleiche mit Verteilungen in der Grundgesamtheit ein (sagen wir, man benutzt nur die von HS für zentral erklärten Variablen Alter, Geschlecht, Schulbildung und Erwerbsstatus und benutzt zusätzlich sinnvolle Schranken für Abweichungen, die sogar die Effekte des multiplen Testens berücksichtigen) so gibt es zwei mögliche Resultate: Es werden keine "signifikanten" Abweichungen festgestellt oder mindestens eine Variable zeigt "signifikante" Abweichungen.

Im ersten Fall erwarten Vertreter des Miniaturbilds, daß nun auch die Verteilungen der übrigen Merkmale der Stichprobe mit der Grundgesamtheit übereinstimmen. Dies könnte aber nur gewährleistet sein, wenn die Verteilung der gewählten "wichtigen" Merkmale die Verteilung der übrigen Merkmale

[7] In ihrer Tabelle 6 führen HS insgesamt 36 Tests durch. Wenn diese Tests unabhängig wären, ergäbe sich statt des von den Autoren reklamierten Signifikanzniveaus von 1% ein Niveau von $1 - 0.99^{36} = 0.31$, was als Fehlerwahrscheinlichkeit der ersten Art kaum zu akzeptieren ist. Um auf ein Gesamtniveau von 1% zu kommen, wäre unter diesen Umständen ein Niveau von 0.03% für die einzelnen Tests anzunehmen.

vollständig bestimmt. Diese Determination soziale Merkmale durch Basismerkmale widerspricht in gewisser Weise der Notwendigkeit von sozialwissenschaftlichen Umfragen und ist auch empirisch nicht zu halten: vgl. hierzu auch Schnell (1992).[8] Träfe diese Hypothe zu, so ließe sich die Verteilung der interessierenden sozialen Merkmale aus der bekannten Verteilung der Basismerkmale bestimmen. Wozu dann noch sozialwissenschaftliche Umfragen?

Bei einem anderen Ausgang des Vergleichs mit der Grundgesamtheit findet man "signifikante" Abweichungen. Bei der dann einsetzenden Ursachenanalyse findet der Sozialforscher immer nachvollziehbare Gründe, warum das Stichprobenergebnis systematisch von dem Ergebnis für die Grundgesamtheit abweicht, z.B. : Unterschiedliche Abgrenzung von Zielpopulation und Grundgesamtheit, unterschiedliche Operationalisierung des Merkmals, unterschiedliche Erhebungszeitpunkte, etc.. In den beiden folgenden zwei Abschnitten werden hierfür einige Beispiele genannt. Über die Größe derartiger systematischer Effekte kann nur spekuliert werden.

Natürlich können die Abweichungen auch durch bisher nicht berücksichtigte Effekte der Feldarbeit entstanden sein. Es lohnt sich in diesem Zusammenhang immer, diese Möglichkeit zu berücksichtigen. Nur besteht ohne entsprechende Feldinformation keine Möglichkeit, entsprechende Hypothesen — etwa HS's These vom Bildungsbias — zu überprüfen.

Resümieren wir: Egal wie der Vergleich mit den Basisdaten der Grundgesamtheit ausgeht, er endet — ohne entsprechende Feldinformation — immer in der Spekulation !

5 Wie genau ist der Mikrozensus ?

Die Mikrozensus-Daten werden von HS aus drei Gründen als Ersatz für die Grundgesamtheit gewählt: a) die hohe Fallzahl b) die Auskunftspflicht der Befragten und c) die konzeptionelle Vergleichbarkeit von Mikrozensus und Allbus hinsichtlich demographischer und sozialstruktureller Kategorien.

Anknüpfend an die Ergebnisse des vorherigen Abschnitts behaupten wir, daß sowohl Mikrozensus als auch Allbus zufällige Stichproben aus derselben Grundgesamtheit sind und allein aufgrund der Zufälligkeit der Stichprobenziehungen Abweichungen voneinander zeigen können. In ihrer univariaten

[8] Selbst HS bemerken auf Seite 331, daß die Korrelation zwischen Schulabschluß und Stellung im Beruf "alles andere als perfekt ist."

Analyse ignorieren HS die stichprobenbedingte Zufälligkeit der Mikrozensusergebnisse. Daneben gibt es jedoch auch erhebungstechnische Gründe, warum man die Mikrozensus–Daten nicht unbedingt mit der Grundgesamtheit identifizieren sollte.

Von HS wird die Unterrepräsentierung von Einpersonenhaushalten im Allbus gegenüber dem Mikrozensus diskutiert. Immerhin treten für alle drei Meßzeitpunkte Abweichungen von ca. 10 Prozentpunkten beim Anteil der Einpersonenhaushalte auf. Zurecht wird darauf hingewiesen, daß die Allbus-Auswahl über die geschätzte Anzahl von Haushalten im Wahlkreis denjenigen Wahlkreisen (Sample-Points) eine höhere Auswahlchance zuschreibt, in denen die Anzahl geschätzter Haushalte die tatsächliche Anzahl übersteigt. Dies ist aber gerade bei Wahlkreisen mit überproportional vielen großen Haushalten der Fall.

Nicht problematisiert wird hingegen die Verteilung der Haushaltsgrößen im Mikrozensus. Hier zeigt jedoch der Vergleich des Mikrozensus 1987 mit der nur drei Monate später durchgeführten Volkszählung trotz nahezu perfekter Übereinstimmung der Verteilung hinsichtlich Alter und Geschlecht eine Abweichung von 800 Tsd. Haushalten (Wedel 1989). Der Mikrozensus wies hierbei fast 600 Tsd. Einpersonenhaushalte mehr aus als die Volkszählung.

Ein Grund für eine mögliche unterschiedliche Messung der Haushaltsverteilung ist die Beobachtung, daß die Größe eines Haushalts in bestimmten Situationen eine ungenaue, für Interpretationen offene Größe ist, vgl. hierzu etwa Wedel (1989), Bartholmai/Melzer/Schulz (1990) sowie Prester (1992). Solche Situationen sind beispielsweise gegeben, wenn mehrere Generationen in einem Mehrfamilienhaus wohnen: Bilden die Großeltern nun einen eigenen Haushalt oder nicht? Die Beantwortung dieser Frage wird trotz minutiöser definitorischer Regelungen häufig im Ermessen des Interviewers stehen. Hierbei waren die Anreize für den Interviewer im Mikrozensus und in der Volkszählung direkt entgegengesetzt: Im Mikrozensus werden freiwillige Interviewer eingesetzt, die für jeden Haushalt bezahlt werden: Je höher die Anzahl der Haushalte desto höher der Verdienst.

In der Volkszählung wurden die Interviewer nicht auf freiwilliger Basis rekrutiert, die Bezahlungsanreize waren gering und jeder neue Haushalt bedeutete Mehrarbeit in Form eines neuen Haushaltsmantelbogens. D.h. die Interviewer hatten den umgekehrten Anreiz, nämlich möglichst wenige Haushalte zu identifizieren.

Halten wir fest: Neben zufälligen gibt es auch erhebungstechnische Gründe,

warum die Verteilungen von Allbus und Mikrozensus voneinander abweichen. Dabei kann auch die amtliche Statistik nicht für sich in Anspruch nehmen, immer mit der Grundgesamtheit übereinzustimmen.

6 Die Abgrenzung von Restkategorien als Ursprung des Bildungsbias

Eine weitere von HS benutzte Variable ist der Bildungsabschluß. In Tabelle 5 von HS ist der Anteil der Restkategorie "Hauptschule[9] ohne Lehre" im Allbus um ca. 11 Prozentpunkte niedriger als im Mikrozensus. Zwar besteht für den Mikrozensus Ankunftspflicht. Es ist jedoch nicht klar, inwieweit dieser Pflicht bei der Beantwortung einzelner Fragen nachgekommen wird.[10] Es ist durchaus vorstellbar, daß die Beantwortung einzelner Fragen schlicht übersehen wird.[11] Auf jeden Fall muß zwischen Unit-Nonresponse, d.h. der Nichtbeantwortung des gesamten Fragebogens und Item-Nonresponse, d.h. der Nichtbeantwortung einzelner Fragen unterschieden werden, und es muß möglich sein, die Nichtbeantwortung einer Frage zu erkennen. Dies ist aber im Mikrosensus bei der Variablen Bildungsabschluß nicht möglich, wo Personen ohne Hauptschulabschluß nur alle Bildungskategorien unangekreuzt lassen können. Die Abnahme des Bestands an Personen ohne Hauptschulabschluß (bzw. ohne Lehrabschluß) um fast 16% in nur 4 Jahren [12] erscheint uns in dieser Größenordnung wenig realistisch unde könnte auch ein Resultat verbesserter Feldarbeit sein.

Die Zusammenfassung der "keine Angabe"-Kategorie mit anderen Kategorien ist willkürlich und kann in der Tat zu systematischen Verzerrungen führen. Nehmen wir beispielsweise an, daß die Ausfälle bezüglich eines Merkmals mit zwei Ausprägungen völlig neutral sind, d.h. für jede Einheit ist die Ausfallwahrscheinlichkeit gleich groß. In der Population betrage das Verhältnis der Merkmale 10 : 90. Dieses Verhältnis überträgt sich auf die

[9]Genauer: Hauptschule oder Kein Abschluß

[10]Leider veröffentlicht das Statistische Bundesamt in seinen Tabellen keine Zahlen über Unit- und Itemnonresponse. In Esser et. al. (1989, Seite 80 u. Seite 281) wird ein Unitnonresponse von 3 Prozent belegt und für Einkommensangaben ein Itemnonresponse von 7,5 Prozent angegeben.

[11]Insbesondere bei schriftlicher Beantwortung des Fragebogens.

[12]Die Abnahme des Bestands dieses Personenkreises von 31,0% (1985) um 4,9 Prozentpunkte auf 26,1% (1989) entspricht einer Reduzierung des 85er Bestands um 16 Prozent.

Nonrespondenten. Bei einer Ausfallrate von x und der Zusammenfassung der kleineren Kategorie mit den Nonrespondenten ergibt sich ein Verhältnis von $(1 + 9x)/(9 - 9x)$ für die beiden Merkmale. Für eine Ausfallrate von $x = 0.30$ erhält man beispielsweise ein Verhältnis von 0.58; also eine deutliche "Unterrepräsentierung" des häufigeren Merkmals. Die gleiche Strategie verwenden HS auch bei dem Merkmal "Stellung im Beruf", wobei die Nichterwerbstätigen als "Restkategorie" fungieren. Hier sind nach Meinung von HS Arbeiter unterrepräsentiert, während die Restkategorie Nichterwerbstätige überrepräsentiert ist. Wie stabil ein derartiger Unterschicht-Bias jedoch gegenüber unterschiedlichen Abgrenzungen der Restkategorie ist, bleibt dahingestellt.

7 Wenn Mikrozensus und Allbus aus unterschiedlichen Grundgesamtheiten stammen

In ihrer multivariaten Analyse rücken HS von ihrer bisherigen Annahme ab, "der Mikrozensus gäbe die 'wahren' Werte wieder" (S. 331). In dieser Analyse gelangen HS zu der Feststellung, "daß die Sozialforschungsumfrage Allbus 1990 nicht aus derselben Grundgesamtheit wie der Mikrozensus stammt. Sie kann damit nicht als Zufallsstichprobe aus dieser Grundgesamtheit angesehen werden" (S. 337).

"Aus welcher Grundgesamtheit ist denn dann die Allbusstichprobe?" wird man fragen und beim genaueren Lesen feststellen, daß beide Stichproben keine Stichproben aus einer Grundgesamtheit sind, sondern Stichproben zweier — eventuell verschiedener — Grundgesamtheiten. Daß Stichproben Realisationen der Grundgesamtheit sind, ist der klassischen Stichprobentheorie fremd.

Wir vermuten, daß HS nicht von der klassischen Stichprobentheorie ausgegangen sind, sondern Modell-basierte Ansätze, die auch unter dem Namen Superpopulationsmodelle bekannt sind (vgl. Cassel et al. (1977), Särndal (1978) sowie Little (1982)) zugrunde gelegt haben. Gemäß dieser Modellklasse sind die Beobachtungsmerkmale einer Einheit (Person oder Haushalt) zufällige Realisationen einer Wahrscheinlichkeitsverteilung. Diese Wahrscheinlichkeitsverteilung beschreibt die Merkmalsverteilung in einer unendlich großen Population, der "Superpopulation", aus der sowohl die Einheiten der Grundgesamheit als auch der Stichprobe in einem Urnenexperiment gezogen wer-

den. Die Parameter, die diese Verteilung charakterisieren, sind unbekannte, fixe Größen.

HS möchten nun testen, ob alle Parameter der von ihnen unterstellten Multinominal-Verteilung für zwei unterschiedliche Stichprobenziehungen identisch sind.[13] Wird die Null-Hypothese, daß alle Verteilungsparameter in 2 Stichproben gleich sind, abgelehnt, kann man - die Korrektheit des zugrundegelegten Modells vorausgesetzt - davon sprechen, daß die beiden Superpopulationen, die diese Verteilungen repräsentieren, unterschiedlich sind. Oder: Die beiden Stichproben entstammen unterschiedlichen (Super-)populationen.[14]

Obwohl die quantitativen Unterschiede zur klassischen Stichprobentheorie mitunter gering sein mögen, vgl. Little (1982), sind die konzeptionellen Unterschiede prinzipieller Natur. Wir meinen, daß das Interesse der amtlichen Statistik wohl eher auf die Population als auf die Superpopulation gerichtet ist und würden dies auch für viele Sozialforscher vermuten, wenn sie sich auf Ergebnisse der amtlichen Statistik beziehen. Bei HS wird dieser Unterschied jedoch verwischt.

Die Gültigkeit der von HS benutzten Teststatistiken basiert wesentlich auf der Annahme, daß die Zuordnung der einzelnen Einheiten zu einer Merkmalskombination unabhängig voneinander ist und alle Einheiten die identische Chance haben, einer Merkmalskombination zugeordnet zu werden. Formal: Die Beobachtungen sind unabhängig und identisch verteilt. Beide Annahmen widersprechen in hohem Maße den Ergebnissen der Sozialforschung über die Bildung von Sozialstrukturmerkmalen in der Realität: Die Merkmale sind abhängig - z.B. über den Haushaltsverbund -, und Gleichheit der Verteilung beschreibt allenfalls eine politische Zielvorstellung denn soziale Realität.

Haben die Personen unterschiedliche Wahrscheinlichkeiten für die Zuordnung von Merkmalen, so ist die resultierende Verteilung für die Stichprobe

[13]HS weisen diese Verteilungsannahme nicht direkt aus. Eine wesentlich klarere Beschreibung des von HS benutzten Ansatzes findet man bei Arminger (1990). Erst hier wird dem Leser die Bezeichnung " Unabhängigkeitsmodell " in Tabelle 7 von HS klar.

[14]Dieses Ergebnis wird von HS nicht weiter interpretiert. Doch es stellt sich sofort die Frage, was bedeutet es, wenn zwei Stichproben aus unterschiedlichen Superpopulationen stammen Berkson (1942) formulierte den entstehenden Zwiespalt folgendermaßen: Suppose I said, "Albinos are very rare in human populations, only one in fifty thousand. Therefore, if you have taken a random sample of 100 from a population and found in it an albino, the population is not human." ... I believe the rational retort would be, " If the population is not human, what is it ? "

nicht mehr multinomial verteilt sondern besitzt eine höhere Varianz. McCullagh/Nelder (1983,p107) sprechen von Überdispersion. Die unter der Annahme der Multinomial–Verteilung gemachten Tests sind dann nicht mehr valide, sie schätzen Unterschiede zu schnell als signifikant ein. Um korrekte Testniveaus zu erhalten, muß der Grad der Überdispersion geschätzt werden. Ein möglicher Schätzer für diesen Grad ist gerade die von HS benutzte Devianz, vgl. McCullagh/Nelder (1983, Seite 83 und Seite 109). Hier findet man sich nun in folgendem Dilemma wieder: Entweder die Devianz mißt die Größe von Modellabweichungen oder den Grad von Überdisperson. Anders ausgedrückt: Entweder die Stichproben stammen aus zwei verschiedenen — jeweils homogenen — Superpopulationen oder sie entstammen aus einer — wenngleich sehr heterogen — Superpopulation. Der "Test" auf Unterschiedlichkeit der Populationsparameter hilft einem also an dieser Stelle überhaupt nicht weiter.

Weiterhin ignoriert der Superpopulations–Ansatz in der von HS benutzten Form die Auswahl der Stichprobe aus der Superpopulation. Jedoch nur wenn die Stichprobenauswahl von den Merkmalen, über die Aussagen gemacht werden sollen, unabhängig ist, behalten die benutzten Teststatistiken ihre Gültigkeit. Diese Voraussetzung ist beim Allbus sicher bezüglich aller Merkmale verletzt, die mit der Haushaltsgröße korrelieren.[15] In diesem Fall stimmt die Verteilung der beobachteten Merkmale, die konditional bzgl. der Stichprobenziehung ist, nicht mit der marginalen Verteilung in der Superpopulation überein, für die man sich eigentlich interessiert. Um die Parameter der Superpopulation in diesem Fall zu schätzen, muß ein gemeinsames Modell für Selektion und Merkmalsausprägung formuliert und geschätzt werden, vgl z.B. Little/Rubin (1987, Kapitel 11). In diesem Fall verliert der Superpopulationsansatz viel von seiner Attraktivität, die darin besteht, die Stichprobenauswahl ignorieren zu können.

Kehren wir nach diesen prinzipiellen Vorbehalten zu der Analyse von HS zurück. Um die Nullhypothese zu testen, ob Mikrozensus und Allbus- Beobachtungen aus einer Verteilung stammen, benutzten HS eine Logit–Analyse in den Variablen Geschlecht (G), Alter (A), Stellung im Beruf (S) und Bildungsabschluß (B). Hierbei wird der Logarithmus der beobachteten Häufigkeiten $m_{G,A,S,B,M}$ beim Mikrozensus mit $m_{G,A,S,B,ALL}$ beim Allbus verglichen.

[15]Die Auswahlwahrscheinlichkeit ist umgekehrt proportional zur (reduzierten) Haushaltsgröße.

Benutzt man die beiden log-linearen Modelle

$$ln(\mu_{G,A,S,B,M}) = \beta_M + \beta_{G,M} + \beta_{A,M} + \beta_{S,M} + \beta_{B,M} + \beta_{G,A,S,B} \qquad (3)$$

bzw.

$$\begin{aligned} ln(\mu_{G,A,S,B,ALL}) &= \beta_{ALL} + \beta_{G,ALL} + \beta_{A,ALL} + \beta_{S,ALL} + \beta_{B,ALL} \\ &+ \beta_{G,A,S,B} \end{aligned} \qquad (4)$$

für die Erwartungswerte μ der beobachteten Häufigkeiten m mit Koeffizientenvektor β. so erhält man die Koeffizienten der logistischen Regression $\lambda_{\cdot} = \beta_{\cdot,M} - \beta_{\cdot,All}$ durch Differenzenbildung von (3) und (4). vgl. McCullagh/Nelder (1983. Seite 140ff). Hierdurch wird jedoch nicht die ursprünglich aufgestellte Nullhypothese getestet. daß keine Unterschiede zwischen beiden Populationen bestehen. Denn trotz gleicher Haupteffekte können Interaktionsterme höherer Ordnung verschieden sein.[16] Das angewandte Verfahren testet lediglich die Haupteffekte von G, A. S und B.

Ein weiterer Beleg. den HS für die Unterschiedlichkeit von Mikrozensus– und Allbus–Population anführen, ist der schlechte Modellfit eines Logit-Modells. das nur aus der Konstanten besteht. In diesem Modell gilt a–priori $\beta_{G,M} = \beta_{G,ALL} \cdots, \beta_{B,M} = \beta_{B,ALL}$. Verglichen werden die unter diesem Modell erwarteten Verhältniszahlen mit den beobachteten Verhältniszahlen $m_{G,A,S,B,ALL}/m_{G,A,S,B,M}$. Diese Verhältniszahlen sind unter dem Modell für alle 600 möglichen Wertekombinationen der Variablen G, A, S und B gleich. Allerdings sind die beobachteten Zellenbesetzungen das Resultat der Verteilung der 3000 Befragten auf die 600 möglichen Wertekombinationen. Die χ^2–Approximation für die Devianz-Statistik, die als Maß für den Modellfit herangezogen wird, ist in diesem Fall sicher nicht angemessen, vgl. Read/Cressie (1988). Obwohl HS diesen Sachverhalt in einer Fußnote streifen ("Strenggenommen wird wegen der vielen Zellen mit kleinen Erwartungswerten nur ein heuristischer Charakter der Analysen beansprucht"), wird hiervon in der Argumentation der Analyse keine Notiz genommen, wo unter Hinweis auf die

[16]Das Endmodell von HS in Tabelle 9 berücksichtigt zwar auch Interaktionen zwischen A und S bzw A und B, jedoch ist das Basismodell in Tabelle 8 nicht mehr in diesem Modell geschachtelt. Trotzdem weisen HS in Tabelle 7 einen statistisch völlig sinnlosen "Test" zwischen diesen beiden Modellen aus.

Devianz von "großen und signifikanten Abweichungen der Allbus-Daten vom Mikrozensus"(S.333) gesprochen wird.[17]

8 Der Kern des Problems: Die Kenntnis der Teilnahmewahrscheinlichkeiten

Die von HS präsentierte Logit-Analyse läßt sich jedoch auch als Versuch interpretieren, die Teilnahmewahrscheinlichkeiten für den Allbus zu bestimmen. Sind diese Wahrscheinlichkeiten bekannt, so lassen sich innerhalb des Randomisierungsansatzes erwartungstreue Schätzer für die Populationswerte angeben.

Der Prozeß der Teilnahme an einer Umfrage läßt sich grob durch drei Stufen kennzeichnen:

- **Stufe 1**: Die Wahrscheinlichkeit der Auswahl von Einheit i durch das Stichprobendesign: $P(D_i = 1)$
- **Stufe 2**: Die Wahrscheinlichkeit der Kontaktaufnahme nach der Auswahl durch das Stichprobendesign: $P(K_i = 1 \mid D_i = 1)$
- **Stufe 3**: Die Wahrscheinlichkeit der Antwortgewährung (Response) nach der Kontaktaufnahme: $P(R_i = 1 \mid D_i = 1, K_i = 1)$

Das Problem von Umfragen besteht darin, daß die kumulativen Ausfälle nach der 1. Stufe (Stichprobendesignstufe) ca. 40% und mehr betragen. Werden die Prozesse, die zu diesen Ausfällen führen, nicht durch eine adäquate Schätzung von $P(K_i = 1 \mid D_i = 1)$ bzw. $P(R_i = 1 \mid D_i = 1, K_i = 1)$ beschrieben, so verlieren Populationsschätzungen die von Randomisierungsansatz behaupteten statistischen Eigenschaften. Umgekehrt gilt aber auch: Kennt man $P(K_i = 1 \mid D_i = 1)$ und $P(R_i = 1 \mid D_i = 1, K_i = 1)$, so können unverzerrte

[17] Ähnlich verfahren HS mit ihrem Pseudo-R^2, für das mangels Verteilungsaussagen ebenfalls lediglich Heuristik beansprucht wird. Doch wir vermögen nicht einzusehen, welche Heuristik eine Aussage wie: "Von empirisch maximal 3 Prozent der Pseudo-R^2 können durch dieses Modell 1.3 Prozent aufgeklärt werden" vermittelt. Sind 1.3 Prozent viel oder wenig? Und was sind eigentlich empirisch maximale Pseudo-R^2 (S.333)?.

Schätzer angegeben werden, selbst wenn die Ausfälle nicht gleichmäßig auftreten. Würde es tatsächlich zutreffen, daß die Ausfälle durch Nichterreichbarkeit und Antwortverweigerung durch ein geringes Bildungsniveau hervorgerufen werden - klassische Repräsentativitätsstudien würden von einem deutlichen "Bildungsbias" sprechen - so liefert eine Populationsschätzung, die auf diesen Teilnahmewahrscheinlichkeiten basiert, weiterhin erwartungstreue Schätzwerte. Allerdings vergrößern die unterschiedlichen Teilnahme-Wahrscheinlichkeiten die Varianz der Schätzer. In diesem Sinne verschlechtert ein ungleichmäßiger Ausfall die Qualität einer Stichprobe, weil die Schätzer weniger effizient werden. Sie sind aber nach wie vor "unbiased".

Der kritische Punkt ist jedoch, Kenntnis darüber zu erlangen, welche Merkmale die Ausfälle verursachen, und die Größe dieses Einflusses zu bestimmen. Dazu müssen Informationen über die nichtteilnehmenden Personen herangezogen werden. Dies wäre in hohem Maße wünschenswert, ist aber vor allem deswegen schwierig, weil bis auf Regionalmerkmale bei ausgefallenen Personen in der Regel keine weiteren Informationen erhoben werden. Eine Versuch, Basismerkmale von Nonrespondenten im Rahmen einer telephonischen Nachbefragung zu erheben, ist aber beim Allbus 1986 unternommen worden, vgl. Erbslöh/Koch (1988).

9 Kann Statistik zaubern ? Oder was passiert bei der Randanpassung ?

HS schätzen Teilnahmewahrscheinlichkeiten jedoch nicht auf Basis von Informationen über den Auswahlprozeß und die Merkmale von Nonrespondenten sondern auf Basis eines Vergleichs von zwei Stichproben (Mikrozensus und Allbus). Ihr Verfahren kann als indirekte Schätzung der Response–Wahrscheinlichkeiten im Rahmen des Postratifizierungsansatzes (vgl. Oh/-Scheuren (1983)) aufgefaßt werden. Dieser Ansatz beruht auf zwei wesentlichen Modellannahmen: a) Die Response–Wahrscheinlichkeiten sind konstant innerhalb der Merkmalskombinationen bestimmter bekannter Variablen, b) die Effekte der einzelnen Variablen kombinieren sich multiplikativ.

Ein wesentlich populäreres Verfahren, das auf denselben Grundannahmen beruht, ist das Verfahren der Randanpassung ("Redressment"). Aus der Sichtweise der Stichprobe als Miniatur der Grundgesamtheit erscheint es konsequent, darauf zu hoffen, daß alle Abweichungen der Stichprobenvertei-

lung von der entsprechenden Verteilung in der Grundgesamtheit verschwinden, wenn die Stichprobe so "transformiert" wird, daß diese Abweichungen für bestimmte, als wesentliche erachtete Merkmale der Grundgesamtheit verschwinden. Wie in Abschnitt 3 erläutert wurde, ist diese Erwartung unrealistisch. Schnell (1992) zeigt für den Allbus, daß eine Randanpassung in demographischen Variablen zu keiner verbesserten Anpassung von anderen Merkmalsverteilungen führt.

Da das Verfahren der Randanpassung trotz seiner Popularität kaum Eingang in die Lehrbücher gefunden hat, sei der postulierte Zusammenhang zu einer impliziten Schätzung der Teilnahmewahrscheinlichkeiten kurz skizziert.

Der Einfachheit halber soll das Verfahren der Randanpassung für zwei Merkmale X_a mit H Ausprägungen und X_b mit K Ausprägungen dargestellt werden.[18] Weiterhin bezeichne $\hat{n}_{hk} (h = 1, \ldots, H; k = 1, \ldots, K)$ die geschätzte Anzahl von Merkmalsträgern mit $X_a = h$ und $X_b = k$ in der Population. Bei der Randanpassung geht es darum, die Schätzwerte $\hat{n}_{hk}$ so zu modifizieren, daß sowohl die Randverteilungen bezüglich des ersten Merkmals als auch des zweiten Merkmals mit den Werten in der Grundgesamtheit $N_{\cdot k}$ bzw. $N_{h\cdot}$ übereinstimmen.

Für diese modifizierten Schätzungen $\tilde{n}_{hk}$ soll also gelten:

$$\sum_{h=1}^{H} \tilde{n}_{hk} = N_{\cdot k} \qquad k = 1, \ldots, K \tag{5}$$

$$\sum_{k=1}^{K} \tilde{n}_{hk} = N_{h\cdot} \qquad h = 1, \ldots, H \tag{6}$$

In einem ersten Schritt wird die Gültigkeit von (5) erreicht, indem man setzt:

$$\begin{aligned} \tilde{n}_{hk}^{(1)} &= \frac{N_{\cdot k}}{\sum_{h=1}^{H} \hat{n}_{hk}} \hat{n}_{hk} \\ &= a_k^{(1)} \hat{n}_{hk} \end{aligned} \tag{7}$$

Hierbei ist der Korrekturfaktor $a^{(1)}$ gerade das Verhältnis der Sollwerte in der Grundgesamtheit und der Istwerte (= geschätzte Werte) durch die Stichprobe.

[18] Mit offensichtlichen Verallgemeinerungen für mehr als zwei Merkmale.

Im nächsten Schritt wird nun - ausgehend von $\hat{n}_{hk}^{(1)}$ - Gleichung (6) angepaßt

$$\begin{aligned} \hat{n}_{hk}^{(2)} &= \frac{N_{h\cdot}}{\sum_{k=1}^{K} \tilde{n}_{hk}^{(1)}} \tilde{n}_{hk}^{(1)} \\ &= b_h^{(1)} a_k^{(1)} \hat{n}_{hk} \end{aligned} \tag{8}$$

Dieses Verfahren wird nun in einer zweiten Runde auf das erste, dann wieder auf das zweite Merkmal angewandt. Dies liefert $a_k^{(2)}$ und $b_h^{(2)}$. Unter geeigneten Voraussetzungen läßt sich zeigen, daß $a_k^{(l)}$ und $b_h^{(l)}$ für $l \to \infty$ konvergieren, vgl. Ireland/Kullback (1968). Für die Grenzwerte a_k und b_h gilt:

$$\sum_{h=1}^{H} a_k b_h \hat{n}_{hk} = N_{\cdot k} \quad , \quad k = 1, \ldots, K \tag{9}$$

$$\sum_{k=1}^{K} a_k b_h \hat{n}_{hk} = N_{h\cdot} \quad , \quad h = 1, \ldots, H \tag{10}$$

Folglich erfüllt die korrigierte Populationsschätzung $\hat{n}_{hk} = a_k b_h \hat{n}_{hk}$ simultan die Randbedingungen.
Ist $\hat{n}_{hk}$ von der Form

$$\hat{n}_{hk} = \sum_{i=1}^{N} \frac{C_i}{P(D_i = 1)} X_i^{hk} \tag{11}$$

wobei X_i^{hk} den Wert 1 annimmt, falls $X_a = h$ und $X_b = k$ für die i-te Einheit gilt, so läßt sich $a_k b_h$ formal als Kehrwert von $P(R_i = 1, K_i = 1 \mid D_i = 1)$ auffassen, falls $a_k b_h \geq 1$. Letztere Bedingung dürfte bei Ausfallraten um 40% fast immer erfüllt sein. Die Modellgleichung lautet dann:

$$P(R_i = 1, K_i = 1 \mid D_i = 1, X_a = h, X_b = k) = \frac{1}{a_k} \cdot \frac{1}{b_h} \tag{12}$$

In ähnlicher Weise läßt sich die logistische Regression von HS als indirekte Schätzung der Auswahlwahrscheinlichkeiten des Allbus auffassen. Für die

Wahrscheinlichkeit $P(\tilde{C}_i = 1)$, daß eine Einheit i über den Allbus statt über den Mikrozensus beobachtet wird, lautet das logistische Modell - wiederum bei Verwendung zweier Merkmale mit Ausprägungen h=1,...,H und k=1,...,K:

$$P(\tilde{C}_i = 1 \mid X_a = h, X_b = k) = \frac{e^{\beta_0+\beta_h+\beta_k}}{1 + e^{\beta_0+\beta_h+\beta_k}} \tag{13}$$

Um die Interpretation von $\tilde{C}_i$ als Auswahlindikator aus der Grundgesamtheit zu sichern, müssen die Unterschiede zwischen dem Mikrozensus und der Grundgesamtheit vernachlässigt werden. Aufgrund des geringen Auswahlsatzes des Allbus ist $e^{\beta_0} \ll 1$, so daß der Nenner in (13) vernachlässigt werden kann, wenn bezüglich der Merkmale X_a und X_b keine sehr großen Differenzen in den Auswahlwahrscheinlichkeiten auftreten. Approximativ gilt damit:

$$P(\tilde{C}_i = 1 \mid X_a = h, X_b = k) \approx e^{\beta_0} e^{\beta_h} e^{\beta_k} \tag{14}$$

Die Schätzung der Teilnahmewahrscheinlichkeiten über den Logitansatz ist damit formal identisch dem Ergebnis der Randanpassung.[19]

Obwohl beide Verfahren als implizite Schätzung eines Haupteffekt-Modells für die Teilnahmewahrscheinlichkeiten interpretiert werden können, darf nicht vergessen werden, daß diese Wahrscheinlichkeiten nur indirekt unter Zuhilfenahme restriktiver Modellannahmen geschätzt werden. Die Gültigkeit des Modellansatzes läßt sich hierbei - mangels Daten über die Teilnahmeverweigerer - nicht überprüfen. Sind die Modell-Bedingungen verletzt, so können die Schätzwerte und die angegebenen Signifikanzniveaus systematische Abweichungen von ihren Erwartungswerten zeigen.[20]

Es sei an dieser Stelle vermerkt, daß das Bild von der Stichprobe als Miniatur der Grundgesamtheit keine Hinweise darauf gibt, welche Variablen in die Randanpassung aufgenommen werden. Außer der Tatsache, daß das betreffende Merkmal für die Grundgesamtheit bekannt sein muß, gibt es nur

[19] Es läßt sich allerdings zeigen, daß diese Schätzung nicht notwendig eine Randanpassung in X_a und X_b liefert.

[20] Da im Allbus die Auswahlwahrscheinlichkeit vom Design her umgekehrt proportional zur (reduzierten) Haushaltsgröße ist, muß die Variable "Haushaltsgröße" in eine multivariate Analyse der Auswahlwahrscheinlichkeiten aufgenommen werden, wenn man das Logitmodell von HS als Schätzung der Auswahlwahrscheinlichkeiten interpretiert. Durch die Fixierung auf einen "Bildungsbias" wird dies bei HS übersehen. Die Auswahl der Variablen ist ganz auf dieses Ziel orientiert.

einen unter Sozialstrukturforschern diffusen Grundkonsens, daß Alter, Geschlecht, Bildung und Erwerbsstatus immer dazugehören. Die Einbettung in den Poststratifizierungsansatz gibt jedoch eine systematische und im allgemeinen andere Antwort: Es sollten die Merkmale sein, die den Auswahlprozeß charakterisieren. Diese Merkmale können aber für die Grundgesamtheit unbekannt sein oder nicht Bestandteil der "klassischen" Anpassungsmerkmale sein.

Fehlt eine der für den Ausfall relevanten Variablen bei der Randanpassung, so liefert dieses Verfahren sehr irreführende Ergebnisse. Wir verweisen an dieser Stelle auf das Beispiel von Rothe (1990), der den über den Allbus geschätzten Anteil von arbeitslosen Personen mit dem entsprechenden Wert im Mikrozensus vergleicht. Hier vergrößert sich durch die Verwendung einer Randanpassung in den Variablen Alter, Geschlecht, Bildung und Stellung im Beruf der Abstand des geschätzten Populationswerts zum Mikrozensus-Wert.[21] Wie Rothe bemerkt, ist dieser Effekt der Randanpassung plausibel, wenn man unterstellt, daß a) Arbeitslose häufiger Nonrespondenten sind als andere Personen und b) Arbeitslose in der Gruppe der Arbeiter häufiger vertreten sind als in anderen Gruppen. Dann sind nämlich in der Stichprobe unterdurchschnittlich viele Arbeiter. Aus diesem Grunde werden die Arbeiter durch die Randanpassung "hochgewichtet". Innerhalb der Gruppe der in der Stichprobe verbliebenen Arbeiten sind jedoch unterdurchschnittlich viele Arbeitslose. Dieses unterdurchschnittliche Verhältnis wird durch die Gewich-

[21]Allbus ungewichtet: 3.59%, Allbus Design gewichtet: 3,20%, Allbus mit Randanpassung: 2.28%, Mikrozensus: 3.70%, vgl. Tabelle 2 in Rothe (1990).

tung noch verstärkt.[22]

Dies mag dazu verführen, sehr viele Merkmale für die Randanpassung zu benutzen. Allerdings wird dabei übersehen, daß durch eine solche Vorgehensweise die Varianz der (geschätzten) Auswahlwahrscheinlichkeiten unter Umständen gewaltig vergrößert wird und damit auch die Varianz der Schätzergebnisse. Es zeigt sich hier der in der Statistik immer wiederkehrende Trade-off zwischen Bias und Varianz.

Der Postratifizierungsansatz steht und fällt mit der Gültigkeit des implizit verwendeten Modells für die Stichprobenausfälle. Sozialstrukturforscher wiegen sich häufig in der trügerischen Sicherheit einer Stichprobe, die durch eine Transformation (sprich Randanpassung) in eine Miniatur der Grundgesamtheit verwandelt wird. Schön wärs, leider gibt es auch im Statistik-Land keine Zauberer. Diese Tatsache wird von Statistikern immer wieder betont und es scheint uns, daß umgekehrt Sozialstrukturforscher an dieser Stelle "schwarze Magie" betreiben, indem sie dem Gewichtungsverfahren Eigenschaften unterstellen, die in dieser Form unhaltbar sind.

10 Gibt es einen Bildungsbias ?

Insgesamt halten wir die Argumente, die HS für einen Bildungsbias anführen, für nicht überzeugend. ("Die Analyse belegt, daß der sogenannte Mittelschichtsbias der Umfrageforschung im Kern ein Bildungsbias und kein Effekt der Klassenlage ist. Er ist theoretisch als Effekt kognitiver Ressourcen zu

[22] Rothes Plausibilitätsbetrachtung sei folgendermaßen präzisiert: Es sei n_{AS} die Anzahl der Personen in der Stichprobe mit Merkmal A (0 = arbeitslos, 1 = beschäftigt) und Merkmal S (0 = Arbeiter, 1 = Angestellter). Die Anzahl der Arbeiter in der Population sei mit $N_{\cdot 0}$ und die Anzahl der Angestellten mit $N_{\cdot 1}$ gegeben. Das Verhältnis κ der Arbeitslosenquote von Stichprobe und angepaßter Stichprobe ist:

$$\kappa = \frac{n_{00}+n_{01}}{n} \frac{N}{n_{00}\frac{N_{\cdot 0}}{n_{00}+n_{10}} + n_{01}\frac{N_{\cdot 1}}{n_{01}+n_{11}}}$$

Setzt man $\frac{n_{00}}{n_{01}} = \mu \frac{n_{\cdot 0}}{n_{\cdot 1}}$, so gilt $\kappa = 1$, falls $\mu = 1$. Dies entspricht gerade der Unabhängigkeit der Merkmale A und S in der Stichprobe. Für $\mu < 1$ ist κ größer als 1. In diesem Fall wird der Arbeitslosenanteil durch die Anpassung "herunter gewichtet". Die Bedingung besagt, daß in der Stichprobe das Verhältnis von Arbeitern zu Angestellten bei den Arbeitslosen kleiner ist als bei den Beschäftigten. Dies kann z. B. dadurch verursacht sein, daß arbeitslose Arbeiter eine besonders hohe Ausfallrate haben.

betrachten. Die partielle Verzerrung in der Variablen Stellung im Beruf läßt sich nach Dichotomisierung in erwerbstätig/nicht erwerbstätig als Effekt der Erreichbarkeit verstehen"; Seite 337.)

Solange beim Mikrozensus nicht zwischen der untersten Bildungskategorie und Item-Nonresponse unterschieden werden kann,[23] ist der doppelt so hohe Anteil von Hauptschülern ohne Lehre im Mikrozensus, der die Basis dieses Resultats ist, unglaubwürdig.

Weiterhin müßten sich derartig gravierende Unterschiede in der Teilnahmebereitschaft auch in der Nonresponsestudie des Allbus gezeigt haben, was aber nicht der Fall war, vgl. Erbslöh/Koch (1988). Es ist interessant zu sehen, daß Erbslöh/Koch selbst bei diesen Ergebnissen nicht an der Existenz eines Mittelschicht-Bias zweifeln, sondern umgekehrt die Validität der Nonresponse-Studie in Frage stellen.[24] Auch beim SOEP zeigte sich bei einer Befragung der Nonrespondenten bei Start des Panels kein Hinweis auf einen Mittelstandsbias, vgl. Rendtel (1988) und Wagner/Schupp/Rendtel (1991).

Daß die Schulbildung — via "kognitive Ressource" — ursächlich die Teilnahmebereitschaft in Umfragen bestimmt, erscheint uns wenig plausibel. Die Erfahrungen aus dem SOEP sprechen eher dafür, daß die Gewährung eines persönlichen Interviews in erster Linie durch das Vertrauen zwischen Befragtem und Interviewer bestimmt wird, vgl. Rendtel (1990). Dieser Sachverhalt ist sicher schwer zu operationalisieren. Und in bestimmten Situationen mag ein mangelndes Vertrauen zwischen Befragtem und Interviewer tatsächlich mit dem Bildungsniveau der Befragten korrelieren.

Bis zum Beweis des Gegenteils halten wir den hier postulierten Bildungsbias für ein bloßes Konstrukt. Um die Abhängigkeit der Antwortbereitschaft vom Bildungsniveau nachweisen zu können, müssen entsprechende Feldinformationen über die Nonrespondenten herangezogen werden. Da HS die (vorhandenen) Informationen nicht nutzen, bleiben ihre Ausführungen über den Ausfallprozeß empirisch unbelegte Spekulation.

---

[23] Dies ist seit dem Mikrozensus 1991 (!) möglich.

[24] "... so wäre für die erfaßten Non-Responses — unter der Annahme, daß sie repräsentativ für sämtliche Nicht-Teilnehmer sind — zu erwarten, daß sie in komplementärer Form diese Verzerrungen der ALLBUS-Haupterhebung widerspiegeln", was aber nicht der Fall ist. Daher die Folgerung: "Insgesamt deuten diese Ergebnisse darauf hin, daß die 'typischen' Non-Responses, zumindest was die hier dargestellten Merkmale betrifft, auch an der telephonischen Zusatzbefragung nicht teilgenommen haben."

11 Sind Sozialstrukturanalysen mit Umfragedaten möglich ?

Kommen wir abschließend zu der von HS gestellten Frage, ob Sozialstrukturanalysen mit Umfragedaten möglich sind.

Wir möchten die Frage wie folgt präzisieren: Ist es möglich mit Umfragedaten soziale Populationsmerkmale bzw. Regressionsbeziehungen zwischen sozialen Charakteristika erwartungstreu bzw. konsistent zu schätzen? Diese Frage soll dabei im Hinblick auf den Selektionsprozeß der Daten behandelt werden. Der Selektionsprozeß hat zwei Komponenten: Einmal das Erhebungsdesign, das bestimmte Einheiten mit eventuell unterschiedlichen Wahrscheinlichkeiten in die Auswahl einbezieht, und zum anderen die feldbedingten Ausfälle, die hauptsächlich von HS problematisiert werden.

Hierbei soll im folgenden zwischen der Schätzung eines Populationsparameters P nach dem Randomisierungsansatz und der Schätzung einer Regressionsgleichung nach dem Modell-basierten Ansatz unterschieden werden.

Das Erhebungsdesign und gegebenenfalls die Variablen des Postratifizierungsansatzes definieren Bevölkerungsgruppen. Sind innerhalb dieser Bevölkerungsgruppen die Auswahlwahrscheinlichkeiten für alle Personen gleich, so können im Rahmen des Randomisierungsansatzes weiterhin erwartungstreue Schätzer für den Populationsparameter P angegeben werden.[25]

Bei der Schätzung einer Regressionsgleichung im Rahmen des Modell-basierten Ansatzes sollen die Parameter der bedingten Verteilung der abhängigen Variablen gegeben die Kovariaten geschätzt werden. Ist diese bedingte Verteilung für die beobachteten und die nicht beobachteten Personen gleich, so können die Parameter dieser Verteilung allein auf Basis der beobachteten Personen geschätzt werden. Der Selektionsprozeß heißt dann ignorierbar, vgl. Dawid (1979 a,b), Rubin (1976), Smith (1991).

Aus dieser Charakterisierung folgt, daß die Ignorierbarkeit des Selektionsprozesses im Modell-basierten Ansatz nicht für eine Umfrage als Ganzes attestiert werden kann oder verworfen werden muß. Vielmehr muß diese Frage für einzelne Analysen getrennt beantwortet werden. Schon der Wechsel der kontrollierenden Variablen kann zu einem unterschiedlichen Resultat

[25] Ist diese Annahme nicht erfüllt, so muß der Randomisierungsansatz verlassen werden und auf Modell-basierte Ansätze zurückgegriffen werden, vgl. Hansen/Madow/Tepping (1983, Seite 791).

hinsichtlich der Ignorierbarkeit der Selektion führen.

Trotz der Betonung der feldbedingten Ausfälle bei HS sollte man nicht übersehen, daß die Design-Auswahl bestimmter Umfragen — z.B. die Überrepräsentation bestimmter soziale Schichten — bei der Regressionsanalyse in bestimmten Fällen nicht ignorierbar sein kann; beispielsweise wenn die Schichtungsvariable nicht als kontrollierende Variable aufgenommen wird oder selbst die zu erklärende Variable ist, vgl. Nathan/Holt (1980).[26]

Wir weisen nochmals darauf hin, daß die Ignorierbarkeit der Selektion auch in Situationen gegeben sein kann, in denen die Randverteilungen von Datensatz und Population große Unterschiede aufweisen. Eine derartige "Verzerrung " äußert sich allenfalls in Effizienzverlusten, schlecht konditionierten Momentenmatrizen. etc..

Damit stellt sich aber die Frage, wie man im Fall einer Regressionsanalyse Anhaltspunkte für die Ignorierbarkeit des Auswahlprozesses gewinnen kann. Hierzu bedarf es einer gewissen Kenntnis, welche Selektionsprozesse realistischerweise überhaupt in Frage kommen.

Wenn spezifische Selektionsprozesse etwa über die Teilnahmebereitschaft bestimmter sozialer Gruppen vermutet werden, so enthält die gegebene Stichprobe keinerlei Informationen über die Charakteristika der nicht teilnehmenden Zielpersonen. Die einzige Quelle für diese benötigten Informationen sind Nonrespondenten Studien. Schon bei Planung der Umfrage sollte man die Möglichkeit einbeziehen. Nonrespondenten dahingehend zu motivieren, wenigstens zu Basisgrößen Angaben zu machen. Weiterhin sollte diese Angaben sowie weitere Feldinformationen — z.B. über die Gründe des Ausfalls. die Anzahl der Kontaktversuche und Merkmale der Interviewer — integraler Bestandteil des Datensatzes werden.

Weiterhin kann man die Sensitivität der Analyseergebnisse unter hypothetischen Ausfallprozessen untersuchen. Vermutet man zum Beispiel, daß eine Analyse durch Nichterreichbarkeit beeinflußt werden kann, so läßt sich dies an der Stichprobe simulieren. Beispielsweise benutzt Schnell (1992) die Dokumentation der Kontaktberichte der Interviewer im Allbus 1980 und variiert die durch die Zahl der Kontaktversuche geschätzte Erreichbarkeit. Sukzessiv werden aus der Stichprobe die Beobachtungen mit mehr als 6, 5, 4, ... bis 2 Kontaktversuchen entfernt und die resultierenden Regressionsschätzungen

[26]Bei der Schätzung von Populationswerten ist die Schichtungsvariable beim Oversampling in der Regel bei der Berechnung der Hochrechungsfaktoren berücksichtigt.

werden miteinander verglichen.

Ganz allgemein kann man einen vermuteten Ausfallprozeß an der Stichprobe simulieren und die jeweils so reduzierte "Stichprobe" als Basis für Regressionsschätzungen wählen. Die Verteilung dieser Schätzwerte um den Wert, den man auf Basis der Originalstichprobe erhält, kann man als Maß für den Einfluß des Ausfallprozesses auf die Schätzung der Regressionsparameter angesehen werden.

Erweist sich ein Selektionsprozeß in einer bestimmten Situation als nicht ignorierbar, so sind zwei Auswege möglich: a) Man versuche den Selektionsprozeß durch die Aufnahme neuer Kontrollvariablen ignorierbar zu machen, b) man schätze ein gemeinsames Modell für Selektionsprozeß und Analysevariablen, vgl Little/Rubin (1987, Kapitel 11 und 12).

Die Aufgabe, ein gemeinsames Modell für Selektion und Merkmalsausprägungen zu formulieren und zu schätzen, wird durch die Stichprobenausfälle erheblich erschwert, da über die ausgefallenen Zielpersonen nur wenig Information vorhanden ist, um solche Modelle zu validieren. Generell gilt, daß die Schätzergebnisse derartiger simultaner Modelle ausgesprochen sensibel auf eine Modifikation des Selektionsmodells reagieren, vgl. Little/Rubin (1987, Seite 225 ff) und Rendtel (1992).

In der Praxis wird hingegen oft die Möglichkeit genutzt, Regressionsanalysen gewichtet zu rechnen.[27] Ohne an dieser Stelle einen langwierigen Disput vom Zaune brechen zu wollen,[28] möchten wir nur bemerken, daß noch immer eine generelle Begründung aussteht, warum ein derartiges Schätzverfahren konsistente Parameter-Schätzungen unter nicht ignorierbarer Selektion liefert. Die über Modelle abgeleiteten Schätzer haben jedenfalls eine gänzlich andere Struktur, vgl. Little/Rubin (1987, Kapitel 11).

Insgesamt ergibt sich damit durchaus die Möglichkeit, Sozialstrukturanalysen mit Umfragedaten durchzuführen. Allerdings erfordert dies eine stärkere Berücksichtigung der möglichen Auswirkungen von Ausfällen als dies üblicherweise bei Analysen geschieht. Hierfür ist allerdings die entsprechende Feldinformation unerläßlich.

[27] Auch HS weisen neben ungewichteten auch gewichtete Logit-Analysen in den Tabellen 8 und 9 aus.

[28] Wir verweisen an dieser Stelle auf Hoem (1989) und Little (1991) sowie die dort ziterte Literatur.

12 Resümee

Die Forderung nach "Repräsentativität" führt unserer Ansicht nach in eine Sackgasse. Keine Umfrage kann die selbstwidersprüchlichen Anforderungen der "Repräsentativität" erfüllen. Daraus könnte geschlossen werden, daß aus Umfragedaten auch nichts Vernünftiges gelernt werden kann. Ein solcher Rigorismus würde konsequent jede Möglichkeit verneinen, aus Beobachtungen zu lernen, die nicht den willkürlichen Kriterien der "Repräsentativität" genügen. Der Vorwurf nicht "repräsentativer" Beobachtungen wurde häufig benutzt, um Untersuchungen wie Fallstudien oder "qualitative" Forschungen zugunsten von Umfrageforschung zu diskreditieren. Die Anwendung solcher "Kriterien" auf die Umfrageforschung selbst, so stellt sich nun heraus, vermag diese ebenfalls als nicht "repräsentativ" zu disqualifizieren.

Wir möchten daher vorschlagen, auf den oft mißbrauchten[29] und nie genau definierten Begriff "repräsentative Stichprobe"[30] zu verzichten. Es kann nur darum gehen zu fragen, was man aus einer gegebenen Umfrage lernen kann und wie dies geschieht. Die Vermeidung des schwammigen Etiketts "repräsentativ" würde die Nutzer von Umfragedaten motivieren, ihre Erwartungen an den Datensatz deutlicher zu formulieren.

Mit der Nichtbenutzung des Wortes "repräsentativ" entfiele auch die damit verbundene "Repräsentativitätsstudie".

Weiterhin plädieren wir für eine veränderte Praxis von Sozialstrukturanalysen. Schon bei Planung der Umfrage sollte man die Möglichkeit einbeziehen, die Nonrespondenten dahingehend zu motivieren, wenigstens zu Basisgrößen Angaben zu machen. Weiterhin sollten diese Informationen sowie weitere Feldinformationen integraler Bestandteil des Datensates werden. Schließlich sollte man bei der Präsentation von Tabellen auch die 'keine–Angabe' Fälle ausweisen und wenigstens für besonders markante Populationswerte Konfidenzintervalle bzw. Varianzen ausweisen.

---

[29] Wir werden den Verdacht nicht los, daß der Hinweis, die Stichprobe sei "repräsentativ", von Nichtstatistikern gerne als Freibrief zu einem völlig unstatistischen Umgang mit Umfragedaten benutzt wird.

[30] "In survey literature, we often come across the term representative sample. To my knowledge the term has never been properly defined." (Basu,1971, Seite 239).

Literatur

[1] Arminger, Gerhard (1990):*Pflicht- versus Freiwilligenerhebung im Mikrozensus.* Allgemeines Statistisches Archiv, 74, 161–187.

[2] Bartholmai, Bernd; Melzer, Manfred; Schulz, Erika (1990):*Privathaushalte und Wohunugsbedarf in Deutschland bis zu Jahr 2000* . DIW Wochenbericht 42/90, 591–597, Berlin.

[3] Basu, Debabrata (1971):*An Essay on the logical foundations of survey sampling , Part one.* In : Godambe,V. ; Sprott,D. (eds) : Foundations of statistical inference . Holt Rinehart Winston, Toronto, 203–242.

[4] Berkson, Joseph (1942):*Tests of Significance considered as evidence.* Journal of the American Statistical Association, 37, 325-335.

[5] Böltken, Ferdinand (1976):*Auswahlverfahren. Eine Einführung für Sozialwissenschaftler.* Teubner, Stuttgart.

[6] Cassel, Claes–Magnus; Särndal, Carl–Erik; Wretman, Jan (1977):*Foundations of Inference in survey sampling.* Wiley, New York.

[7] Cochran, William (1977): *Sampling Techniques, Third Edition.* Wiley, New York.

[8] Dawid, A. (1979a): *Conditional Independence in Statistical Theory* . Journal of the Royal Statistical Society, B, 41, 1–31.

[9] Dawid, A. (1979b):*Some Misleading Arguments Involving Conditional Independence.* Journal of the Royal Statistical Society, B, 41, 249–252.

[10] Erbslöh, Barbara; Koch, Achim (1988):*Die Non–Response–Studie zum Allbus 1986: Problemstellung, Design, erste Ergebnisse* . ZUMA Nachrichten Nr. 22 , Mannheim.

[11] Esser, Hartmut; Grohmann, Heinz; Müller, Walter; Schäffer, Karl-August (1989):*Mikrozensus im Wandel.* Metzler-Poeschel, Stuttgart.

[12] Hansen, Morris; Madow, William; Tepping, Benjamin (1983): *An evaluation of model-dependent and probability sampling inferences in sample surveys.* Journal of the American Statistical Association, 70, 776–807.

[13] Hoem, Jan (1989):*The Issue of Weights in Panel Surveys of Individual Behavior.* In: Kasprzyk, Daniel et al. (eds): Panel Surveys, Wiley, New York, 539–565.

[14] Horvitz, D. ;Thompson, D. (1952): *A Generalization of Sampling without Replacement from a Finite Universe.* Journal of the American Statistical Association, 47, 663–685.

[15] Ireland, C. and Kullback, Solomon (1968): *Contingency tables with given marginals.* Biometrica, 55, 179–188.

[16] Kirschner, Hans-Peter (1984): *Allbus 1980 — Stichprobenplan und Gewichtung.* In: Mayer, Karl-Ulrich; Schmidt, Peter (Hrsg:), Allgemeine Bevölkerungsumfrage der Sozialwissenschaften - Beiträge zu methodischen Problemen des Allbus 1980, Frankfurt, S. 114–182.

[17] Kruskal, William; Mosteller, Federick (1979a):*Representative sampling. I: Non-scientific Literature* ,International Statistical Review, 47, 13–24.

[18] Kruskal, William; Mosteller, Federick (1979b):*Representative sampling, II: Scientific Literature, Excluding Statistics* , International Statistical Review, 47, 111-127.

[19] Kruskal, William; Mosteller, Federick (1979c):*Representative sampling. III: The Current Statistical Literature* , International Statistical Review, 47, 245–265.

[20] McCullagh, Peter; Nelder, John (1983):*Generalized Linear Models.* Chapman Hall, London.

[21] Nathan, G. ; Holt, D. (1980):*The Effect of Survey Design on Regression Analysis.* Journal of the Royal Statistical Society, B, 42, 377–386.

[22] Lang, Serge (1981): *The File case Study in Correction (1977-1979).* Springer, New York.

[23] Little, Roderick (1982):*Models for Nonresponse in Sample surveys.* Journal of the American Statistical Association, 77, 237–250.

[24] Little, Roderick (1991):*Inference with Survey Weights.* Journal of Official Statistics, 7, 405–424.

[25] Little, Roderick; Rubin, Donald (1987):*Statistical Analysis with Missing Data*. Wiley, New York.

[26] Oh, H. Lock; Scheuren, Federick (1983): *Weighting Adjustment for Unit Nonresponse*. In: Madow et al. (eds): Incomplete Data in Sample Surveys, Vol 2, Academic Press, ,New York, 143–184.

[27] Prester, Hans-Georg (1992):*Die Zahl der Haushalte — ein Problem der Marktforschungspraxis* . Planung und Analyse , Heft 2/92 , 50–54.

[28] Read, Timothy; Cressie, Noel (1988):*Goodness–of–Fit Statistics for Discrete Multivariate Data* . Springer, New York.

[29] Rendtel, Ulrich (1988): *Repräsentativität und Hochrechnung der Datenbasis* . in: Krupp, Hans–Jürgen; Schupp, Jürgen (Hrg): Lebenslagen im Wandel: Daten 1987: Campus, Frankfurt/New York, 289–308.

[30] Rendtel, Ulrich (1990): *Teilnahmebereitschaft in Panelstudien: Zwischen Beinflussung, Vertrauen und Sozialer Selektion.* Kölner Zeitschrift für Soziologie und Sozialpsychologie, 42, 280-299.

[31] Rendtel, Ulrich (1991):*Die Schätzung von Populationswerten in Panelerhebungen* . Allgemeines Statistisches Archiv, 75, 225–244.

[32] Rendtel, Ulrich (1992):*On the Choice of a Selection–Model When Estimating Regression Models With Selectivity* . DIW Discussion Papers No. 53; Berlin.

[33] Rothe, Günter (1990):*Wie (un)gewichtig sind Gewichtungen ? Eine Untersuchung am Allbus 1986* . ZUMA Nachrichten Nr. 26, Mannheim.

[34] Rubin, Donald (1976):*Inference and Missing Data*. Biometrica, 63, 581–592.

[35] Särndal, Carl–Erik (1978):*Design–based and Model–based Inference in Survey Sampling*. Scandinavian Journal of Statistics, 5, 27–52.

[36] Schnell, Rainer (1992): *Die Homogenität als Voraussetzung für "Repräsentativität" und Gewichtungsverfahren*. erscheint in: Zeitschrift für Soziologie.

[37] Schnell, Rainer (1991):*Wer ist das Volk ? Zur faktischen Grundgesamtheit bei "allgemeinen Bevölkerungsumfragen": Undercoverage, Schwererreichbare und Nichterreichbare.* Kölner Zeitschrift für Soziologie und Sozialpsychologie, 43, 106–137.

[38] Smith, Terence M. F. (1991):*Post-stratification.* The Statistician, 40, 315–323.

[39] Wedel, Edgar (1989):*Haushalte 1987 — Methode und Ergebnis der Volkszählung* , Wirtschaft und Statistik, Heft 5/1989, 273–276.

[40] Wolter, Kirk (1985): *Introduction to Variance Estimation*, Springer, New York.

[41] Wagner, Gert; Schupp, Jürgen; Rendtel, Ulrich (1991):*Das Sozio-ökonomische Panel — Methoden der Datenproduktion und -aufbereitung im Längsschnitt.* In: DIW Diskussionspapier Nr. 31 , Berlin (erscheint 1992 in : Hauser, R. et al. (Hrsg.) : Mikroanaytische Grundlagen der Gesellschaftspolitik — Erhebungsverfahren, Analyseverfahren und Mikrosimulation).