

Die Berechnung der Varianz von Populationsschätzern im Scientific Use File des Mikrozensus ab 1996

Rendtel, Ulrich; Schimpl-Neimanns, Bernhard

Veröffentlichungsversion / Published Version

Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:

GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:

Rendtel, U., & Schimpl-Neimanns, B. (2001). Die Berechnung der Varianz von Populationsschätzern im Scientific Use File des Mikrozensus ab 1996. *ZUMA Nachrichten*, 25(48), 85-116. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-208020>

Nutzungsbedingungen:

Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:

This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

DIE BERECHNUNG DER VARIANZ VON POPULATIONSSCHÄTZERN IM SCIENTIFIC USE FILE DES MIKROZENSUS AB 1996*

ULRICH RENDTEL UND BERNHARD SCHIMPL-NEIMANNS

Erstmals wurden im Scientific Use File des Mikrozensus 1996 Stichprobeninformationen bereitgestellt, die eine Berechnung der Varianz von Populationsschätzern ermöglichen. Nach der Darstellung der Ziehung des Mikrozensus und des Scientific Use Files wird in Anlehnung an Särndal et al. (1992) ein methodisches Konzept zur Berechnung der Varianz entwickelt und auf die Schätzung von Totals, Verhältnis- und Mittelwerten angewendet. Für ausgewählte Merkmale werden Varianzschätzungen auf der Basis des Scientific Use Files mit Ergebnissen des Statistischen Bundesamtes verglichen. Zur Behandlung von hochgerechneten, an die Bevölkerungsfortschreibung angepassten Mikrozensusergebnissen wird eine Regressionschätzung vorgestellt. Erhebliche Differenzen zwischen angepassten und nicht angepassten Schätzergebnissen werfen die Frage verzerrter Populationsschätzungen auf. Abschließend wird der Nutzen von Approximationen der Varianz über die Design-Effekte diskutiert. Es zeigt sich, dass dieses Hilfsmittel auch für das Scientific Use File zu brauchbaren Ergebnissen führt, in Einzelfällen aber mit erheblichen Über- bzw. Unterschätzungen der Varianz verbunden ist. Mit dem Vorliegen der Stichprobeninformationen ab dem Scientific Use File des Mikrozensus 1996 ist es jedoch nicht mehr nötig, diese grobe Varianzschätzung zu verwenden.

* Diese Arbeit entstand als Resultat eines Gastaufenthalts des Erstautors im Oktober 1999 bei ZUMA. Für anregende Diskussionen danken wir Wolf Bihler (Statistisches Bundesamt), Ralf Münnich (Universität Tübingen) sowie Siegfried Gabler, Sabine Häder und Michael Wiedenbeck (ZUMA). Eine erste Fassung des Manuskripts wurde auf der 2. Nutzerkonferenz „Forschung mit dem Mikrozensus“ am 12.-13.10.2000 in Mannheim vorgetragen und liegt in einer kürzeren Version vor (Rendtel/Schimpl-Neimanns 2000). Wir danken Wolf Bihler, Ulrich Pötter und Götz Rohwer sowie zwei Gutachtern der ZUMA-Nachrichten für hilfreiche Kommentare zu einer früheren Fassung.

The Scientific Use File of the German Microcensus 1996 was the first to provide anonymized sample information necessary for calculating variances. The paper outlines the sampling design of the 1996 Microcensus and the selection procedure of the Scientific Use File. Following Särndal et al. (1992), a procedure is presented which uses the available design information to calculate the variance of parameter estimates for totals, ratios, and means. The variance estimates based on the Scientific Use File are compared with results from the Federal Statistical Office. A regression estimator (group mean model) is presented for the post-stratification of the Microcensus estimates. Large differences between the post-stratified data and the unadjusted data raise the question as to which of the two population estimates is biased. We then discuss the behaviour of the linear regression of the design effect, which is frequently used as a tool for variance estimation. In general terms this provides reasonable results, but in individual cases leads to either considerable over- or underestimation. The sample design information now available for the Microcensus means researchers no longer need to rely on such coarse approximations.

1. Einleitung

Mit einem Stichprobenumfang von einem Prozent der Personen und Haushalte werden im Mikrozensus jährlich vielfältige Informationen über die demographische, soziale und wirtschaftliche Struktur der Bevölkerung erhoben (vgl. Emmerling/Riede 1997; Lüttinger/Riede 1997). Der Forschung steht eine faktisch anonymisierte 70-Prozent-Substichprobe des Mikrozensus als Scientific Use File mit über 500.000 Personen in rund 230.000 Haushalten zur Verfügung.¹ Dieser große Stichprobenumfang erlaubt die differenzierte Analyse auch kleinerer Bevölkerungsgruppen. Neben der Stichprobengröße ist der sehr kleine Unit-Nonresponse hervorzuheben. Auf Grund der Auskunftspflicht liegt die Teilnahmequote der Haushalte bei rund 97 Prozent. In diesem Zusammenhang und unter der Annahme, dass mit Mikrozensusdaten Verteilungs- und Zusammenhangsmaße für die Population verlässlich geschätzt werden können, dient der Mikrozensus häufig als Referenzstatistik für die normalen Bevölkerungsumfragen, die vielfach Ausschöpfungsquoten unter 60 Prozent aufweisen.

Bei der Berechnung von Stichprobenfehlern ist zu berücksichtigen, dass der Mikrozensus keine uneingeschränkte Zufallsauswahl, sondern eine geschichtete Klumpenstichprobe ist. Weil die Scientific Use Files des Mikrozensus bis einschließlich 1995

1 Vgl. zu den Anonymisierungsmaßnahmen im Mikrozensus Müller et al. (1991).

keine Informationen zum Stichprobenplan enthielten, konnten die Nutzer die Varianzschätzungen nur unter der Annahme einer uneingeschränkten Zufallsauswahl durchführen. Da aber die Klumpung in der Regel mit einer Varianzvergrößerung verbunden ist, wird hierbei der Stichprobenfehler unterschätzt. Nur sehr eingeschränkt war mit Hilfe veröffentlichter Fehlerrechnungen des Statistischen Bundesamtes (1998a) eine Korrektur für Design-Effekte möglich.

Aus Sicht der Nutzer der Scientific Use Files war deshalb eine sachgerechte Lösung dieses Problems erforderlich. Um so wichtiger wurde dies, nachdem die Zahl der Mikrozensusnutzer in Folge eines vom Bundesministerium für Bildung und Forschung (BMBF) finanzierten Pilotprojekts (siehe Köhler et al. 2000) stark anstieg. Der im Vergleich zu früher vor allem in finanzieller Hinsicht wesentlich erleichterte Datenzugang hat maßgeblich dazu beigetragen, dass der Mikrozensus von der Forschung zunehmend für die Bearbeitung verschiedenster Fragestellungen herangezogen wird.²

Um den vielfältigen Forschungszwecken bei der Analyse von Scientific Use Files gerecht zu werden, reichen die nur für wenige Merkmale vom Statistischen Bundesamt veröffentlichten Design-Effekte nicht aus. Damit Varianzschätzungen auch für die von den Forschern selbst gebildeten Variablen durchgeführt werden können, sind Angaben zur Schichtung und Klumpung im Scientific Use File notwendig. Auf Nachfrage von Mikrozensusnutzern stellt nun das Statistische Bundesamt ab dem Mikrozensus 1996 Informationen über den Stichprobenplan in anonymisierter Form zur Verfügung. Damit ist in der Forschungspraxis erstmals eine adäquate Berechnung der Stichprobenfehler möglich geworden. In diesem Beitrag wird diskutiert, wie man selbst Varianzschätzungen unter Berücksichtigung des Stichprobendesigns mit den faktisch anonymisierten Mikrozensusdaten vornehmen kann und welche Unterschiede ggf. zu den Fehlerrechnungen der statistischen Ämter bestehen.

Dieser Beitrag konzentriert sich auf den Stichprobenfehler, der wie bei allen Stichproben für die Beurteilung der Qualität der Schätzungen von Bedeutung ist. Grundsätzlich sind aber auch systematische Fehler zu beachten, die bei der Berechnung des Stichprobenfehlers unberücksichtigt bleiben. Systematische Fehler entstehen durch Untererfassung, Befragungsausfälle, falsche Angaben und Kodierfehler etc. Systematische Fehlerquellen werden in diesem Beitrag im Zusammenhang mit dem

2 Siehe u.a. die auf zwei Nutzerkonferenzen zum Mikrozensus vorgestellten Arbeiten in Lüttinger (1999) sowie im World Wide Web (URL http://www.gesis.org/Dauerbeobachtung/Mikrodaten/Veranstaltungen/Veranstaltungen_index.htm).

Unit-Nonresponse und der Anpassung an die Bevölkerungsfortschreibung angesprochen.

Im Folgenden wird zunächst das Erhebungsdesign des Mikrozensus skizziert. Daran schließt sich die Darstellung der Ziehung des Scientific Use Files an. Danach wird gezeigt, wie die Varianz von Totals sowie von Verhältnis- und Mittelwerten bestimmt werden kann. Weiterhin wird die Anpassung der Mikrozensusergebnisse an die Bevölkerungsfortschreibung mit Hilfe des von Särndal et al. (1992) entwickelten Regressionsansatzes behandelt. Anschließend wird die bisher verwendete Methode der linearen Approximation der Design-Effekte untersucht, die auch für die Scientific Use Files vor dem Erhebungszeitpunkt 1996 noch relevant ist. In den abschließenden Bemerkungen fassen wir die wichtigsten Ergebnisse zusammen und nennen Möglichkeiten, wie die Datenbereitstellung verbessert werden kann.

2. Das Erhebungsdesign des Mikrozensus

Das Scientific Use File, das im Folgenden mit dem Kürzel FAMZ für faktisch anonymisierter Mikrozensus bezeichnet wird, ist eine 70-prozentige Substichprobe des Mikrozensus (MZ). Um die Varianz von Populationsschätzungen auf Basis des FAMZ bestimmen zu können, ist es deshalb notwendig, zunächst das Erhebungsdesign des MZ darzustellen (siehe Krug et al. 1999: 304ff.; Meyer 1994; Statistisches Bundesamt 1999).³ Im MZ wurde 1990 ein neuer Stichprobenplan eingeführt, der im Vergleich zum Stichprobenplan für den MZ 1972-1989 eine tiefere regionale Schichtung und eine Verringerung der Klumpengröße aufweist. Zunächst wurden Primäreinheiten (Primary Sampling Units = PSU's) gebildet, die als Auswahlbezirke bezeichnet werden. Sie bestehen durchschnittlich aus neun benachbarten Wohnungen, die in einer Gebäudegruppe oder innerhalb eines größeren Gebäudes liegen.

Die Bildung dieser PSU's erfolgte innerhalb von 4 Gebäudeklassen (sogenannte fachliche Schichtung) auf Basis von Ergebnissen der Volkszählung 1987 bzw. des Zentralen Einwohnerregisters in den neuen Bundesländern ab 1991. Der MZ 1996 enthält über 40.000 PSU's. Da alle Haushalte eines ausgewählten PSU befragt werden, ist der MZ eine Klumpenstichprobe (Cluster Sample). Tendenziell vergrößert

³ Siehe hierzu ergänzend auch die Darstellungen der statistischen Landesämter: Frank/Kafurke 1990; Reinders 1993; Schmidt 1990; Werner 1994.

die Klumpung die Varianz der Populationschätzer im Vergleich zu einer uneingeschränkten Zufallsauswahl.⁴

Das zweite Element des MZ ist die regionale Schichtung, die den umgekehrten Effekt hat, nämlich die Varianz zu verringern. Großstädte über 200 Tsd. Einwohner und sonstige Regionen über 250 Tsd. Einwohner bilden eigene Schichten. Innerhalb der Schichten werden darüber hinaus Schichtuntergruppen identifiziert, die aus Regionaleinheiten mit mindestens 100 Tsd. Einwohnern bestehen.

Vor der eigentlichen Auswahl wurden die PSU's innerhalb der fachlichen und regionalen Schichten nach den Merkmalen Bundesland, Regierungsbezirk, Kreis, Gemeindegroßenklasse und Gemeinde angeordnet. Innerhalb der Gemeinden wurden die PSU-Nummern, welche die regionale Anordnung innerhalb der Schichtuntergruppen widerspiegeln, zu einer Anordnung der PSU's der Auswahlgesamtheit verwendet. Jeweils 100 aufeinander folgende PSU's wurden zu einer Zone zusammengefasst.

Der Kern des Ziehungsverfahrens besteht in einer zufälligen Zuordnung von jeweils 100 PSU's einer Zone zu den Zahlen 0 bis 99. Für jede Zone wurde eine neue zufällige Permutation der Zahlen 0 bis 99 verwendet. Die PSU's zu einer Zahl zwischen 0 bis 99 bildeten dann jeweils eine MZ-Stichprobe.

Nach dem Stichprobendesign des MZ wird eine Zerlegung der 1-Prozent-Stichproben in 0,25-Prozent-Stichproben benötigt, da im Rahmen des Rotationsverfahrens jährlich ein Viertel des Bestandes ausgetauscht wird.⁵ Hierzu wurden jeweils vier aufeinanderfolgende Zonen einer Zufallszahl von 1 bis 4 zugeordnet. Dabei wurde für jeden Block von vier aufeinanderfolgenden Zonen eine Permutation der Zahlen 1 bis 4 verwendet. PSU's mit gleicher Nummer zählten zum gleichen Rotationsviertel, so dass nach diesem Verfahren die Auswahlgesamtheit der PSU's in Rotationsviertel zerlegt wurde. Aus dieser Auswahlgesamtheit von 100 1-Prozent-Stichproben der PSU's wurden 20 Vorratsstichproben über eine Ziehung aus der Urne ausgewählt.

Interpretiert man dieses Verfahren innerhalb des Schemas der klassischen Stichprobentheorie, so entsprechen die Zonen Schichten. Für die Berechnung der Gesamt-

4 Der Stichprobenfehler ist in der Regel um so größer, je homogener die Klumpen hinsichtlich der interessierenden Merkmale, je größer die Klumpen, und je unterschiedlicher die Klumpengrößen sind.

5 Der MZ ist als rotierende Panelstichprobe angelegt, bei der ein Auswahlbezirk bzw. die darin wohnenden Haushalte vier Jahre lang befragt werden, wobei jährlich ein Viertel der PSU's ausgetauscht wird. Wegziehende Personen bzw. Haushalte werden allerdings nicht weiterbefragt, sondern durch die nachziehenden Haushalte ersetzt (Prinzip der Flächenstichprobe).

varianz würde die Varianz der PSU-Totals innerhalb einer Zone benötigt. Da aber aus jeder Zone jeweils nur eine Primäreinheit gezogen wurde, kann die Varianz der PSU-Totals innerhalb einer Zone nicht berechnet werden. Dieser Ansatz führt deshalb nicht weiter. Eine Möglichkeit mit dieser Schwierigkeit umzugehen, besteht darin, die Schichten größer zu definieren und die Zonenstruktur des Ziehungsverfahrens zu ignorieren; das heißt, für diese größeren Schichten wird eine uneingeschränkte Zufallsauswahl angenommen. Diese Strategie wird auch in den Varianzberechnungen des Statistischen Bundesamtes angewendet. Innerhalb jeder Gebäudeklasse bilden die insgesamt über 200 Regionalschichten jeweils eine Schicht bei der Varianzberechnung. Die Annahme größerer Schichten und das Ignorieren der kleinräumigen Struktur bei dem Ziehungsverfahren führt in der Tendenz zu einer Überschätzung der Varianz. Vor dem Hintergrund, dass auf Grund fehlender Detailangaben zum Stichprobendesign im FAMZ die Varianz vermutlich nur annähernd berechnet werden kann, ist es vielleicht überraschend, dass schon im Fall des MZ - so, wie er in den statistischen Ämtern vorliegt - die Stichprobenvarianz ebenfalls nur näherungsweise bestimmt werden kann.

Die Benutzung eines vereinfachten Berechnungsverfahrens wird auch durch einen zweiten Sachverhalt nahegelegt. In den Jahren nach der Ziehung der MZ-Stichproben sind Neubauten entstanden bzw. ganze Flächen neu bebaut worden. Die Neubauten werden bei der Aktualisierung der Stichproben in einer zusätzlichen „Neubauschicht“ berücksichtigt und unterliegen einem gesonderten Ziehungsverfahren.

Bei der Aktualisierung der Stichprobe werden zunächst auf Kreis- bzw. Gemeindeebene Primäreinheiten auf Basis der Bautätigkeitsstatistik gebildet. Diese PSU's werden nach Gebäudegrößenklassen geschichtet und innerhalb jeder Schicht nach der Reihenfolge ihrer Bildung durchnummeriert. Im Unterschied zu dem oben beschriebenen Verfahren erfolgt die Ziehung der Neubau-PSU's aber durch systematisches Ziehen mit festem Intervall bei zufälligem Startpunkt.

Für das systematische Ziehungsverfahren ist bekannt, dass keine erwartungstreuen Varianzschätzer existieren (vgl. Wolter 1985: 248ff.). Man muss sich auch in diesem Fall mit einer Vereinfachung begnügen. Erfahrungen aus Simulationsstudien zeigen, dass die Approximation über die Varianz einer in den Schichten uneingeschränkten Zufallsstichprobe in der Regel zu einer Überschätzung der Varianz führt (vgl. Wolter 1985: 282).

Da alle Stichproben nach dem selben Verfahren ermittelt wurden, ergibt sich die Möglichkeit, die Variation der Schätzergebnisse über die Rotationsviertel für die Varianzschätzung zu benutzen. Allerdings ist im FAMZ die Zugehörigkeit der Haus-

halte zu den Rotationsgruppen bisher nicht dokumentiert. Bei Kenntnis der Rotationsgruppenzugehörigkeit könnte man die Varianz beispielsweise über das Jackknife-Verfahren analysieren (vgl. Wolter 1985: 153ff) und hierbei Effekte der Schichtung durch Anordnung sowie des systematischen Ziehens bei der Neubauschicht berücksichtigen.

3. Die Ziehung des Scientific Use Files aus dem MZ 1996

Aus Datenschutzgründen konnte die Regionalschicht als regionales Schichtungsmerkmal nicht in den FAMZ aufgenommen werden.⁶ Darüber hinaus wurde im Vergleich zum MZ ein anderes Ziehungsverfahren verwendet, das aus folgenden Schritten besteht.

Zunächst wurden die Haushalte des MZ nach Bundesland, Regierungsbezirk, Gemeindegrößenklasse, Zahl der Personen im Privathaushalt, Auswahlbezirksnummer und Haushaltsnummer angeordnet. Diese Anordnung unterscheidet sich von der Schichtung im MZ hauptsächlich hinsichtlich der Zahl der Personen im Privathaushalt in der Sortierfolge. In dieser Anordnung wurden die Haushalte neu durchnummeriert. In die Auswahl von 70 Prozent aller Haushalte wurden alle Haushalte – einschließlich aller Personen in diesen Haushalten – übernommen, deren letzte Platzziffer der Haushaltsnummer von 2, 5 und 9 verschieden war. Im Anschluss an die Stichprobenziehung wurden die Haushalte umsortiert und erhielten eine neue, fortlaufende Nummerierung, so dass die alten Haushaltsendziffern nicht mehr rekonstruierbar sind.

Dieses vom MZ abweichende Verfahren erschwert die Berechnung der Stichprobenvarianz. Insbesondere ist die fehlende Kenntnis über die Anzahl der Haushalte pro PSU im MZ beziehungsweise der Verlust der Information über die letzte Platzziffer beim Auswahlverfahren insofern bedauerlich, als man aus der Varianz der Schätzergebnisse über die sieben verbliebenen Platzziffern Rückschlüsse auf die zusätzliche Varianzkomponente gewinnen könnte, die durch die 70-Prozent-Auswahl verursacht wird.

⁶ Zur Zeit stehen der Forschung Scientific Use Files des Mikrozensus nur als sogenanntes Grundfile zur Verfügung, das durch tiefe fachliche Gliederungen und sehr grobe regionale Gliederungen charakterisiert ist. Die Bereitstellung eines faktisch anonymisierten Mikrozensus-Regionalfile durch die statistischen Ämter ist in Vorbereitung.

4. Die Berechnung der Varianz im Scientific Use File

Obwohl das Scientific Use File somit im strengen Sinne über ein 2-phasiges Ziehungsverfahren auf Haushaltsebene gewonnen wurde, ist das Verfahren in guter Näherung durch ein zweistufiges Auswahlverfahren beschreibbar.⁷ Hierbei entspricht die erste Stufe der geschichteten Auswahl der PSU's im Mikrozensus. Die zweite Stufe kann als eine einfache 70-Prozent-Auswahl von Haushalten aus einer PSU betrachtet werden.

Wie im letzten Abschnitt festgehalten, ist die Annahme fester Stichprobenumfänge für die zweite Auswahlstufe durch das systematische Ziehungsverfahren nicht immer gegeben und bildet eine Quelle für die Variation der Schätzergebnisse, die mangels entsprechender Information ignoriert werden muss. Weiterhin sind aufgrund des systematischen Ziehungsverfahrens die Ziehungen der Haushalte zwischen den einzelnen PSU's nicht voneinander unabhängig. Es kann jedoch vermutet werden, dass die aus beiden Sachverhalten resultierende Unterschätzung der Varianz eine untergeordnete Bedeutung hat (vgl. hierzu auch Särndal et al. 1992: 349ff).

Zusammenfassend ist festzuhalten, dass der FAMZ nicht alle Schichtungsmerkmale der ersten Auswahlstufe enthält, sondern nur die Merkmale Bundesland, Gemeindegroßenklasse sowie die Gebäudeschicht. Grundsätzlich führt die Nichtberücksichtigung von Schichtungsmerkmalen zu einer Überschätzung der Varianz, so dass man bei der Bildung von Konfidenzintervallen konservativ bleibt.

Die im Folgenden benutzte Notation lehnt sich an das Lehrbuch von Särndal et al. (1992) an. Es bezeichne $h \in \{1, \dots, H\}$ den Schichtindex. Die PSU's werden mit $i \in \{1, \dots, N_h\}$ indiziert, wobei N_h die Anzahl der PSU's in der h -ten Schicht ist. Haushalte werden mit $k \in \{1, \dots, N_i\}$ indiziert, wobei N_i die Anzahl der Haushalte im i -ten PSU ist. Der Merkmalswert $y_{h,i,k}$ bezieht sich somit auf Haushalt k in PSU i in Schicht h ; y_k verweist auf den Merkmalswert von Haushalt k , unabhängig von dessen PSU-Nummer und der Schichtzugehörigkeit.

⁷ Mehrphasige unterscheiden sich von mehrstufigen Stichproben darin, dass bei Phasenauswahlen die Auswahlseinheiten (z.B. 1. Phase: Auswahl der MZ-Haushalte, 2. Phase: 70-Prozent-Auswahl der FAMZ-Haushalte) gleich sind, während mehrstufigen Auswahlen unterschiedliche Auswahlseinheiten zu Grunde liegen (z.B. 1. Stufe: PSU's, 2. Stufe: Haushalte). Siehe zum Vergleich von Varianzschätzungen bei 2-phasigen und 2-stufigen Stichproben Särndal et al. 1992: 349-350).

Für ein 2-stufiges geschichtetes Ziehungsverfahren müssen unterschiedliche *Grundgesamtheiten* berücksichtigt werden. Es sei U_I die Menge aller Primäreinheiten im Erhebungsgebiet (Anzahl = N_I). $U_{I,h}$ sei die Menge der Primäreinheiten in Schicht h (Anzahl = $N_{I,h}$). Die Menge aller Haushalte im i -ten PSU sei U_i (Anzahl = N_i). Schließlich bezeichne U die Menge aller Haushalte im Erhebungsgebiet (Anzahl = N).

Den Grundgesamtheiten entsprechen die unterschiedlichen *Stichproben*: s_I ist die Stichprobe der PSU's vom Umfang n_I . Diese Stichprobe verteilt sich auf die H Schichten als $s_{I,h}$ mit dem Umfang $n_{I,h}$.

Die Stichprobe der Haushalte aus PSU i ist s_i und besitzt den Umfang n_i . Schließlich bezeichnet s die Stichprobe aller Haushalte mit dem Umfang n .

Weiterhin bezeichnet $\pi_{I,i}$ die *Ziehungswahrscheinlichkeit* von PSU i und $\pi_{k|i}$ die bedingte *Ziehungswahrscheinlichkeit* von Haushalt k aus PSU i , wenn PSU i gezogen wurde. Im Falle des MZ gilt $\pi_{I,i} = 0,01$. Für den FAMZ nehmen wir $\pi_{k|i} = 0,7$ an.

5. Schätzung von Gesamtwerten qualitativer Merkmale (Totals)

Bevor wir zeigen, wie sich diese Stichprobeninformationen für die Varianzschätzung verwenden lassen, soll ein einführendes Beispiel aus der Situation vor der Weitergabe von Angaben zur Schichtung und Klumpung im FAMZ die Ausgangsproblematik und den möglichen Informationsgewinn verdeutlichen.

Nehmen wir an, ein Forscher sei im Rahmen einer Analyse von Lebenslagen allein lebender Personen an der Zahl weiblicher 1-Personenhaushalte in der Grundgesamtheit interessiert. Der Einfachheit halber wird bei der Hochrechnung die Anpassung der Ergebnisse an die Bevölkerungsfortschreibung nicht berücksichtigt (siehe Abschnitt 8). Die Auswertung des FAMZ 1996 ergibt, dass unter den insgesamt 229.221 Privathaushalten 49.071 weibliche 1-Personenhaushalte sind. Nach Design-Gewichtung der Fallzahlen mit dem Kehrwert der *Ziehungswahrscheinlichkeiten* ($1/(0,01 \cdot 0,7)$) ergibt sich hochgerechnet ein Gesamtwert (Total) von 7.010,1 Tsd. Haushalten. Um die stichprobenbedingte Unsicherheit dieser Schätzung zu berücksichtigen, musste ein Nutzer bislang von einer uneingeschränkten Zufallsauswahl ausgehen. Unter Annahme einer Binomialverteilung berechnet man den relativen Standardfehler und erhält einen Wert von 0,4 Prozent (siehe Statistisches Bundesamt 1998a: 17; siehe auch Gleichung (28) im Abschnitt 9). Um die Schichtung und

Klumpung des MZ näherungsweise zu berücksichtigen, ist dieser Wert mit dem Design-Effekt zu multiplizieren. Verwendet man die Näherungsfunktion des Statistischen Bundesamtes (1998a: 17), beträgt der Design-Effekt für 1-Personenhaushalte 1,36. Für das interessierende Merkmal erhält man schließlich den Wert 0,54 ($=0,4 \cdot 1,36$) als Schätzung des relativen Standardfehlers. Ein 95-Prozent-Konfidenzintervall für den geschätzten Populationswert reicht somit von circa 6.936 bis 7.084 Tsd. Haushalten. Würde man, wie vom Statistischen Bundesamt empfohlen, den für den MZ berechneten design-basierten Standardfehler an den geringeren Stichprobenumfang des FAMZ anpassen und entsprechend mit 1,2 multiplizieren, verbreiterte sich das Konfidenzintervall auf circa 6.921 bis 7.099 Tsd. Haushalte.

Wie können nun die ab dem FAMZ 1996 vorliegenden Informationen zum Stichprobendesign für die direkte Varianzschätzung verwendet werden? Der Gesamtwert (Total) eines Merkmals y ist gegeben durch:

$$(1) \quad t = \sum_{k \in U} y_k = \sum_{i \in U_i} t_i \quad \text{wobei } t_i = \sum_{k \in U_i} y_k$$

$$(2) \quad = \sum_{h=1}^H t_h \quad \text{wobei } t_h = \sum_{i \in U_{i,h}} t_i$$

Hierbei ist (1) die Darstellung über die Totals der Primäreinheiten und (2) die Darstellung über die Schichttotals.

Der π -Schätzer von t basiert auf dem Kehrwert der Ziehungswahrscheinlichkeiten. Man erhält für das PSU-Total die Schätzung:

$$(3) \quad \hat{t}_i = \sum_{k \in s_i} \frac{y_k}{\pi_{k|i}} = \frac{1}{0,7} \sum_{k \in s_i} y_k$$

Das Schichttotal t_h lässt sich schätzen durch:

$$(4) \quad \hat{t}_h = \sum_{i \in s_{i,h}} \frac{\hat{t}_i}{\pi_{I,i}} = 100 \sum_{i \in s_{i,h}} \hat{t}_i$$

Man erhält damit als Schätzung für das Gesamt-Total:

$$(5) \quad \hat{t} = \sum_{h=1}^H \hat{t}_h = \frac{100}{0,7} \sum_{k \in s} y_k$$

Wir wollen im folgenden $V(\hat{t})$, die Varianz von \hat{t} , bestimmen. Bei Unabhängigkeit der Ziehung zwischen den Schichten erhält man:

$$(6) \quad V(\hat{t}) = \sum_{h=1}^H V_h(\hat{t}_h)$$

$$(7) \quad \hat{V}(\hat{t}) = \sum_{h=1}^H \hat{V}_h(\hat{t}_h)$$

Das heißt, es genügt, die Varianz des geschätzten Totals \hat{t}_h für die einzelnen Schichten zu bestimmen. Bei einem 2-stufigen Ziehungsverfahren, mit dem auf jeder Stufe eine einfache (simple: SI) Stichprobe gezogen wird (Typ: (SI, SI) in der Notation von Särndal et al. 1992: 142), erhält man:

$$(8) \quad V_{SI,SI}(\hat{t}_h) = N_{I,h}^2 \frac{1-f_h}{n_{I,h}} S_{U_{I,h}}^2 + \frac{N_{I,h}}{n_{I,h}} \sum_{i \in U_{I,h}} N_i^2 \frac{1-f_i}{n_i} S_{U_i}^2$$

wobei $f_h = \frac{n_{I,h}}{N_{I,h}} = 0,01$, $f_i = \frac{n_i}{N_i} = 0,7$ und:

$$(9) \quad S_{U_{I,h}}^2 = \frac{1}{N_{I,h} - 1} \sum_{i \in U_{I,h}} (t_i - \bar{t}_{U_{I,h}})^2$$

= Varianz der PSU - Werte in der Schicht h
= "Varianz between PSU"

$$(10) \quad S_{U_i}^2 = \frac{1}{N_i - 1} \sum_{k \in U_i} (y_k - \bar{y}_{U_i})^2$$

= Varianz der y - Werte im i -ten PSU
= "Varianz within PSU"

Für die gegebenen Werte vereinfacht sich $V_{SI,SI}$ zu:

$$(11) \quad V_{SI,SI} = 100^2 \times 0,99 \times n_{I,h} \times \text{"Varianz between PSU"} \\ + 100 \sum_{i \in U_{I,h}} \frac{0,3}{0,7^2} n_i \times \text{"Varianz within PSU"}$$

Indem in Gleichung (8) bzw. (11) die Between- und Within-Varianzen durch ihre Stichprobenpendants ersetzt werden, ergibt sich die Varianzschätzung $\hat{V}_{SI,SI}$:

$$(12) \quad S_{S_{I,h}}^2 = \frac{1}{n_{I,h} - 1} \sum_{i \in S_{I,h}} \left(\hat{t}_i - \hat{\bar{t}}_{S_{I,h}} \right)^2 \\ = \text{Varianz der PSU - Werte in Schicht } h$$

$$(13) \quad S_{S_i}^2 = \frac{1}{n_i - 1} \sum_{k \in S_i} \left(y_k - \bar{y}_{S_i} \right)^2 \\ = \text{Varianz der } y \text{ - Werte im } i \text{ - ten PSU}$$

Man beachte allerdings, dass die beiden Varianzterme in Gleichung (11) durch die Stichprobenpendants nicht erwartungstreu geschätzt werden (vgl. Särndal et al. 1992: 142). Die Between-Komponente wird im Mittel durch die Varianz der geschätzten PSU-Totals überschätzt. Lediglich die Summe der beiden Stichproben-terme ist eine erwartungstreue Schätzung für $V_{SI,SI}$.

Die Berechnung des zweiten Teils von $\hat{V}_{SI,SI}$ verlangt die Ermittlung der Standardabweichung der y-Werte innerhalb von circa 40.000 PSU's. Dies ist zwar rechenaufwendig, jedoch kein prinzipielles Problem. Särndal et al. (1992: 139ff.) schlagen als Näherung für $\hat{V}_{SI,SI}$ die Berechnung lediglich des ersten Teils über die Between-Varianzen vor. Dies ist gerechtfertigt, wenn die Auswahlwahrscheinlichkeit der Primäreinheiten klein ist. Wie die numerischen Ergebnisse in Tabelle 1 belegen, ist im FAMZ die Within-Varianz eine zu vernachlässigende Größe im Vergleich zur Between-Varianz. Für die Schätzung der Varianz können deshalb Standardprozeduren verwendet werden, die mittlerweile in Statistikpaketen wie STATA und SAS implementiert sind. In diesen Prozeduren wird lediglich die erste Komponente der Between-PSU Varianz innerhalb der Schichten berechnet.

Die Tabelle 1 zeigt für einige ausgewählte Merkmale die Schätzung von Totals auf Basis des FAMZ und den Vergleich der Schätzgenauigkeit mit den Angaben für den MZ (Statistisches Bundesamt 1998b). Weiterhin werden die relativen Standardfehler miteinander verglichen sowie die in Abschnitt 9 behandelte Näherungsfunktion auf Basis der Design-Effekte. Für das eingangs geschilderte Beispiel weiblicher 1-Personenhaushalte finden wir bei Anwendung des direkten Schätzverfahrens mit 0,47 Prozent im Vergleich zu 0,54 Prozent bei der bisher notwendigen Vorgehensweise einen deutlich geringeren relativen Standardfehler. Damit erhalten wir mit einem 95-Prozent-Konfidenzintervall von circa 6.946 bis 7.075 Tsd. Haushalten eine präzisere Schätzung.

Betrachten wir die geschätzten Totals in Tabelle 1, fällt auf, dass die Werte für den FAMZ stets kleiner sind als für den MZ. Bei den meisten Merkmalen liegen die MZ-Werte weit außerhalb des konventionellen 95-Prozent-Konfidenzintervalls der FAMZ-Werte. Diese systematische Abweichung ist aber kein Hinweis auf eine verzerrte Substichprobenziehung, sondern darauf zurückzuführen, dass das Statistische Bundesamt für seine Berechnungen eine Kompensationsgewichtung für die circa 2,5 Prozent Nonresponse-Fälle auf Haushaltsebene benutzt. Die für die einzelnen Haushalte unterschiedlichen Gewichte liegen jedoch im FAMZ nicht vor und können folglich auch nicht für die Schätzung von Totals benutzt werden.⁸

Der Vergleich der Standardfehler zeigt, dass die FAMZ-Werte stets über den MZ-Werten liegen. Dies war auch angesichts des um den Faktor 0,7 niedrigeren Stichprobenumfangs zu erwarten. Die Reduktion des Stichprobenumfangs würde bei einer uneingeschränkten Zufallsauswahl zur Vergrößerung des Standardfehlers um rund 20 Prozent führen. Empirisch ergibt sich jedoch nur eine Vergrößerung um durchschnittlich 9 Prozent.⁹ Der Grund für die Halbierung des Fallzahleffekts liegt in der Verringerung des Klumpeneffekts, der sich aus der Auswahl von Haushalten unter approximativer Erhaltung der Schichtung bei der Substichprobenziehung ergibt.

⁸ Bezogen auf das Standardauswertungsprogramm von über 450 Merkmalen betragen die FAMZ-Totals durchschnittlich 97,7 Prozent der entsprechenden MZ-Totals.

⁹ Das Statistische Bundesamt hat auf einem Workshop zur Nutzung des FAMZ 96 für die Berechnung des Stichprobenfehlers empfohlen, den Design-Effekt mit dem Faktor 1,2 zu multiplizieren. Statt des Faktors $(1/0,7)^{1/2} = 1,195$ beträgt der Faktor für den FAMZ bezogen auf die über 450 Merkmale des Standardprogramms durchschnittlich 1,09.

Tabelle 1: Schätzung von Gesamtwerten (Totals) für ausgewählte Merkmale

Merkmal	Total (in 1000)		Standardfehler (in 1000)			relativer Standardfehler (in Prozent)			Design- Effekt		
	FAMZ	MZ	FAMZ			MZ	Näherung			FAMZ	MZ
			Insges.	Betw.	Within		FAMZ	MZ	Näherung		
Ausländische Erwerbspersonen ¹⁾	2.283,7	2.330,6	29,4	29,4	1,2	27,3	1,29	1,17	0,92	1,66	1,83
Sofort verfügbare Erwerbslose ¹⁾	2.976,7	3.034,2	24,1	24,0	1,2	21,4	0,81	0,71	0,81	1,20	1,26
Überwiegender Lebensunt. durch Erwerbstätigkeit ¹⁾	29.607,9	30.285	81,4	81,4	3,2	77,7	0,28	0,26	0,30	1,63	1,85
Weibliche Erwerbst., monatl. Nettoeink. <600 DM ¹⁾	1.545,4	1.585,4	15,9	15,8	0,8	13,9	1,03	0,88	1,11	1,08	1,12
Privathaushalte mit 1 Person, weiblich ²⁾	7.010,1	7.259,6	33,1	33,1	1,5	32,0	0,47	0,44	0,54	1,19	1,35

Quelle: FAMZ: Mikrozensus 1996 (faktisch anonymisierte 70%-Substichprobe); MZ: Frei hochgerechnete Werte nach Kompensation der bekannten Ausfälle (Fehlerrechnungen zum Mikrozensus 1996; Statistisches Bundesamt 1998b).

Näherung: Aus den Besetzungszahlen der Tabellenfelder geschätzter relativer Standardfehler unter Verwendung der veröffentlichten Zuschlagsfaktoren zum Mikrozensus 1990 (Statistisches Bundesamt 1998a: 17); Näherungsfunktion; s. Abschnitt 9.

Subpopulationen: 1) Bevölkerung am Hauptwohnsitz; 2) Privathaushalte.

Die Betrachtung des relativen Standardfehlers (=Standardfehler/Total) belegt die hohe Präzision von Ergebnissen auf Basis des FAMZ, die ihresgleichen bei den üblichen sozialwissenschaftlichen Erhebungen sucht. Selbst bei relativ schwach besetzten Merkmalen, wie zum Beispiel erwerbstätigen Frauen mit einem Nettoeinkommen unter 600 DM, liegt der relative Standardfehler bei 1 Prozent.¹⁰ Wie aber der Vergleich der FAMZ-Totals mit den durch das Kompensationsgewicht hochgerechneten MZ-Werten gezeigt hat, liegen bereits bei kleinen Stichprobenausfällen die MZ-Schätzwerte weit außerhalb des Konfidenzintervalls, das man für den

¹⁰ Nicht zu verwechseln mit einem Prozentpunkt.

FAMZ erhält. Dieser Sachverhalt erklärt sich aus dem Fehlen von Nonresponse-Informationen im FAMZ, das einen Bias in zweifacher Hinsicht bewirkt.

Zum einen werden die Populationstotals systematisch unterschätzt. Diese Unterschätzung ließe sich beheben, wenn die Kompensationsgewichtung im FAMZ erhalten wäre.¹¹ Die Kompensationsgewichte geben die geschätzten Ausfallraten innerhalb der 401 regionalen Untergruppen für 19 verschiedene Merkmalskombinationen an.¹² Die Kompensationsgewichte könnten für eine Korrektur der Inklusionswahrscheinlichkeiten $\pi_{k|i}$ auf der zweiten Auswahlstufe verwendet werden. Ersatzweise kann man die durchschnittliche Ausfallrate von rund 2,5 Prozent verwenden. Multipliziert man die FAMZ-Schätzwerte in Tabelle 1 mit dem Faktor 1,025 so überdecken alle Konfidenzintervalle die entsprechenden Werte aus dem MZ.

Zum anderen bewirkt das Ausblenden des Nonresponse jedoch auch eine gewisse Unterschätzung des Stichprobenfehlers, da die Streuung der Stichprobenvarianz aufgrund des Nonresponse ignoriert wird. Hinsichtlich der geringen Ausfallrate dürfte dieser Effekt aber zu vernachlässigen sein.

Auch wenn die genannten Probleme die Varianzschätzung beeinträchtigen, kann zusammenfassend festgehalten werden, dass die direkte Varianzberechnung im Vergleich zur bisher notwendigen Vorgehensweise für die Nutzer des FAMZ erhebliche Vorteile einer präziseren Varianzschätzung unter Berücksichtigung des Stichprobendesigns besitzt.

6. Schätzung eines Verhältniswerts

Für die im Mikrozensus überwiegenden qualitativen Merkmale werden neben Gesamtwerten häufig auch Verhältnis- oder Anteilswerte betrachtet. Da bei dieser Verhältnissschätzung sowohl für das Zähler- als auch das Nennermerkmal Stichprobenfehler zu berücksichtigen sind, ist die Berechnung nicht ganz so einfach wie bei Totals.

¹¹ Das Grundfile enthält als Gewichtungsvariable lediglich die Gewichtung nach Anpassung an die Bevölkerungsfortschreibung, die sowohl die unterschiedlichen Ausfallraten als auch die Soll-Ist Anpassung nach der Bevölkerungsfortschreibung berücksichtigt; vgl. hierzu nachfolgenden Abschnitt 8.

¹² Auf Haushaltsebene (d.h. für Haushaltsbezugspersonen) werden die Merkmalsgruppen über die Haushaltsgröße (1,2,3≤), die Nationalität (Deutsche/Ausländer), den Wohnsitz (Haupt-/Nebenwohnsitz; nur für Deutsche) sowie für 1-Personenhaushalte das Geschlecht und das Alter (≤60, >60) gebildet.

Es sei t_y ein Total bezüglich eines Merkmals y und t_z ein Total bezüglich eines Merkmals z . Die zu schätzende Größe wird durch $\hat{R} = \hat{t}_y / \hat{t}_z$ dargestellt. Zur Berechnung der Varianz von \hat{R} wird eine Taylorentwicklung der Funktion $f(t_y, t_z) = t_y / t_z$ benutzt:

$$\begin{aligned}
 f(\hat{t}_y, \hat{t}_z) &\approx f(t_y, t_z) + \frac{\partial f}{\partial t_y} \Big|_{t_y} (\hat{t}_y - t_y) + \frac{\partial f}{\partial t_z} \Big|_{t_z} (\hat{t}_z - t_z) \\
 &= f(t_y, t_z) + \frac{1}{t_z} (\hat{t}_y - t_y) - \frac{t_y}{t_z^2} (\hat{t}_z - t_z) \\
 &= \text{Konstante} + \frac{1}{t_z} (\hat{t}_y - R \hat{t}_z) \\
 (14) \quad &= \text{Konstante} + \frac{1}{t_z} \sum_{k \in s} \frac{u_k}{\pi_k} \quad \text{wobei } u_k = (y_k - R z_k)
 \end{aligned}$$

Damit ist bis auf die Konstante $1/t_z$ lediglich die Varianz des π -Schätzers für das U-Total zu bestimmen. Wegen der linearen Approximation der Schätzfunktion wird allerdings nur eine Näherung der Varianz von \hat{R} bestimmt, die für große Beobachtungsumfänge mit $V(\hat{R})$ übereinstimmt (Bezeichnung: Asymptotische Varianz $AV(\hat{R})$):

$$(15) \quad AV(\hat{R}) = \frac{1}{t_z^2} V(\hat{t}_u)$$

Zur Berechnung von $V(\hat{t}_u)$ kann man wieder Gleichung (8)-(11) verwenden. Man hat dabei lediglich y_k durch u_k zu ersetzen. $S_{U_i}^2$ misst die Varianz der Hilfsgrößen u_k innerhalb eines PSU. Die Hilfsgröße u_k kann als Abweichung der y_k/z_k Werte von dem Populationswert R interpretiert werden.

Eine asymptotisch erwartungstreue Schätzung von $AV(\hat{R})$ erhält man, indem die unbekanntenen Populationswerte t_z und R durch die Schätzwerte \hat{t}_z und \hat{R} ersetzt und die Gleichungen (12) und (13) benutzt werden. Hierbei ist \hat{t}_i gegeben durch:

$$(16) \quad \hat{t}_i = \sum_{k \in s_i} \frac{u_k}{\pi_{k|i}} = \frac{1}{0,7} \sum_{k \in s_i} (y_k - \hat{R}z_k)$$

Der Mittelwert der unter (16) geschätzten PSU Werte über alle PSU's der Schicht h entspricht dem Ausdruck $\hat{t}_{s_i, h}$ in Gleichung (12).

Die Schätzung der Varianz von \hat{R} bedarf damit keiner gesonderten Programmierung. Man hat lediglich statt des Merkmals y_k das Merkmal $u_k = y_k - \hat{R}z_k$ zu verwenden und darf am Schluss die Division durch \hat{t}_z^2 nicht vergessen.

7. Die Varianz von Populationsmitteln

Um auch für quantitative Variablen, wie beispielsweise Alter oder Zahl der Kinder im Haushalt etc., die Varianz von Populationsmitteln zu bestimmen, kann wie folgt vorgegangen werden. Der Populationsmittelwert zu einem Merkmal y ist gegeben durch:

$$\bar{y}_U = \frac{1}{N} \sum_{k \in U} Y_k = \frac{t_y}{N}$$

Falls der Populationswert N bekannt ist, lässt sich \bar{y}_U schätzen durch:

$$\hat{y}_{U, \pi} = \frac{\hat{t}_y}{N}$$

Eine geringere Varianz hat jedoch häufig das gewichtete Stichprobenmittel \hat{y}_s (vgl. Särndal et al. 1992: 182):

$$\hat{y}_s = \frac{\sum_{k \in s} Y_k / \pi_k}{\sum_{k \in s} 1 / \pi_k}$$

Dieser Schätzer wird von den meisten Programmpaketen benutzt, sobald eine GewichtungsvARIABLE verwendet wird. In dieser Form ist \hat{y}_s ein Spezialfall von \hat{R} mit $z_k = 1$, so dass die Varianzschätzung für Verhältniswerte \hat{R} benutzt werden kann.

Man erhält mit $\hat{N} = \sum_{k \in s} 1 / \pi_k$ für das Hilfsmerkmal u_k den folgenden Wert:

$$(17) \quad u_k = y_k - \frac{\hat{t}_y}{\hat{N}} \cdot 1 = y_k - \hat{y}_s$$

Das Hilfsmerkmal u_k misst also die Abweichung der y -Werte vom gewichteten Stichprobenmittel. Der Faktor $1/\hat{t}_z^2$ ist durch $1/\hat{N}^2$ gegeben.

Die Tabelle 2 zeigt für die Merkmale von Tabelle 1 den geschätzten Anteil (in Prozent) der Merkmalsträger an der Bevölkerung am Hauptwohnsitz beziehungsweise den Privathaushalten insgesamt. Die letzte Spalte in Tabelle 2 zeigt den Wert für den Standardfehler, wenn man die Insgesamt-Werte als bekannte Größe benutzt, indem man beispielsweise die im letzten Abschnitt geschätzten Standardfehler der Totals auf Anteilswerte umrechnet. Man erhält das auf den ersten Blick verblüffende Ergebnis, dass diese Schätzung in allen Fällen ungenauer ist als die Schätzung mit variablem Nenner. Dieser Effekt ist um so größer, je häufiger das Merkmal in der Population vertreten ist, das heißt je stärker Zähler- und Nennermerkmal kovariieren. Beispielsweise wird der Anteil der Personen mit überwiegendem Lebensunterhalt aus Erwerbstätigkeit um rund 30 Prozent genauer geschätzt, wenn der Nenner ebenfalls geschätzt wird. Für das Beispielmerkmal weibliche 1-Personenhaushalte wirkt sich dieser Effekt praktisch nicht aus. Sowohl im Fall eines geschätzten wie auch bekannten Nennermerkmals beträgt das 95-Prozent-Konfidenzintervall 21,2 bis 21,6 Prozent. Man beachte auch hier die extreme Genauigkeit; selbst bei relativ seltenen Merkmalen.

Tabelle 2: Schätzung von Anteilswerten für ausgewählte Merkmale

Merkmal Y	Anteil R (in %)	Standardfehler (in %)	
		N gesch.	N bek.
Ausländische Erwerbspersonen ¹⁾	3,2	0,040	0,041
Sofort verfügbare Erwerbslose ¹⁾	4,1	0,033	0,034
Überwiegender Lebensunterhalt durch Erwerbstätigkeit ¹⁾	41,3	0,081	0,114
Weibliche Erwerbstätige mit einem monatlichen Nettoeinkommen unter 600 DM ¹⁾	2,2	0,021	0,022
Privathaushalte mit 1 Person, weiblich ²⁾	21,4	0,093	0,101

Quelle: Mikrozensus 1996 (faktisch anonymisierte 70%-Substichprobe).

N. gesch./bek.: Standardfehler bei Verwendung der Fallzahl im Nenner (Z) als zu schätzende bzw. bekannte Größe.

Z-Merkmal: 1) Bevölkerung am Hauptwohnsitz; 2) Privathaushalte.

8. Die Varianz von Populationsschätzern nach der Anpassung an die Bevölkerungsfortschreibung

Wir sind bisher bei der Hochrechnung ausschließlich von den Stichprobenwerten ausgegangen. Von den statistischen Ämtern werden jedoch die Mikrozensusergebnisse anhand von Eckzahlen der laufenden Bevölkerungsfortschreibung hochgerechnet. Mit der Bereitstellung dieser Hochrechnungsfaktoren werden zwei Ziele verfolgt: Zum einen soll die Varianz der Populationsschätzungen verringert werden. Zum anderen soll ein möglicher Bias aufgrund der Stichprobenausfälle von rund 2,5 Prozent kompensiert werden. Zu diesem Zweck wird ein Hochrechnungsfaktor auf der Basis von regionalen Eckzahlen zu sechs Anpassungsklassen¹³ gebildet, der im Wesentlichen das Verhältnis der Sollzahlen aus der Bevölkerungsfortschreibung zu den Ist-Zahlen der im Mikrozensus Befragten abbildet (vgl. Heidenreich 1994). Die Nutzer der faktisch anonymisierten Mikrozensusdaten können für diese sogenannte gebundene Hochrechnung auf die im Datensatz enthaltenen Hochrechnungsfaktoren für Personen und Haushalte zurückgreifen. Bei deren Verwendung sollten sich die FAMZ-Ergebnisse nicht wesentlich von den veröffentlichten Gesamtwerten der amtlichen Statistik unterscheiden. Im Gegensatz zu der Veröffentlichung angepasster MZ-Ergebnisse basieren die Fehlerrechnungen des Statistischen Bundesamtes nur auf den Stichprobenwerten, korrigiert um den Unit-Nonresponse auf Haushaltsebene. Die Anpassung der Fallzahlen an die Bevölkerungsfortschreibung bleibt bei der Berechnung des Stichprobenfehlers außer Acht. Es ist deshalb zu fragen, in welcher Weise die Anpassung an die Bevölkerungsfortschreibung („Gewichtung“) bei der Varianzschätzung berücksichtigt werden kann.

Formal kann die Verwendung von Gewichten, die aus der Anpassung der MZ-Fallzahlen an die Bevölkerungsfortschreibung resultieren, als Regressionsschätzung interpretiert werden. Der hier benutzte Regressionsschätzer basiert auf dem Group Mean Modell, das sich wie folgt darstellen lässt.

¹³ Die Anpassung der MZ-Ergebnisse an die Bevölkerungsfortschreibung nach den Klassen Geschlecht in Kombination mit Staatsangehörigkeit (Deutsche/Ausländer) erfolgt regional auf der Ebene von Anpassungsschichten; das sind regionale Einheiten mit wenigstens 500.000 Einwohnern. Bei der Anpassung der Ergebnisse für Soldaten (einschl. Bundesgrenzschutz und Bereitschaftspolizei) und Wehrpflichtige werden Bestandsmeldungen aus dem Verteidigungs- bzw. Innenministerium auf Regierungsbezirks- bzw. Länderebene herangezogen.

Die Grundgesamtheit U lässt sich in G disjunkte Teilmengen U_g , $g \in \{1, \dots, G\}$ zerlegen. Für die Elemente innerhalb jeder Gruppe g gilt:

$$(18) \quad E_{\xi}(y_k) = \beta_g \quad k \in U_g$$

$$(19) \quad V_{\xi}(y_k) = \sigma_g^2 \quad k \in U_g$$

Hierbei bezeichnet $E_{\xi}(\cdot)$ die Erwartungswertbildung bezüglich einer Verteilung ξ über den Merkmalen und $V_{\xi}(\cdot)$ die entsprechende Varianz. Die Erwartungswertbildung bezieht sich hier also nicht auf den Wert eines Populationsschätzers bezüglich des Erhebungsdesigns, sondern auf die Realisierung der Merkmalswerte bei einer gegebenen Stichprobe.¹⁴

Die OLS-Schätzung von β_g auf Basis der Daten für die Grundgesamtheit ist:

$$(20) \quad B_g = \frac{1}{N_g} \sum_{k \in U_g} y_k = \bar{y}_{U_g}$$

wobei: $N_g =$ Anzahl Elemente von U_g . Die Gruppendifinition liefert eine Zerlegung der Stichprobe: $s_g = s \cap U_g$. Als Schätzer für B_g benutzt man:

$$(21) \quad \hat{B}_g = \frac{\sum_{k \in s_g} y_k / \pi_k}{\sum_{k \in s_g} 1 / \pi_k} = \frac{1}{\hat{N}_g} \sum_{k \in s_g} \frac{y_k}{\pi_k}$$

Hierbei ist \hat{N}_g der geschätzte Wert für den Umfang der Gruppe g . Im Fall des FAMZ sind die Zichungswahrscheinlichkeiten aller Einheiten gleich. Folglich gilt:

$$(22) \quad \hat{B}_g = \bar{y}_{s_g} \quad g = (1, \dots, G)$$

¹⁴ Die Verwendung des statistischen Modells wird nur für die Herleitung des Populationsschätzers benötigt. Die Eigenschaften des resultierenden Schätzers werden jedoch wieder unter dem klassischen Design-basierten Ansatz ermittelt. Särndal et al. (1992: 238) sprechen daher von einem ‚model assisted‘ Ansatz. Der Vorteil der hier gewählten Darstellung liegt in allgemein hergeleiteten Formeln für die Schätzung der Varianz von Regressionsschätzern.

Es bezeichne \hat{y}_k den durch das Regressionsmodell geschätzten Wert von y_k . Im vorliegenden Fall erhält man:

$$(23) \quad \hat{y}_k = \hat{B}_g = \bar{y}_{s_g} \quad k \in s_g$$

Der Regressionsschätzer \hat{t}_{reg} hat für das Group Mean Modell die folgende Gestalt:

$$(24) \quad \hat{t}_{reg} = \sum_{k \in U} \hat{y}_k = \sum_{g=1}^G \sum_{k \in U_g} \hat{B}_g = \sum_{g=1}^G N_g \hat{B}_g = \sum_{g=1}^G \sum_{k \in s_g} \frac{N_g}{\hat{N}_g} \cdot \frac{y_k}{\pi_k} = \sum_{g=1}^G \sum_{k \in s_g} w_g \frac{y_k}{\pi_k}$$

Der Faktor $w_g = N_g / \hat{N}_g$ beschreibt das Verhältnis von $N_g =$ Umfang von Gruppe g in der Grundgesamtheit (= Soll - Vorgabe) zu $\hat{N}_g =$ geschätzter Umfang von Gruppe g (= Ist - Wert).

Bei der Anpassung an die Bevölkerungsfortschreibung wird für N_g jeweils der Wert für bestimmte Gruppen gemäß dieser Fortschreibung gewählt. Im FAMZ ist eine GewichtungsvARIABLE w_k enthalten, die für jede Person $k \in s_g$ den jeweiligen Wert N_g / \hat{N}_g („Soll durch Ist“) annimmt.¹⁵ Der Regressionsschätzer lässt sich damit als ein „gewichtetes Mittel“ darstellen:

$$(25) \quad \hat{t}_{reg} = \sum_{k \in s} w_k \frac{y_k}{\pi_k}$$

Eine wesentliche Eigenschaft des Regressionsschätzers liegt allgemein darin, dass er für Merkmale, die in die Soll/Ist-Anpassung eingingen, unter jeder Stichprobe wieder die Soll-Werte liefert (vgl. Särndal et al. 1992: 324). Folglich hat \hat{t}_{reg} für diese Merkmale die Varianz Null. Allerdings ist der FAMZ nur eine 70-Prozent-Substichprobe aus dem MZ und wurde nicht extra an die Bevölkerungsfortschreibung angepasst. Daher wird die Varianzschätzung für den Regressionsschätzer auch im Fall der Anpassungsmerkmale positive Werte liefern.

¹⁵ Darüber hinaus liegt ein Personengewicht für die Hochrechnung der Unterstichprobe vor, in der Merkmale der Arbeitskräftestichprobe der EU sowie Ergänzungs- und Zusatzprogramme erhoben werden. Für Haushalte bzw. Familien wurde das Haushaltsgewicht als arithmetisches Mittel der Personenfaktoren im Haushalt gebildet.

Für die Herleitung von $V(\hat{t}_{reg})$ wird wieder eine Taylorentwicklung von \hat{t} benutzt. Der lineare Teil der Taylorentwicklung ist durch die folgende Hilfsgröße u_k gegeben (vgl. Särndal et al. 1992: 331):

$$(26) \quad u_k = \frac{N_g}{\hat{N}_g} (y_k - \hat{B}_g) = w_k (y_k - \bar{y}_{s_g}) \quad k \in s_g$$

Dieser lineare Anteil ist also die mit w_k gewichtete Abweichung des Merkmalswerts y_k von dem jeweiligen Gruppenmittelwert \bar{y}_{s_g} in der Stichprobe. Als asymptotische Varianz $V(\hat{t}_{reg})$ wird die Varianz dieses Hilfsmerkmals u verwendet. Bei der praktischen Berechnung hat man also lediglich in Gleichung (8) - (13) y_k durch u_k zu ersetzen.¹⁶

In Tabelle 3 wird für die Merkmale von Tabelle 1 der Einfluss der Anpassung an die Bevölkerungsfortschreibung auf die Schätzung von Totals und ihrer Varianz dargestellt. Es zeigt sich, dass der relative Standardfehler durch die Anpassung etwas verringert wird. Für Merkmale, die eng mit den Anpassungsklassen zusammenhängen, fällt die Reduktion stärker aus. Beispielsweise reduziert sich der relative Standardfehler für das Merkmal „1-Personenhaushalte, weiblich“ von 0,47 auf 0,27 Prozent; ähnlich stark auch beim Merkmal „Ausländische Erwerbsperson“. Jedoch besteht ein bemerkenswerter Unterschied zwischen den Schätzwerten in der Größenordnung von teilweise mehreren Millionen, bei weiblichen 1-Personenhaushalten um fast 900 Tsd. Haushalte. Dies zeigt die Problematik bei der Verwendung der Hochrechnungsfaktoren auf Basis der Bevölkerungsfortschreibung: Sie führen insgesamt betrachtet praktisch kaum zu einer Verringerung der Varianz, sondern verdecken vielmehr einen Bias. Entweder liefert der MZ und damit auch der FAMZ verzerrte Populationsschätzer oder aber die Bevölkerungsfortschreibung produziert ihrerseits verfälschte Schätzungen.

¹⁶ Für Haushaltsauswertungen können die Anpassungsklassen nach den Eigenschaften der Haushaltsbezugs person gebildet werden.

Tabelle 3: Schätzung von Bevölkerungstotalen und relativer Standardfehler (in Prozent) für ausgewählte Merkmale mit und ohne Anpassung an die Bevölkerungsfortschreibung

Merkmal Y	Total (in 1000)		relativer Standardfehler (in %)	
	mit Anp.	ohne Anp.	mit Anp.	ohne Anp.
Ausländische Erwerbspersonen ¹⁾	3.609,7	2.283,7	0,72	1,29
Sofort verfügbare Erwerbslose ¹⁾	3.490,3	2.976,7	0,79	0,81
Überwiegender Lebensunterhalt durch Erwerbstätigkeit ¹⁾	33.806,1	29.607,9	0,20	0,28
Weibliche Erwerbstätige mit einem monatlichen Nettoeinkommen unter 600 DM ¹⁾	1.740,7	1.545,4	1,00	1,03
Privathaushalte mit 1 Person, weiblich ²⁾	7.896,6	7.010,1	0,27	0,47

Quelle: Mikrozensus 1996 (faktisch anonymisierte 70%-Substichprobe)

Subpopulation — verwendeter Hochrechnungsfaktor:

- 1) Bevölkerung am Hauptwohnsitz — Personen-Hochrechnungsfaktor
- 2) Privathaushalte — Haushalts-/Familienhochrechnungsfaktor.

Als potentielle Quelle dieses Bias kommen beim Mikrozensus Probleme hinsichtlich der Realisierung der Stichprobe in Frage, wie beispielsweise Aktualität des Stichprobenplans, Erreichbarkeit der Haushalte und Unit-Nonresponse usw. In diesem Zusammenhang darf vermutet werden, dass die Untererfassung der Soldaten und Wehrpflichtigen, die sich in einem Soll/Ist-Verhältnis von ca. 1,6 widerspiegelt, durch Nichterreichbarkeit verursacht wird.¹⁷ Aber auch die Bevölkerungsfortschreibung ist nicht fehlerfrei. Insbesondere bei den Ausländern wird vermutet, dass Wegzüge nur unzureichend erfasst sind und die Bevölkerungsfortschreibung deshalb zu hohe Ausländerzahlen ausweist. Mit der Gewichtung werden die Mikrozensusergebnisse auf Ergebnisse der Bevölkerungsfortschreibung adjustiert, was zur Übertragung von Fehlern der Bevölkerungsfortschreibung auf den Mikrozensus führen kann. Die Anpassung führt beim FAMZ zu einer Hochgewichtung der Deutschen um einen Faktor von circa 1,1 und der Ausländer sogar um über 1,5.

¹⁷ Gemäß dem Stichprobenplan umfassen die Auswahlbezirke (PSU's) keine Kasernen. Soldaten und Wehrpflichtige sind somit nicht in den Kasernen, sondern nur an ihrem Haupt- bzw. Nebenwohnsitz erreichbar.

Eine „Korrektur“ in dieser Größe lässt sich einerseits nicht aus den Ergebnissen der Feldarbeit herleiten (siehe Heidenreich 1994: 116). Andererseits zeigen Vergleiche der Ergebnisse der Bevölkerungsfortschreibung und der Volkszählung 1987 wesentlich kleinere Abweichungen (Heidenreich 1989: 328; Jäger 1992: 105f.; vgl. auch Krug et al. 1999: 234). Die beträchtlichen Differenzen zwischen MZ und Bevölkerungsfortschreibung können somit keinesfalls überwiegend auf Fehler der Bevölkerungsfortschreibung zurückgeführt werden. Zur Klärung dieser Frage fehlen jedoch systematische Untersuchungen über die Fehlerquellen und zu Abweichungen zwischen Mikrozensus und anderen Bevölkerungsstatistiken.

9. Design-Effekte

Wie am Anfang des vierten Abschnitts beispielhaft dargestellt, waren bisher die Nutzer des FAMZ bei der Schätzung der Varianz von Totals auf die Verwendung von Ergebnissen der Fehlerrechnung des Statistischen Bundesamtes (1998a) angewiesen, in denen Design-Effekte („Zuschlagsfaktoren“) berichtet werden. Ein Design-Effekt beschreibt das Verhältnis des design-basierten Standardfehlers des MZ zum Standardfehler einer Stichprobenziehung gleichen Umfangs, die jedoch ohne Klumpung und Schichtung durchgeführt worden wäre – also unter der Annahme einer uneingeschränkten Zufallsauswahl. Kern dieses Ansatzes ist eine lineare Regression der Design-Effekte auf den geschätzten Populationsanteil der durch das Tabellenfeld definierten Merkmalsträger.

Durch die Substichprobenziehung verringert sich im FAMZ die Zahl der Haushalte und Personen pro Auswahlbezirk im Vergleich zum MZ, und damit auch der Klumpeneffekt. Es ist deshalb davon auszugehen, dass die für den MZ veröffentlichten Design-Effekte nicht einfach auf den FAMZ übertragbar sind und ihre Verwendung zu einer Überschätzung der Varianz im FAMZ führt. Wir wollen daher prüfen, ob die Ziehung des FAMZ zu anderen Design-Effekten führt. Außerdem soll untersucht werden, wie gut die lineare Approximation der Design-Effekte ist.

Es bezeichne U_d eine Teilmenge (Domain) der Grundgesamtheit, die das Feld einer Tabelle charakterisiert. N_d sei die Anzahl von U_d in der Grundgesamtheit und n_d in der Stichprobe.

Mit Hilfe der Indikatorfunktion

$$y_k = \begin{cases} 1 & \text{falls } k \in U_d \\ 0 & \text{sonst} \end{cases}$$

kann man die Schätzung von Domains auf die Schätzung von Totals zurückführen, da in diesem Falle $N_d = \sum_{k \in U} y_k = t_y$ gilt. Für die Schätzung des Populationsanteils $P_d = N_d/N$ durch $\hat{P}_d = \hat{t}_y/N$ erhält man im Fall einer uneingeschränkten Zufallsauswahl (vgl. Särndal et al. 1992: 70):

$$(27) \quad \hat{V}_{SI}(\hat{P}_d) = \frac{1-f}{n-1} p_d(1-p_d)$$

wobei $p_d = n_d/n$ das Stichprobenpendant zu P_d ist.

Der relative Standardfehler ist der Variationskoeffizient von \hat{P}_d . Für die uneingeschränkte Zufallsauswahl ist er gegeben durch:

$$(28) \quad cv_{SI} = \frac{\sqrt{\hat{V}_{SI}(\hat{P}_d)}}{\hat{P}_d} = \sqrt{\frac{1-f}{n-1} \frac{1-p_d}{p_d}}$$

Der Design-Effekt $k(\hat{p}_d)$ ist dann definiert durch:¹⁸

$$(29) \quad k(\hat{p}_d) = \frac{\sqrt{\hat{V}(\hat{p}_d)}}{\sqrt{\hat{V}_{SI}(\hat{p}_d)}}$$

Dieses Verhältnis kann für jede Subpopulation anders ausfallen. Das Statistische Bundesamt benutzt für drei Merkmalsgruppen (Bevölkerung und Erwerbstätige (B/E), Ausländer und Erwerbstätige in der Landwirtschaft (A/L) und Haushalte (H)) eine unterschiedliche einfache lineare Regression von $k(\hat{p}_d)$ auf \hat{p}_d :

$$(30) \quad k(\hat{p}_d) \approx a + b\hat{p}_d$$

¹⁸ Häufig wird der Design-Effekt als das Verhältnis der Varianzen definiert; so zum Beispiel bei Särndal et al. (1992: 54).

Für gegebene Werte von a und b erhält man unter Verwendung von (29) und (30) die folgende Näherung für den relativen Standardfehler in Abhängigkeit von p_d :

$$(31) \quad cv = k(p_d)cv_{SI} = (a + bp_d) \sqrt{\frac{1-f}{n-1} \frac{1-p_d}{p_d}}$$

Tabelle 4 vergleicht zusammenfassend die auf Basis des FAMZ jeweils für die einzelnen Gruppen ermittelten Regressionskoeffizienten mit den für den MZ 1990 veröffentlichten Werten und den Berechnungen für den MZ 1996. Insgesamt verlaufen die Geraden für den FAMZ 1996 flacher als für den MZ 1996. Bezogen auf die über 450 Merkmale des Standardauswertungsprogramms beträgt die Reduktion etwa 10 Prozent und ist für die drei Merkmalsgruppen B/E, A/L und H in etwa gleich. Diese Verringerung des Design-Effekts hängt mit der Reduktion des Klumpeneffekts bei der Substichprobenziehung zusammen. Lässt man einmal außer acht, dass die in der Näherungsfunktion verwendeten Koeffizienten für den MZ 1990 veraltet sind und sich nur auf das frühere Bundesgebiet beziehen, kann durch Vergleich der MZ90- und FAMZ-Werte in Tabelle 4 festgestellt werden, dass der relative Standardfehler des FAMZ für die hier ausgewählten Merkmale in brauchbarer Näherung wiedergegeben wird.

Tabelle 4: Vergleich der Koeffizienten einer einfachen linearen Regression des Design-Effekts auf den geschätzten Bevölkerungsanteil

Gruppe	Datenbasis	Konstante	Steigung
Bevölkerung und Erwerbstätige (B/E)	FAMZ	1.009	1.84
	MZ96	1.042	2.44
	MZ90	1.136	1.61
Ausländer und Erwerbstätige in der Landwirtschaft (A/L)	FAMZ	1.088	21.69
	MZ96	1.162	25.47
	MZ90	1.169	25.04
Haushalte (H)	FAMZ	0.988	1.01
	MZ96	1.009	1.60
	MZ90	1.119	1.14

Datenbasis:

FAMZ: Mikrozensus 1996 (faktisch anonymisierte 70%-Substichprobe).

MZ96: Eigene Berechnungen auf Basis unveröffentlichter Fehlerrechnungen zum Mikrozensus 1996 (Statistisches Bundesamt 1998b).

MZ90: Fehlerrechnung zum Mikrozensus 1990 (Statistisches Bundesamt 1998a: 22).

Es bleibt noch zu überprüfen, wie gut die lineare Beziehung in Gleichung (30) zu den berechneten Design-Effekten passt.¹⁹ Die Abbildungen 1 bis 3 zeigen die Regressionen der Design-Effekte für die Merkmale des MZ-Standardauswertungsprogramms für die Gruppen Bevölkerung und Erwerbstätige (B/E), Ausländer und Erwerbstätige in der Landwirtschaft (A/L) sowie Haushalte (H). Für das Beispiel weibliche 1-Personenhaushalte wurde empirisch ein Design-Effekt von 1,19 ermittelt (siehe Tabelle 1). Verwendet man die lineare Approximation, erhält man mit einem Wert von 1,20 eine sehr gute Schätzung. Hingegen streuen die empirisch ermittelten Design-Effekte für die Merkmalsgruppe Bevölkerung und Erwerbstätige (B/E) wesentlich stärker. Für einen Anteilswert von rund 20 Prozent liegt das Minimum der Design-Effekte bei 1,08 und das Maximum bei 1,56; durch die Regression wird für diesen Anteil ein Design-Effekt von rund 1,38 geschätzt. Die auf Basis direkter Varianzschätzungen ermittelten Konfidenzintervalle können sich folglich um mehrere Zehntausend Personen von dem auf der linearen Approximation basierenden Konfidenzintervall unterscheiden und zu verschiedenen Aussagen führen.

Insgesamt zeigt sich, dass die lineare Approximation zwar ein brauchbares Modell zur Beschreibung des Design-Effekts liefert, aber bei einzelnen Merkmalen doch beträchtliche Abweichungen des jeweiligen Design-Effekts von der Regressionsgeraden vorliegen. In Einzelfällen führt also die Verwendung der Design-Effekte zu erheblichen Über- bzw. Unterschätzungen der Varianz, wobei über die Richtung des Fehlers keine Aussage gemacht werden kann. Mit dem Vorliegen der Information über den Auswahlbezirk und die Gebäudeschicht ist es jedoch nicht mehr nötig, sich auf diese Varianzabschätzung zu verlassen. Nutzer der Mikrozensus 1991-1995, in denen diese Stichprobeninformationen nicht enthalten sind, können ersatzweise die hier berichteten Design-Effekte für den FAMZ 1996 bei Varianzschätzungen verwenden.

¹⁹ Für den MZ sind bisher nur für 1990 die entsprechenden Regressionskoeffizienten veröffentlicht worden. Eine Überprüfung der Angemessenheit des einfachen Regressionsmodells ist bis auf den Hinweis, daß die Abweichungen der berechneten von den geschätzten Design-Effekten im Mittel kleiner als 15-20 Prozent sind (Statistisches Bundesamt 1998a: 17), bisher nicht dokumentiert worden.

Abbildung 1: Regression des Design-Effekts (k) für die Merkmalsgruppe Bevölkerung und Erwerbstätige (B/E) auf den Anteilswert für 346 Merkmale des FAMZ

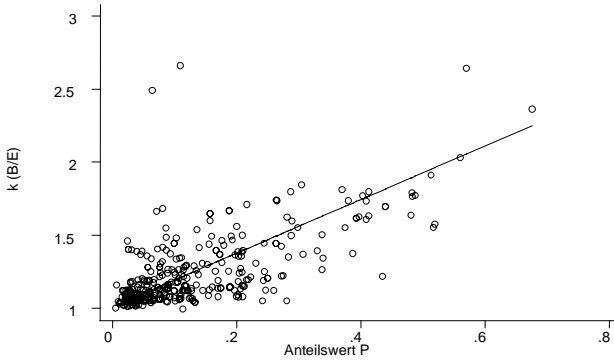


Abbildung 2: Regression des Design-Effekts (k) für die Merkmalsgruppe Ausländer und Erwerbstätige in der Landwirtschaft (A/L) auf den Anteilswert für 18 Merkmale des FAMZ

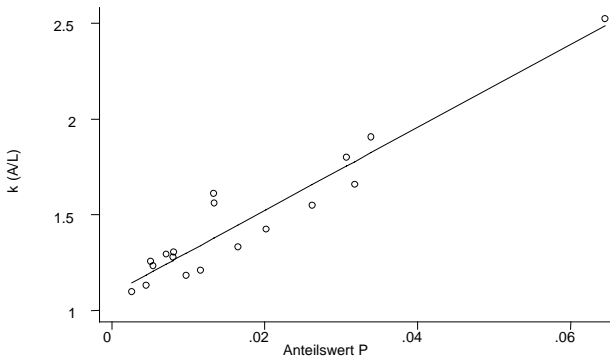
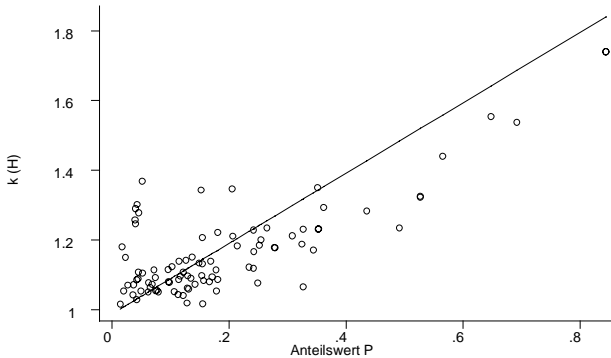


Abbildung 3: Regression des Design-Effekts (k) für die Merkmalsgruppe Haushalte (H) auf den Anteilswert für 94 Merkmale des FAMZ



10. Abschließende Bemerkungen

Im vorliegenden Beitrag wurde gezeigt, dass die ab dem Scientific Use File des Mikrozensus 1996 mögliche Varianzschätzung gegenüber dem bisher verwendeten Näherungsverfahren zu niedrigeren Standardfehlern führt. Dieses Ergebnis hängt hauptsächlich mit der Reduzierung des Klumpeneffekts infolge der Substichprobenziehung zusammen. Die wichtigsten Voraussetzungen für die verbesserte Varianzschätzung sind mit der Weitergabe anonymisierter Informationen über die Zugehörigkeit zum Auswahlbezirk und zur Gebäudeschicht sowie einigen groben Regionalmerkmalen erfüllt. Dies eröffnet den Nutzern die Möglichkeit, eine wesentliche Eigenschaft dieses Datensatzes, nämlich die hohe Präzision der Schätzergebnisse bei zugleich sehr geringem Nonresponse, effizient auszuschöpfen. Neben Methoden für die Schätzung von Totals, Anteils- und Mittelwerten wurde eine Regressionsschätzung für Totals entwickelt. Damit können die an die Bevölkerungsfortschreibung angepassten Mikrozensusergebnisse statistisch angemessen behandelt werden. Es hat sich gezeigt, dass die Varianzschätzungen für das Scientific Use File des Mikrozensus ab 1996, die nur unter Berücksichtigung der ersten Auswahlstufe der PSU's durchgeführt werden, eine sehr gute Näherung der Gesamtvarianz erge-

ben. Für diese vereinfachte Schätzung können deshalb auch Standardprozeduren der Statistikpakete STATA und SAS verwendet werden.²⁰

Die Analysen haben gleichzeitig die Grenzen der klassischen Stichprobentheorie deutlich gemacht, die sich ausschließlich auf den unter der Annahme einer idealen Umsetzung des Stichprobenplans berechenbaren Stichprobenfehler konzentriert und die in der Praxis auftretenden systematischen Fehler – Befragungsausfälle, Antwort- und Verkodungsfehler etc. – außer Acht lässt. Insbesondere die Differenzen der Mikrozensussergebnisse vor und nach der Anpassung an die Bevölkerungsfortschreibung weisen darauf hin, dass für Aussagen zur Qualität statistischer Daten ein großer Bedarf an systematischen Untersuchungen besteht. Vor dem Hintergrund, dass die nächste Volkszählung nicht mehr in der Form einer primärstatistischen Totalerhebung durchgeführt wird, sondern statt dessen verschiedene Datenquellen – unter anderem Ergebnisse des Mikrozensus – in einem registergestützten System zusammengeführt werden, sind Fragen der Datenqualität von zentraler Bedeutung. Hier ist aber nicht nur die amtliche Statistik gefordert.²¹

Damit auch von Seiten der akademischen Forschung Methodenstudien durchgeführt werden können, ist es wünschenswert, weitere feldbedingte Merkmale im Scientific Use File zur Verfügung zu stellen. In bezug auf Varianzschätzungen ist die Zugehörigkeit zu den Rotationsgruppen im Mikrozensus von besonderem Interesse. Bei Kenntnis der Rotationsgruppen kann man die Varianz der Schätzungen durch die Varianz der Schätzergebnisse über die Rotationsgruppen schätzen. Eine Identifikation der Zugehörigkeit zur Rotationsgruppe ist außerdem für den Vergleich von Kennwerten auf der Basis einzelner Querschnittsangaben (zum Beispiel Veränderungen 1996-1997) für Varianzschätzungen nötig. In diesem Zusammenhang sei angemerkt, dass die ab dem Mikrozensus 1996 mögliche Zusammenführung der einzelnen Mikrozensusserhebungen zu einem rotierenden Panel für die Forschung von allergrößtem Interesse ist. Es wäre wünschenswert, wenn auch die Scientific Use Files zu einem Rotationspanel zusammengeführt werden können.

20 Eine Dokumentation der hier verwendeten SAS-, SPSS- und STATA-Programme, die auch die zweite Auswahlstufe berücksichtigen, ist als ZUMA-Methodenbericht in Vorbereitung. Die Programme werden in Kürze auf den WWW-Seiten der Abteilung Mikrodaten bei ZUMA (www.gesis.org/dauerbeobachtung/mikrodaten) zur Verfügung gestellt.

21 In anderen Ländern wird das Thema Datenqualität schon seit längerer Zeit verstärkt untersucht. Beispielsweise hat das statistische Amt im Vereinigten Königreich hierzu eine Methods and Quality Division eingerichtet und arbeitet auf diesem Gebiet eng mit Statistikern der Universität Southampton zusammen (Holt/Jones 1998).

Hinsichtlich der Varianzschätzung beziehen sich weitere Verbesserungsmöglichkeiten des Datensatzes auf die Bereitstellung des Kompensationsgewichts und von Informationen über die Ziehung der 70-Prozent-Substichprobe aus dem Mikrozensus. Bei Kenntnis der Zahl der Haushalte vor der Substichprobenziehung ist es möglich, die durch diese Auswahlstufe bedingte Varianzkomponente zu schätzen. Da die Merkmale Kompensationsgewicht, Rotationsgruppe und Zahl der MZ-Haushalte pro PSU in bezug auf den Datenschutz kein grundsätzliches Problem darstellen dürften, würde die Weitergabe der Variablen neue methodische und inhaltliche Analysen des Scientific Use File des Mikrozensus bieten.

Korrespondenzadresse

*Prof. Dr. Ulrich Rendtel
Johann Wolfgang Goethe Universität
Institut für Statistik und Mathematik, Fachbereich Wirtschaftswissenschaften (02)
Mertonstraße 17, R138B, 60054 Frankfurt am Main
E-Mail: rendtel@em.uni-frankfurt.de*

Literatur

- Emmerling, D./Riede, T., 1997: 40 Jahre Mikrozensus. *Wirtschaft und Statistik* (3): 160-174.
- Frank, E./Kafurke, A., 1990: Die Mikrozensusstichprobe ab 1990 auf neuer Auswahlgrundlage. *Baden-Württemberg in Wort und Zahl* (4): 154-164.
- Heidenreich, H.-J., 1989: Erwerbstätigkeit im April 1988. *Wirtschaft und Statistik* (6): 327-339.
- Heidenreich, H.-J., 1994: Hochrechnung des Mikrozensus ab 1990. S. 112-123 in: S. Gabler/J.H.P. Hoffmeyer-Zlotnik/D. Krebs (Hrsg.): *Gewichtung in der Umfragepraxis*. Opladen: Westdeutscher Verlag.
- Holt, T./Jones, T., 1998: Quality work and conflicting quality objectives. Paper for the 84th DGINS conference in Stockholm 28-29 May 1998 <URL: http://www.statistics.gov.uk/methods_quality/downloads/qualitywork.doc; letzter Zugriff: 03.11.00>.
- Jäger, M., 1992: Im Westen was Neues? - Im Osten was Besseres? Möglichkeiten der Nutzung von Daten der Einwohnermelderegister für statistische Zwecke. S. 103-124 in: *Statistisches Bundesamt (Hrsg.): Volkszählung 2000 - oder was sonst?* Band 21 der Schriftenreihe Forum der Bundesstatistik. Stuttgart: Metzler-Poeschel.

Köhler, S./Schimpl-Neimanns, B./Schwarz, N., 2000: Pilotprojekt zur Erleichterung der Nutzungsmöglichkeiten von faktisch anonymisierten Mikrodaten. *Wirtschaft und Statistik* (1): 30-37.

Krug, W./Nourney, M./Schmidt, J., 1999: *Wirtschafts- und Sozialstatistik. Gewinnung von Daten.* München: Oldenbourg (5., völlig neubearb. Auflage).

Lüttinger, P., (Hrsg.), 1999: *Sozialstrukturanalysen mit dem Mikrozensus.* ZUMA Nachrichten Spezial, Band 6. Mannheim: ZUMA.

Lüttinger, P./Riede, T., 1997: Der Mikrozensus: amtliche Daten für die Sozialforschung. *ZUMA-Nachrichten* Nr. 41: 19-43.

Meyer, K., 1994: Zum Auswahlplan des Mikrozensus ab 1990. S. 106-111 in: S. Gabler/J.H.P. Hoffmeyer-Zlotnik/D. Krebs (Hrsg.), *Gewichtung in der Umfragepraxis.* Opladen: Westdeutscher Verlag.

Müller, W./Blien, U./Knoche, P./Wirth, H. u.a., 1991: *Die faktische Anonymität von Mikrodaten.* Band 19 der Schriftenreihe Forum der Bundesstatistik, herausgegeben vom Statistischen Bundesamt. Stuttgart: Metzler-Poeschel.

Reinders, M., 1993: Fehlerrechnung zum Mikrozensus 1990. *Statistische Rundschau Nordrhein-Westfalen* (8): 398-404.

Rentel, U./Schimpl-Neimanns, B., 2000: *Varianzschätzungen für den faktisch anonymisierten Mikrozensus.* *Jahrbücher für Nationalökonomie und Statistik* 220(6): 759-776.

Särndal, C.-E./Swensson, B./Wretman, J., 1992: *Model Assisted Survey Sampling.* New York: Springer.

Schmidt, G., 1990: Der Auswahlplan des Mikrozensus ab 1990. *Bayern in Zahlen* (6): 217-221.

Statistisches Bundesamt, 1998a: *Fachserie 1, Bevölkerung und Erwerbstätigkeit. Reihe 4.1.1, Stand und Entwicklung der Erwerbstätigkeit 1996 (Ergebnisse des Mikrozensus).* Stuttgart: Metzler-Poeschel.

Statistisches Bundesamt, 1998b: *Fehlerrechnung Mikrozensus 1996 (nach Kompensation der bekannten Ausfälle).* Wiesbaden (unveröffentlichte Tabellen; StBA VIII C; September 1998).

Statistisches Bundesamt, 1999: Zum Auswahlplan des Mikrozensus ab 1990. S. E2 49-54 in: *Arbeitsunterlagen zum Mikrozensus. Das Erhebungsprogramm des Mikrozensus seit 1957 (Loseblattsammlung).* Wiesbaden.

Werner, J., 1994: *Regionalisierung des Mikrozensus. Baden-Württemberg in Wort und Zahl* (6): 278-285.

Wolter, K., 1985: *Introduction to Variance Estimation.* New York: Springer.