Mikrodaten-Tools

Zur Berechnung des Stichprobenfehlers im Mikrozensus

Bernhard Schimpl-Neimanns
Jörg Müller

Juni 2001

ZUMA
Quadrat B2,1
Postfach 12 21 55
68072 Mannheim
Telefon: 0621-1246-263
Telefax: 0621-1246-100
E-mail: Schimpl-Neimanns@zuma-mannheim.de

1 Einleitung

Der Stichprobenfehler ist ein wichtiger Indikator für die Beurteilung der Qualität von Schätzungen auf Basis von Stichproben. Da in Stichproben nicht alle Einheiten (Haushalte, Personen) der Population befragt werden, sondern die Befragung nur bei einer Zufallsauswahl der Population durchgeführt wird, ist bei einem Schätzwert immer auch die zufallsbedingte Variation zu berücksichtigen. Im Folgenden wird darüber informiert, wie mit Hilfe der ab dem Scientific Use File des Mikrozensus 1996 vorliegenden Informationen zum Stichprobendesign der Stichprobenfehler berechnet werden kann und es werden für eine Reihe von Merkmalen Ergebnisse von design-basierten Varianzschätzungen berichtet. Diese Ergebnisse können näherungsweise auch für die Berechnung des Stichprobenfehlers bei den Scientific Use Files des Mikrozensus 1991-1995 herangezogen werden.

Für die Berechnung des Stichprobenfehlers nach Formeln der klassischen Stichprobentheorie wird angenommen, daß die in der Stichprobe vorliegenden Informationen korrekt sind, d.h. systematische Fehler, die z.B. durch Antwortausfälle und falsche Angaben etc. entstehen, bleiben außer Acht. Das Ausmaß des Stichprobenfehlers hängt vom Stichprobendesign ab. Dessen zentrale Elemente sind:

- Auswahlsatz bzw. die Inklusionswahrscheinlichkeit, die angibt, mit welcher Wahrscheinlichkeit ein Element der Population für die Stichprobe ausgewählt wird,
- Schichtung der Auswahlgesamtheiten, die zur Verringerung des Stichprobenfehlers im Vergleich zur uneingeschränkten Zufallsauswahl beiträgt, und
- Klumpung der Auswahleinheiten, d.h. alle Elementareinheiten eines für die Stichprobe ausgewählten Klumpens gelangen in die Auswahl. Tendenziell vergrößert die Klumpung den Stichprobenfehler im Vergleich zu einer uneingeschränkten Zufallsauswahl.

2 Das Stichprobendesign des Mikrozensus ab 1990

Ab 1990 wurde im Mikrozensus ein neuer Stichprobenplan eingesetzt, dessen wichtigste Kennzeichen wie folgt skizziert werden können:

- Auswahlgesamtheit: Ergebnisse der Volkszählung 1987 (altes Bundesgebiet) und des Zentralen Einwohnerregisters 1991 (neue Bundesländer; ab 1991)
- Klumpung: Die Bildung von Auswahlbezirken als Klumpen (Cluster, Primary Sampling Units (PSU's)) erfolgt innerhalb von vier Gebäudegrößenklassen (Gebäudeschicht, fachli-

che Schichtung). Zur Berücksichtigung von Neubauten und Aktualisierung der Stichprobe werden Neubaubezirke auf Basis der Bautätigkeitsstatistik gebildet.

- Schichtung: Die PSU's der Auswahlgesamtheit werden innerhalb der Gebäudeschicht nach den Merkmalen Bundesland, Regierungsbezirk, Kreis, Gemeindegrößenklasse und Gemeinde angeordnet.
- Auswahlverfahren: Jeweils 100 aufeinander folgende PSU's der Auswahlgesamtheit werden zu einer Zone zusammengefasst und danach einer Zufallszahl von 0 bis 99 zugeordnet. Eine Zone (=100 PSU's) bildet eine mögliche Stichprobe. Da der Mikrozensus eine rotierende Panelstichprobe ist, werden die Zonen des Weiteren in Rotationsviertel (Blöcke) zerlegt. Aus der Auswahlgesamtheit von einhundert 1-Prozent-Stichproben der PSU's werden zwanzig 1-Prozent-Vorratsstichproben gezogen, wobei aus jeder Zone genau eine PSU in die Stichprobe gelangt. Für die vier Gebäudeschichten der Grundauswahl kann das Verfahren als uneingeschränkte Zufallsauswahl beschrieben werden. Für die Neubauschicht wird ein systematisches Ziehungsverfahren mit festem Intervall bei zufälligem Startpunkt angewendet. Im Durchschnitt umfasst ein Auswahlbezirk 9 Wohnungen.

Die Scientific Use Files des Mikrozensus wurden als 70-Prozent-Substichproben wie folgt erstellt:

- Anordnung (nachträgliche Schichtung) der Haushalte nach Bundesland, Regierungsbezirk, Gemeindegrößenklasse, Zahl der Personen im Privathaushalt, Auswahlbezirksnummer, Haushaltsnummer
- Vergabe neuer, fortlaufender Haushaltsnummern
- Auswahlverfahren: Übernahme aller Haushalte in das Scientific Use File, deren letzte Platzziffer der Haushaltsnummer von zwei, fünf und neun verschieden ist. (Abschließend werden Auswahlbezirke und Haushalte neu nummeriert.)

Nach der Bereitstellung von Informationen zum Stichprobendesign ab dem Mikrozensus 1996 können die Stichprobenfehler direkt ermittelt werden. Wie dabei vorgegangen werden kann, wird unten gezeigt. Da dieses Verfahren nicht für die Mikrozensen 1991-1995 anwendbar ist, soll zunächst dargestellt werden, wie mit Hilfe von Design-Effekten der Stichprobenfehler von Gesamtwerten (Totals) näherungsweise berechnet werden kann.

3 Berechnung des Stichprobenfehlers mit Hilfe von Design-Effekten

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.) Die Auswertung des Scientific Use File des Mikrozensus 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*0,7)) ergibt sich hochgerechnet ein Total (t) von 7.010,1 Tausend Haushalten. Man berechnet zunächst den Standardfehler (s) unter Annahme einer uneingeschränkten Zufallsauswahl (simple random sample: SI) und der Binomialverteilung des Totals mit:

$$s_{SI} = ((1-f) / (n-1) * p\,(1-p))^{1/2} = 0{,}00085$$

wobei $f = 0{,}01$

$n = 229.221$

$p = 49.071 / 229.221 = 0{,}214$

Der relative Standardfehler ist der Variationskoeffizient (cv) von p. Er beträgt für dieses Beispiel 0,4 Prozent:

$$cv_{SI} = s_{SI}/p = 0{,}00398$$

Um die Schichtung und Klumpung des MZ näherungsweise zu berücksichtigen, ist dieser Wert mit dem Design-Effekt zu multiplizieren. Der Design-Effekt (k) ist definiert als das Verhältnis des design-basierten Standardfehlers (s) zum Standardfehler einer Stichprobenziehung gleichen Umfangs, die jedoch ohne Klumpung und Schichtung durchgeführt worden wäre – also unter der Annahme einer uneingeschränkten Zufallsauswahl (s_{SI}):

$$k(p) = s(p) / s_{SI}(p)$$

Das Statistische Bundesamt schätzt den Design-Effekt („Zuschlagsfaktor“) für drei Gruppen von Merkmalen mittels einer linearen Regression der empirisch ermittelten Design-Effekte (k) auf den geschätzten Populationsanteil (p) der durch das Tabellenfeld definierten Merkmalsträger.

Der relative Standardfehler, mit dem annähernd das Stichprobendesign berücksichtigt wird, ergibt sich dann nach Multiplikation des geschätzten Design-Effekts (k=a+b*p) mit dem unter der Annahme einer uneingeschränkten Zufallsauswahl geschätzten relativen Standardfehler

(s_{SI}). Verwendet man die auf Basis des Mikrozensus 1990 veröffentlichte Näherungsfunktion des Statistischen Bundesamtes (1998a: 17; siehe auch Tabelle 1), beträgt der Design-Effekt für 1-Personenhaushalte 1,36 (= 1,119 + 1,14 * 0,214). Für das Beispielmerkmal erhält man als Schätzung des relativen Standardfehlers:

$$cv = k * cv_{SI} = (a + b(p)) * cv_{SI} = 1{,}36 * 0{,}00398 = 0{,}54 \text{ Prozent}$$

Der auf das Total (t) bezogene näherungsweise geschätzte Standardfehler (s) beträgt somit (in Tausend) 37,9 (=7.010,1*0,54%). Ein 95-Prozent-Konfidenzintervall für das geschätzte Total von 7.010,1 Tausend weiblichen 1-Personen-Haushalten ergibt:

$$(t \pm 1{,}96*s) = 6.935{,}9 – 7.084{,}3 \text{ Haushalte (in Tausend)}$$

Wie der Vergleich der auf Basis der Scientific Use Files des Mikrozensus 1996 berechneten Design-Effekte mit den Ergebnissen des Statistischen Bundesamtes in Tabelle 1 zeigt, verlaufen die Geraden für das Scientific Use File flacher als für den Original-Mikrozensus. Bezogen auf die über 450 Merkmale des Standardauswertungsprogramms des Statistischen Bundesamtes beträgt die Reduktion etwa 10 Prozent und ist für die drei Merkmalsgruppen in etwa gleich. Diese Verringerung des Design-Effekts hängt mit der Reduktion des Klumpeneffekts zusammen, der bei der Substichprobenziehung durch die Auswahl von Haushalten aus einer PSU und unter approximativer Erhaltung der Schichtung verringert wird. Die Verwendung der vom Statistischen Bundesamt veröffentlichten Design-Effekte bei der näherungsweisen Berechnung des Stichprobenfehlers führt zu einer Überschätzung der Varianz und ist deshalb nicht zu empfehlen.

Nutzer der Scientific Use Files der Mikrozensen 1991-1995, die keine entsprechenden Design-Informationen (Auswahlbezirksnummer, Gebäudeschicht) enthalten, können ersatzweise die für das Scientific Use File des Mikrozensus 1996 ermittelten Ergebnisse verwenden. Da die Files 1991-95 nach dem gleichen Substichprobenziehungsverfahren wie 1996 erstellt wurden, dürften die Design-Effekte für 1996 näher bei den für die Daten 1991-95 nicht bekannten Design-Effekten liegen als die amtlichen Ergebnisse für den Mikrozensus 1990.[1]

[1] Dies trifft nicht für das Scientific Use File des Mikrozensus 1989 zu. Beim Stichprobenplan für die Mikrozensen 1972-1989 betrug die durchschnittliche Klumpengröße noch etwa 22 Wohnungen (ab 1996: ca. 9 Wohnungen). Darüber hinaus war das Bundesland das einzige regionale Schichtungsmerkmal.

Tabelle 1: Vergleich der Koeffizienten einer einfachen linearen Regression des Design-Effekts auf den geschätzten Bevölkerungsanteil

Gruppe	Datenbasis	Konstante	Steigung
Bevölkerung und Erwerbstätige (B/E)	FAMZ97	1,007	1,87
	FAMZ96	1,006	1,84
	MZ96	1,042	2,44
	MZ90	1,136	1,61
Ausländer und Erwerbstätige in der Landwirtschaft (A/L)	FAMZ97	1,085	20,79
	FAMZ96	1,088	21,69
	MZ96	1,162	25,47
	MZ90	1,169	25,04
Haushalte (H)	FAMZ97	0,989	1,02
	FAMZ96	0,988	1,01
	MZ96	1,009	1,60
	MZ90	1,119	1,14

Quelle:
FAMZ97: Scientific Use File des Mikrozensus 1997 (faktisch anonymisierte 70%-Substichprobe)
FAMZ96: Scientific Use File des Mikrozensus 1996 (faktisch anonymisierte 70%-Substichprobe)
MZ96: Mikrozensus 1996 (eigene Berechnungen auf Basis unveröffentlichter Fehlerrechnungen des Statistischen Bundesamtes (1998b)
MZ90: Mikrozensus 1990 (Statistisches Bundesamt 1998a: 17)

Weitere Untersuchungen zur Güte der mittels linearer Regression geschätzten Design-Effekte im Vergleich zu direkt berechneten Werten haben gezeigt, daß die lineare Approximation zwar ein brauchbares Modell zur Beschreibung von Design-Effekten liefert, aber bei einzelnen Merkmalen doch beträchtliche Abweichungen der jeweiligen, direkt berechneten Design-Effekte von der Regressionsgeraden vorliegen. Da in Einzelfällen somit die Verwendung von Design-Effekten zu erheblichen Über- und Unterschätzungen des Stichprobenfehlers führen kann, ist für die Nutzer der Scientific Use Files des Mikrozensus ab 1996 die direkte Varianzschätzung die bessere Alternative. Wie diese mit Hilfe des Statistikpakets STATA vorgenommen werden kann, wird im Folgenden kurz dargestellt.

4 Direkte Berechnung des Stichprobenfehlers ab dem Scientific Use File des Mikrozensus 1996

Die folgende Kurzdarstellung lehnt sich eng an die Arbeiten von Rendtel/Schimpl-Neimanns (2001) und Schimpl-Neimanns/Rendtel (2001) an. Für das Scientific Use File, das als 70-prozentige Substichprobe vorliegt, kann man vereinfachend von einem zweistufigen Auswahlverfahren mit einer einfachen Auswahl auf jeder Stufe ausgehen. Auf der ersten Stufe werden PSU's (Auswahlbezirke) und auf der zweiten Stufe jeweils 70 Prozent der Haushalte einer PSU ausgewählt. Auf Grund der sehr kleinen Auswahlwahrscheinlichkeit der Primäreinheiten

(1. Stufe) von 1 Prozent reicht es für die Varianzberechnung aus, nur diese erste Stufe zu berücksichtigen. Damit können in der Praxis Standardprozeduren von STATA oder SAS verwendet werden. Im ZUMA-Methodenbericht (Schimpl-Neimanns/Rendtel 2001) werden Programme zur Berechnung der Varianz von Populationsschätzern für Totals, Anteils- und Mittelwerte sowie für die Hochrechnung von Mikrozensus-Ergebnissen nach Anpassung an die Bevölkerungsfortschreibung dokumentiert, in denen auch die Varianzen der zweiten Auswahlstufe geschätzt werden.

Im Folgenden soll die Verwendung von Standardprozeduren von STATA für die Varianzschätzung skizziert werden. Zur Abgrenzung der Schichtung können die Variablen Bundesland (EF1), Gemeindegrößenklasse (EF708) und die Gebäudeschicht (EF712) in einer kombinierten Variablen gebildet werden. Es sind dann die Primäreinheiten (PSU's; EF3 Auwahlbezirksnummer) und Inklusionswahrscheinlichkeiten etc. zu definieren. Mit dem Kommando svytotal wird für ein Merkmal y für eine Subpopulation (z) eine Varianzschätzung angefordert. Die Ausgabe von Design-Effekten wird durch das Schlüsselwort „deff" erreicht, wobei zu beachten ist, daß sich diese Design-Effekte nicht auf das Verhältnis der Standardfehler, sondern auf das Verhältnis von Varianzen beziehen. Das Schlüsselwort „ci" fordert ein Konfidenzintervall an. Weitere Ausgabe- und Berechnungsoptionen sind möglich.

Beispiel einer Varianzschätzung mit STATA

```
* Konstruktion und Definition der Schichtung
generate schicht = ef1*100 + ef708*10 + ef712
svyset strata schicht
* Definition der Primäreinheiten
svyset psu ef3
* Definition des Stichprobengewichts:
*   Auswahlsatz der 70%-Stichprobe (einfache Hochrechnung)
generate gew = 100/0.7
svyset pweight gew
* Definition der Endlichkeitskorrektur
generate  r = 0.01
svyset fpc r
svytotal y, ci deff subpop(z)
```

Ergebnisse auf Basis der Scientific Use Files des Mikrozensus 1996 und 1997 für Merkmale des Standardtabellenprogramms des Statistischen Bundesamtes, die mit STATA nach obigem Vorgehen berechnet wurden, werden auf den WWW-Seiten von ZUMA bereit gestellt. Für die geschätzten Totals wurde die Anpassung der Mikrozensusergebnisse an die laufende Be-

völkerungsfortschreibung nicht berücksichtigt, sondern es wurde lediglich die sogenannte freie Hochrechnung durchgeführt. In den Tabellen sind die Definitionen der interessierenden Merkmale und der jeweils verwendeten Subpopulationen zur Nachberechnung genannt.

- Design-basierte Schätzung von Gesamtwerten (Totals) für das Scientific Use File des Mikrozensus 1997 (Tabellen)

  (URL: www.gesis.org/Dauerbeobachtung/GML/Service/Mikrodaten-Tools/varianz/mz97_totals.htm)

- Design-basierte Schätzung von Gesamtwerten (Totals) für das Scientific Use File des Mikrozensus 1996 (Tabellen)

  (URL: www.gesis.org/Dauerbeobachtung/GML/Service/Mikrodaten-Tools/varianz/mz96_totals.htm)

Literatur

Rendtel, U./Schimpl-Neimanns, B., 2001: Die Berechnung der Varianz von Populationsschätzern im Scientific Use File des Mikrozensus ab 1996. ZUMA-Nachrichten 48: 85-116. (URL www.gesis.org/Publikationen/Zeitschriften/ZUMA_Nachrichten/documents/-pdfs/zn48_10-bernhard.pdf)

Schimpl-Neimanns, B./Rendtel, U., 2001: SAS-, SPSS- und STATA-Programme zur Berechnung der Varianz von Populationsschätzern im Mikrozensus ab 1996. ZUMA-Methodenbericht 2001/04. (URL www.gesis.org/Publikationen/Berichte/ZUMA_Methodenberichte/documents/pdfs/tb01_04.pdf)

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.