8 Konstruktion von Resampling-Gewichten für die Varianzschätzung

8.1 Einleitung

Zur Schätzung von Populationsparametern sind die in Kapitel 7 beschriebenen Survey-Gewichte ausreichend. Zur Berechnung jeweils korrekter Varianzen bzw. Standardfehler der Schätzer sind jedoch die in diesem Kapitel beschriebenen Resampling-Gewichte erforderlich. Die Stichprobenziehung des HFCS weist mehrere komplexe Merkmale, wie etwa Stratifizierung, Mehrstufigkeit, „Ziehen Proportional to Size" in der ersten Stufe oder „Ziehen ohne Zurücklegen" in der zweiten Stufe, auf. Darüber hinaus werden die Design-Gewichte mit Non-Response- und Poststratifizierungsanpassungen weiterverarbeitet. Bleiben diese Merkmale in der statistischen Analyse unberücksichtigt, kommt es zu Verzerrungen der geschätzten Varianzen der Punktschätzer. Wird z. B. die Stratifizierung außer Acht gelassen, ergeben sich zu große Standardfehler; bleiben die Stufen des Cluster-Sampling unbeachtet, sind die Standardfehler zu klein. Werden außerdem die Design-Gewichte nicht berücksichtigt, sind Beobachtungen mit geringer Auswahlwahrscheinlichkeit in den Stichprobenverteilungen der Statistiken unter- und jene mit hoher Auswahlwahrscheinlichkeit überrepräsentiert (siehe Kolenikov, 2010).

Berücksichtigt die statistische Analyse jedoch das komplexe Erhebungsdesign mit all seinen Merkmalen, so ergibt sich häufig das Problem, dass die mathematischen Funktionen der Varianzschätzer nicht bekannt sind. Daher ist es erforderlich, bei der Analyse eigens für die Varianzschätzung entwickelte Verfahren anzuwenden. Generell lassen sich zwei Kategorien von Varianzschätzungsverfahren unterscheiden: *Resampling-Methoden und Linearisierungsverfahren.*[1]

Bis vor Kurzem wurde in der Literatur dem Linearisierungsverfahren als der weniger rechenintensiven Methode der Vorzug gegeben. Dieses Verfahren hat jedoch den großen Nachteil, dass aufgrund von Datenschutzbestimmungen die für die Linearisierung erforderliche Information nicht vollständig zur Verfügung gestellt werden darf. Das Problem, dass aus Datenschutzgründen gewisse Informationen nicht zur Verfügung stehen, lässt sich beispielsweise durch die Verwendung von Resampling-Gewichten umgehen. Da sich Resampling-Gewichte aus zahlreichen Variablen zusammensetzen und ihre Werte auf Informationen beruhen, die für den Nutzer des Datensatzes nicht verfügbar sind (z. B. Stratum- und PSU-Variablen), kann er auch keine Rückschlüsse auf die Identität einzelner Befragter ziehen (siehe Stata Library, Replicate Weights).

Das Linearisierungsverfahren ist außerdem für die Varianzschätzung von nichtlinearen Statistiken (Mediane, Quartile usw.) ungeeignet, weil es Ableitungen stetiger Funktionen erfordert; Quantilsfunktionen z. B. sind jedoch unstetig. Resampling-Gewichte sind für die Varianzschätzung derartiger Statistiken hingegen gut geeignet (siehe Heeringa et al., 2010).

Aus erwähnten Datenschutzgründen und weil die HFCS-Daten insbesondere die Analyse von Verteilungsparametern, wie z. B. Medianen und Quantilen, ermöglichen, werden zur Varianzschätzung im HFCS Resampling-Gewichte

[1] *Ein ausführlicher Überblick über Varianzschätzungsmethoden findet sich in Levy und Lemeshow (2008) und Heeringa et al. (2010).*

verwendet.[2] Im folgenden Abschnitt wird beschrieben, wie die Resampling-Gewichte für den HFCS in Österreich konstruiert wurden.

8.2 Erstellung von Resampling-Gewichten

8.2.1 Die Resampling-Methode

Bei der Resampling-Methode geht es darum, die Varianz eines geschätzten Populationsparameters zu schätzen. Dabei werden in einem ersten Schritt Populationsparameter für einzelne Untergruppen von Stichprobenbeobachtungen, sogenannte Resamples, geschätzt. Durch die Berechnung der Variabilität dieser geschätzten Populationsparameter über alle Resamples ergibt sich in einem zweiten Schritt die gewünschte Varianz des geschätzten Gesamtpopulationsparameters (siehe Levy und Lemeshow, 2008).

Statt eine ganze Stichprobe je Resample zu speichern, ist es praktischer, die finalen Erhebungsgewichte zu variieren. Anstatt etwa eine Stichprobenbeobachtung zu entfernen, um ein bestimmtes Resample zu konstruieren, kann man ihr in diesem Resample ein Gewicht von null zuweisen. Die Gewichte der übrigen Beobachtungen in demselben Stratum müssen dann hinaufgesetzt werden, um Verzerrungen der Gesamtsummen pro Resample *r* zu vermeiden (siehe Kolenikov, 2010). Resampling-Gewichte $w_i^{(r)}$ für $r = 1, \ldots, R$ werden gemeinsam mit dem HFCS-Datensatz veröffentlicht.

Resamples können auf unterschiedliche Weise erstellt werden. In der Literatur zu Erhebungen werden drei Hauptkategorien von Resampling-Methoden unterschieden: *Balanced Repeated Replication, Jackknife* und *Bootstrapping*. Obwohl die Varianzschätzer all dieser Resampling-Methoden in den meisten Fällen mit zunehmender Stichprobengröße zueinander konvergieren, sind Bootstrapping und Balanced Repeated Replication laut den Ergebnissen von Simulationsstudien besser zur Quantilschätzung geeignet als Jackknife (siehe Kovar et al., 1988). Da Balanced Repeated Replication nur in Designs mit exakt zwei primären Stichprobeneinheiten (PSUs) pro Stratum funktioniert, was auf den HFCS für Österreich nicht zutrifft, fiel letztlich die Entscheidung für das von Rao und Wu (1988) vorgeschlagene und von Rao et al. (1992) erweiterte *(Rescaling)-Bootstrap-Verfahren*. Dieses Verfahren entspricht auch den Vorgaben des Household Finance and Consumption Network der EZB.

Nach diesem Verfahren werden Resamples durch wiederholtes Ziehen mit Zurücklegen der PSUs innerhalb eines Stratums generiert. Durch Imitation des ursprünglichen Stichprobenverfahrens sollen so Näherungswerte für die Stichprobenverteilungen der relevanten Statistik ermittelt werden.

8.2.2 Das Sampling-Error-Calculation-Modell

Um das ursprüngliche Stichprobenverfahren zu imitieren, wird ein Sampling-Error-Calculation-Modell erstellt, das das komplexe Stichprobendesign (siehe Kapitel 6) vereinfacht nachbildet (siehe Heeringa et al., 2010).

Eine notwendige Vereinfachung des Sampling-Error-Calculation-Modells gegenüber dem ursprünglichen Stichprobenverfahren besteht im HFCS für Österreich darin, Strata mit einer einzigen PSU zusammenzulegen, da das

[2] *In Kombination mit multiplen Imputationen ist jedoch in der Literatur die Varianzschätzung von nichtlinearen Statistiken anhand von Resampling-Gewichten noch weitgehend unerforscht..*

Bootstrap-Verfahren mindestens zwei PSUs pro Stratum erfordert. Aufgrund der spezifischen Stratifizierung des HFCS-Stichprobendesigns sind in der Stichprobe Strata mit nur einer PSU relativ häufig: In 50 von 185 Strata wurde nur eine einzige PSU gezogen. Für das Sampling-Error-Calculation-Modell wird jedes Stratum mit nur einer PSU mit dem geografisch nächstliegenden Stratum gepaart, sodass jeweils ein gemeinsames Pseudo-Stratum entsteht. Hierbei wird darauf Bedacht genommen wie viele PSUs sich in diesem geografisch nächstliegenden Stratum befinden. Es wird dabei immer zum nächstgelegenen Stratum mit der kleineren Anzahl an PSUs aggregiert, wodurch die Häufigkeit der notwendigen Aggregation minimiert wird. Auch wenn die geschätzte Varianz durch Strata-Zusammenlegung nach oben verzerrt wird, sollten durch das Zusammenlegen geografisch benachbarter Strata die PSUs im Pseudo-Stratum sehr homogen bleiben. Somit sollte die Verzerrung der geschätzten Varianz so gering wie möglich ausfallen. In diesem Zusammenhang ist darauf hinzuweisen, dass Verzerrungen von Standardfehlern nach oben zu einem Verlust von Teststärke führen, was im Allgemeinen jedoch akzeptabler ist als negative Verzerrungen von Standardfehlern, die zu Ergebnissen führen, die zu oft als statistisch signifikant gelten.

Tabelle 17

HFCS-Design-Strata und HFCS-Pseudo-Strata im Vergleich

	Design-Strata	Pseudo-Strata
Anzahl der Strata	185,0	135,0
Durchschnittliche Größe	3,3	4,6
Mediane Größe	2,0	2,0
Minimale Größe	1,0	2,0
Maximale Größe	37,0	37,0

Quelle: HFCS Austria 2014, OeNB.

Anmerkung: Stratumgrößen gemessen an PSUs pro Stratum.

Tabelle 17 zeigt, inwiefern sich die Stratumgröße (gemessen an der Anzahl der pro Stratum gezogenen PSUs) verändert, wenn statt des ursprünglichen HFCS-Stichprobendesigns das HFCS-Sampling-Error-Calculation-Modell zum Einsatz kommt: Durch Zusammenlegung der Strata im Sampling-Error-Calculation-Modell verringert sich ihre Anzahl von 185 auf 135, d. h., die Stratifizierung ist nach wie vor sehr hoch. Außerdem erhöht sich die durchschnittliche Stratumgröße von 3,3 PSUs auf 4,6 PSUs pro Stratum.

Eine weitere Vereinfachung des HFCS-Sampling-Error-Calculation-Modells gegenüber dem ursprünglichen Stichprobendesign besteht in der Annahme, dass die Stichprobenvarianz in erster Linie auf die erste Stufe der Stichprobenziehung zurückgeht (d. h. die Auswahl der PSUs und nicht der privaten Haushalte in den einzelnen PSUs). Daher wird die zweistufige Stichprobenziehung auf eine einstufige Stichprobenziehung reduziert, bei der alle privaten Haushalte der Bruttostichprobe innerhalb der gezogenen PSUs in die Resample-Stichprobe Eingang finden.

Darüber hinaus werden alle PSUs mit gleicher Wahrscheinlichkeit in den Resamples gezogen. Das Sampling-Error-Calculation-Modell vereinfacht das Stichprobenverfahren also insofern, als die Ziehungswahrscheinlichkeit der PSUs nicht abhängig von der Größe der PSU gemessen an der Anzahl von Haushalten ist.

Weitere Vereinfachungen sind im Sampling-Error-Calculation-Modell nicht erforderlich. So werden die Gewichtsanpassungen für Antwortausfall und Post-Stratifizierung auf dieselbe Weise wie in den ursprünglichen Gewichtungsverfahren

durchgeführt (siehe Kapitel 7); außerdem wird eine Endlichkeitskorrektur[3] vorgenommen.

8.2.3 Erstellung von Resampling-Gewichten

Der Algorithmus zur Konstruktion der Resampling-Gewichte im HFCS besteht aus folgenden Schritten:

Schritt 1: Innerhalb jedes Pseudo-Stratums h werden m_h PSUs mit Zurücklegen gezogen.

Schritt 2: Durch Anpassung der Design-Gewichte der gezogenen Beobachtungen wird ein neuer Satz von Resampling-Gewichten generiert. Dabei werden dieselben Gewichtsanpassungen für Antwortausfall und Poststratifizierung (siehe Abschnitte 7.2.3 und 7.2.4) durchgeführt wie bei den finalen Erhebungsgewichten; außerdem wird eine Endlichkeitskorrektur vorgenommen.

Schritt 3: Durch R-malige Wiederholung von Schritt 1 und 2 werden $r = 1, \ldots, R$ Sätze von Resampling-Gewichten generiert.

In Schritt 1 wird die Anzahl der PSUs m_h, die pro Stratum mit einer Anzahl von ausgewählten PSUs in der Bruttostichprobe n_h gezogen wird, auf $m_h = n_h - 1$ gesetzt. Diese Entscheidung wird häufig getroffen, da sie die Effizienz der Bootstrap-Schätzer gewährleistet, ohne zu Überschreitungen der natürlichen Parameter-Bandbreiten zu führen (siehe Kolenikov, 2010).

In Schritt 2 müssen die finalen Erhebungsgewichte angepasst werden, da einige PSUs dupliziert sein können und einige unter Umständen gar nicht gezogen wurden. Daher werden die einzelnen Resamples im Hinblick auf die Zielpopulation verzerrt sein, weshalb zur Generierung der Resampling-Gewichte die Design-Gewichte auf dieselbe Weise angepasst werden müssen wie bei der Konstruktion der finalen Survey-Gewichte (siehe Kapitel 7). Zusätzlich ist eine Endlichkeitskorrektur (siehe Fußnote 3) erforderlich, weil die Stichprobenziehung der SSUs im ursprünglichen HFCS-Stichprobendesign ohne Zurücklegen vorgenommen wird.[4]

Je höher schließlich in Schritt 3 die Anzahl der Resamples R ist, desto genauer sind die Standardfehlerschätzungen. Wir verwenden $R = 1.000$, also einen Wert im oberen Bereich der in der Literatur üblicherweise empfohlenen Bandbreite (siehe Kolenikov, 2010).

Tabelle 18 zeigt deskriptive statistische Angaben zu einer Auswahl von Resampling-Gewichten des HFCS. Mittelwert und Gesamtsumme der Resampling-Gewichte bleiben aufgrund der homogenen Gewichtsanpassungen unverändert. Im Vergleich zu den finalen Survey-Gewichten des HFCS weisen die Resampling-Gewichte außerdem kleinere Minimalwerte auf, jedoch ist keiner davon gleich null. Diese Werte entsprechen nicht ausgewählten PSUs, denen statt eines Gewichts von null aufgrund der Endlichkeitskorrektur ein kleines positives

[3] *Die Endlichkeitskorrektur berücksichtigt die Varianzreduktion, die dann auftritt, wenn aus einer endlichen Population Stichproben ohne Zurücklegen gezogen werden. Dies ist in der zweiten Stufe des HFCS-Stichprobendesigns in Österreich vorgesehen.*

[4] *Im HFCS-Stichprobendesign werden die PSUs „mit Zurücklegen" gezogen und die SSUs „ohne Zurücklegen". Obwohl das Sampling-Error-Calculation-Modell die zweite Stufe ignoriert, wurde hier trotzdem eine Endlichkeitskorrektur vorgenommen, um dem Umstand Rechnung zu tragen, dass in der Stichprobe Haushalte nicht doppelt vorkommen dürfen. Durch die Endlichkeitskorrektur wird der Bias einer höheren Variabilität der Replicate Weights vermindert.*

Gewicht zugewiesen wird. Dass die Resampling-Gewichte zudem höhere Maximalwerte aufweisen als die finalen Erhebungsgewichte, liegt an den durchgeführten Gewichtsanpassungen: Da einige PSUs in den Resamples nicht gezogen werden, und um dieselben geschätzten Populationsgrößen wie in der ursprünglichen Stichprobe zu erhalten, müssen die Gewichte der Beobachtungen in den gezogenen PSUs erhöht werden.

Tabelle 18

Auswahl von Resampling-Gewichten im HFCS

	Mittelwert	Median	Minimum	Maximum	Gesamtsumme
Finale Erhebungsgewichte	1.289	1.207	287	4.360	3.862.526
1. Satz von Resampling-Gewichten	1.289	1.040	7	14.374	3.862.526
2. Satz von Resampling-Gewichten	1.289	989	10	11.418	3.862.526
3. Satz von Resampling-Gewichten	1.289	1.023	8	10.852	3.862.526
998. Satz von Resampling-Gewichten	1.289	1.104	10	8.369	3.862.526
999. Satz von Resampling-Gewichten	1.289	985	6	11.201	3.862.526
1.000. Satz von Resampling-Gewichten	1.289	974	7	10.349	3.862.526

Quelle: HFCS Austria 2014, OeNB.

Anmerkung: Sämtliche Statistiken beschränken sich auf die erfolgreich interviewten Haushalte.

8.3 Abschließende Bemerkungen

Es wurden 1.000 Sätze von Resampling-Gewichten konstruiert, mit deren Hilfe HFCS-Datennutzer die Standardfehler von Punktschätzern im HFCS korrekt schätzen können. Dies ist deshalb erforderlich, weil aufgrund des komplexen Erhebungsdesigns, das u. a. Stratifizierung, unterschiedliche Stufen von Cluster-Sampling und Gewichtungsanpassungen vorsieht, Verzerrungen der Varianzschätzer auftreten, wenn der Datennutzer diese Design-Merkmale unberücksichtigt lässt.

Die korrekte Berechnung der Standardfehler mithilfe der Resampling-Gewichte erfordert zwar mehr Rechenleistung als Analysen ohne Resampling-Gewichte, doch ist es in der Praxis nicht erforderlich, für die Varianzschätzung alle 1.000 Sätze von Resampling-Gewichten zu verwenden. So kann man eine Varianzschätzung z. B. rascher, aber weniger präzise, mit weniger Resamples durchführen. Wie viele Resamples man verwendet, hängt von der Art des Schätzers und der Größe der untersuchten Population ab. Zur Schätzung der Mittelwerte der Gesamtpopulation werden beispielsweise in der Regel weniger Resamples erforderlich sein als zur Schätzung der Medianwerte spezifischer Populationsuntergruppen.

Eine Anleitung zur korrekten Verwendung der Resampling-Gewichte in Stata findet sich in Kapitel 9 *User Guide*.