



Daniel Assoulin und Jann Potterat, 25. Januar 2013

Strukturerhebung 2010: Erstellen von Schätzungen mittels dem Makro %CalVarSE

1 Einführung

Im August 2012 hat das BFS das Dokument „Anleitung zur Berechnung von Hochrechnungen anhand der Strukturerhebung 2010“ publiziert. Das Dokument beschreibt unter anderem die Referenzpopulation, das Zustandekommen der Stichprobe sowie die Konstruktion der Hochrechnungsgewichte für die im Juni 2012 publizierten Resultate der Strukturerhebung. Zudem enthält das Dokument den methodischen Hintergrund zu den im Juni verwendeten Schätzverfahren, welche auf der SAS-Prozedur SURVEYFREQ basierten. Die in SURVEYFREQ implementierten Standardverfahren für die Varianzschätzung liefern eine konservative (tendenziell zu hohe) Schätzung der Varianz, weil sie den (normalerweise varianzreduzierenden) Einfluss der Kalibrierung nicht berücksichtigen. Aus diesem Grund hat das BFS das Macro %CalVarSE entwickelt, welches Varianzschätzungen liefert, die den Kalibrierungseffekt berücksichtigen. Hervorzuheben ist, dass sich die Unterschiede zwischen den Resultaten von SURVEYFREQ und %CalVarSE nur auf die Varianzschätzungen, nicht aber auf die Schätzwerte (Häufigkeiten, Anteile) beziehen. Im Vergleich zu SURVEYFREQ, welches zur Hochrechnung kategorialer Variablen gedacht ist, ermöglicht %CalVarSE auch Schätzungen (Totale, Mittelwerte, Quotienten) für numerische Variablen. Ziel dieses Dokumentes ist es, Funktionsweise und Methodik von %CalVarSE kurz vorzustellen. Es kann als Nachtrag zum im August publizierten Schreiben betrachtet werden.

2 Das SAS-Macro %CalVarSE

Das SAS-Macro %CalVarSE ermöglicht folgende Schätzungen (auf Ebene Schweiz oder Untersuchungsbe- reich), wobei nebst den Schätzungen jeweils auch eine Varianzschätzung unter Berücksichtigung des Kali- brierungseffektes mitgeliefert wird:

1. die Schätzung einer Summe / eines Totals einer numerischen Variablen. Beispiel: Schätzung der insge- samt für den Arbeitsweg aufgebrauchten Zeit. Das heisst wir interessieren uns für das Total der Variablen `y=timetowork`.
2. die Schätzung eines Mittelwertes einer numerischen Variablen. Beispiel: Schätzung der mittleren, je Person, für den Arbeitsweg aufgebrauchten Zeit. Das heisst, wir interessieren uns für den Mittelwert der Variablen `y=timetowork`.
3. die Schätzung von Häufigkeiten in Kreuztabellen. Beispiel in Anhang F: die Schätzung der Anzahl Schweizer/innen ohne Migrationshintergrund und Ausbildung Sekundarstufe 2 (Sek2).
4. die Schätzung von Anteilen in Kreuztabellen. Beispiel: Anteil Personen mit Ausbildung Sek2 bei Schwei- zer/innen ohne Migrationshintergrund.

5. die Schätzung von Quotienten. Bei Quotienten betrachtet man gewöhnlicherweise zwei numerische Variablen, aber auch Mittelwerte können in %CalVarSE wie Quotienten behandelt werden. Beispielsweise entspricht die mittlere Zeit für den Arbeitsweg in einem Untersuchungsbereich dem Quotienten zwischen dem Total der Variablen timetowork und der Anzahl (Total) Personen in diesem Bereich.

%CalVarSE vs SURVEYFREQ

- Für Anzahlen und Anteile liefern SURVEYFREQ und %CalVarSE identische Schätzungen.
- Varianzschätzungen mittels %CalVarSE berücksichtigen den Effekt der Kalibrierung, Varianzschätzungen mittels SURVEYFREQ nicht.
- Während SURVEYFREQ nur Schätzungen für kategorielle Variablen liefert, können mit %CalVarSE auch Totale, Mittelwerte und Quotienten für numerische Variablen geschätzt werden. Für all diese Fälle liefert %CalVarSE auch eine Varianzschätzung unter Berücksichtigung der Kalibrierung.

Im nächsten Abschnitt wird die Handhabung von %CalVarSE beschrieben. In Anhang D wird auch der Aufruf von %CalVarSE vorgestellt.

2.1 Arbeiten mit %CalVarSE

Folgende Parameter müssen dem Makro %CalVarSE übergeben werden

Tab. 1: Makro-Variablen %CalVarSE

Makrovariable	Charakter	Beschrieb
datain	oblig.	Inputdatensatz SE2010 mit angehängten Kalibrationsvariablen
kateg	oblig.	Was ist zu schätzen? kateg=1: Summe/Total numerische Variable, kateg=2: Mittelwert numerische Variable, kateg=3: Häufigkeiten (Anzahl Personen in Kreuztabelle) kateg=4: Quotient (ratio) , oder Anteil
var	oblig.*	Analysevariable (für kateg=1,2,3). Bei kateg=3; mehrere Variablen möglich, mit * dazwischen
ratio	oblig.*	Analysevariablen für kateg=4: Quotient mit / dazwischen z.B. Miete pro Quadratmeter: ratio=Nettomiete/Wohnfläche
domain	optional	Klassifikationsvariablen (Domains) (für kateg=1,2,4), mehrere Domains möglich, mit * dazw. Achtung: Bei kateg=3 nicht erlaubt: Domains müssen alle in var stehen!
dataout	oblig.	Name der Output csv-Datei (wird in pfadexp abgelegt)
pfadexp	oblig.	Exportpfad csv Output.
strata	oblig.	Schichtvariable(n). Falls mehrere, ein Leerschlag als Trennzeichen dazwischen
weight	oblig.	Kalibriertes Hochrechnungsgewicht (weight2010)
weight_glm	optional	Gewicht für Residuenberechnung in GLM. Es wird gleich weight2010 gesetzt.
		oblig.* heisst: entweder var oder ratio muss angegeben werden!

Beispiele für den Aufruf von %CalVarSE befinden sich in Abschnitt D des Anhanges.

3 Varianz und Konfidenzintervall

Varianzen und Konfidenzintervalle werden sowohl in SURVEYFREQ wie auch in %CalVarSE simultan zu den geschätzten Werten berechnet. Die Varianzschätzung erfolgte bedingt auf die Nettostichprobengrösse in den verschiedenen Zonen ¹ *h.* Das heisst, die Zonen (Variable Zone) werden in der Varianzschätzung als Schichten berücksichtigt und die Menge der antwortenden Personen werden als einfache Zufallsstichproben innerhalb dieser Schichten behandelt ². Aus der Quadratwurzel der Varianz ergibt sich die Standardabweichung der Schätzung. Die Standardabweichung bildete die Grundlage zur Ermittlung der Vertrauensintervalle zum Niveau 95%. Dabei verwendet SURVEYFREQ: Schätzwert $\pm t(97.5, df) \times$ Standardabweichung ³ und %CalVarSE die Normalverteilungsapproximation: Schätzwert $\pm 1.96 \times$ Standardabweichung.

Die im Juni 2012 erstellten Schätzungen von Häufigkeiten sowie deren Varianzen basierten auf SURVEYFREQ. SURVEYFREQ liefert für die geschätzten Häufigkeiten (und Anteile) eine vereinfachte Approximation der Varianzschätzung mittels %CalVarSE. Aus diesem Grund wird in Anhang C zuerst das Schätzverfahren für Häufigkeiten (Totale) basierend auf der Prozedur SURVEYFREQ erklärt, bevor dann in Anhang E technische Details für die Schätzung mittels %CalVarSE gegeben werden. Sowohl in SURVEYFREQ wie auch in %CalVarSE spielen Population- und Stichprobengrössen in den betrachteten Zonen (Schichten) eine wichtige Rolle. Darum wird in Anhang A eine Übersicht zu Stichproben- und Populationgrössen je Zone geliefert. Eine aktualisierte Übersicht zu den Kalibrierungsvariablen findet man in Anhang G.

¹ Dabei entspricht eine Zone im Normalfall einem Kanton. Ausnahmen:

- Zürich gesplittet in Stadt / Restkanton;
- Bern gesplittet in Stadt, Berner Jura und Restkanton.

² Streng genommen handelt es sich hierbei um eine etwas vereinfachende Annahme, welche der Realität aber Nahe kommt.

³ $t(97.5, df)$ bezeichnet das 97.5% Quantil der t-Verteilung mit $df = (\text{Anzahl Beobachtungen} - \text{Anzahl Schichten})$ Freiheitsgraden, welches von der Prozedur SURVEYFREQ automatisch berechnet wird. Vorausgesetzt, dass die Stichprobengrösse nicht zu klein ist, führt die Verwendung des entsprechenden Quantils der Normalverteilung (1.96) auf sehr ähnliche Resultate.

Anhang

A Population und Stichprobe

Tab. 2: Populations und Stichprobengrössen nach Schicht (Zone/Kanton+)

Nr.	Zone h	Populationstotal bei Ziehung	Stichprobengrösse		Populationstotal für Kalibrierung
		N_h	Brutto n_h	Netto (Antwortende) m_h	$\hat{N}_h$
1	AG00	517419	34035	30398	509900
2	AI00	13064	437	389	12814
3	AR00	45061	1580	1424	44128
4	BE00	687613	24302	21913	674582
5	BE01	109887	15918	13019	107178
6	BE02	43329	3243	2917	42576
7	BL00	235112	8233	7410	231553
8	BS00	162278	6088	5143	153239
9	FR00	229692	7230	6612	224578
10	GE00	384421	25419	20703	359847
11	GL00	33045	1091	935	32272
12	GR00	166227	5605	5012	162865
13	JU00	58419	4194	3784	57490
14	LU00	318589	21602	19353	310985
15	NE00	144211	10536	9411	141068
16	NW00	35174	1168	1064	34624
17	OW00	29958	918	845	29328
18	SG00	403896	13681	12101	395992
19	SH00	65771	2296	2026	64315
20	SO00	218624	7525	6676	215052
21	SZ00	123911	4032	3660	121348
22	TG00	209726	13702	12277	206477
23	TI00	287619	19833	17678	281693
24	UR00	29893	1080	992	28915
25	VD00	594981	39565	34393	573030
26	VS00	265670	8188	7304	256526
27	ZG00	95696	5910	5304	93173
28	ZH00	843951	28100	25196	830524
29	ZH01	326081	48817	39282	323181
Total		6679318	364328	317221	6519253

mit BE01=Stadt Bern, BE02=Berner Jura, BE00: Rest KT Bern
und ZH01=Stadt Zürich, ZH00: Rest KT ZH

B Aufruf von Surveyfreq in SAS

Die ersten Resultatetabellen der SE2010, wie jene in Abbildung 1, die im Juni 2012 publiziert wurden, sind mit der Prozedur `proc surveyfreq` der Software SAS berechnet worden. Der Aufruf in SAS geschieht folgendermassen:

```
ods output crosstabs=resultate;
proc surveyfreq data=res.se2010_net total=plan;
strata Zone;
tables var1 * var2 /col cl clwt row;
weight weight2010;
run;
ods output close;
```

Dabei werden folgende in Tabelle 3 aufgelistete Variablen und Dateien verlangt:

Tab. 3: Variablen/Daten surveyfreq

crosstabs=	Output wird in diese Datei geschrieben (resultate)
data=	Datei der Antwortenden (res.se2010_net: 317'221 Zeilen)
total=	Datei mit Populationstotal bei Ziehung je Schicht N_h (plan: 29 Schichten)
strata	Schichtvariable (Zone)
tables	Hochrechnungsvariablen für Kreuztabellen (mehrere mit * dazwischen. Hier var1*var2)
weight	Hochrechnungsgewicht (weight2010)

C Details zur Schätzung mit SURVEYFREQ

Bezeichnet man mit w_i das Hochrechnungsgewicht (weight2010) für Person i , so lässt sich die Schätzung der Anzahl Schweizer/innen ohne Migrationshintergrund (CHoMig) und Ausbildung Sekundarstufe 2 (Sek2) wie folgt schreiben:

$$\hat{N}_{CHoMig, Sek2} = \sum_{i \in r} w_i I(i \in CHoMig \text{ und } i \in Sek2) \quad (1)$$

Dabei ist

- $I(i \in CHoMig \text{ und } i \in Sek2)$ die Indikatorfunktion, welche den Wert 1 annimmt, falls die Person i Schweizer/in ohne Migrationshintergrund ist und Ausbildung Sekundarstufe 2 hat. Erfüllt die Person diese Bedingung nicht, ist die Funktion 0.
- r die Menge der antwortenden Personen

Die Varianz von $\hat{N}_{CHoMig, Sek2}$ wird folgendermassen geschätzt:

$$\hat{V}(\hat{N}_{CHoMig, Sek2}) = \sum_h \frac{m_h}{m_h - 1} \left(1 - \frac{m_h}{N_h}\right) \sum_{i \in r_h} (w_i I(i \in CHoMig \text{ und } i \in Sek2) - \frac{\hat{N}_{CHoMig, Sek2, h}}{m_h})^2 \quad (2)$$

Dabei bezeichnet

- $\hat{N}_{CHoMig, Sek2, h}$ analog zu $\hat{N}_{CHoMig, Sek2}$ die geschätzte Anzahl Schweizer/innen ohne Migrationshintergrund mit Ausbildung Sekundarstufe 2, aber in Schicht/Zone h
- N_h das Populationstotal in Schicht/Zone h .
- r_h die Menge der antwortenden Personen in Schicht resp. Zone h und m_h die Anzahl Personen in r_h (Anzahl vorhandene Antworten in h).

D Aufruf von %CalVarSE in SAS

In der folgenden Tabelle 4 sind ein paar Beispiele für den Aufruf in SAS aufgelistet. Vor dem Aufruf müssen jedoch folgende Punkte kontrolliert werden:

- Es müssen alle 17 Kalibrierungsvariablen (CHVar1 - CHVar7 und KtVar1 - KtVar10), das Hochrechnungsgewicht (weight2010), die Schichtungsvariable (zone) und natürlich alle Analysevariablen in der Datei der Strukturerhebung vorhanden sein.
- Die Datei mit den Schichttotalen bei Ziehung enthält die Schicht als kategorielle (zone) und als numerische Variable (zone_num), sowie die Schichttotalen $N_{h\cdot}$. Achtung: Letztere muss _total_ heissen. (Die Schichttotalen dürfen nicht mit den Kalibrationstotalen verwechselt werden!)
- Für die Schätzung von Quotienten in Untersuchungsbereichen wird in SAS ein Hilfsmakro von SAS.com benötigt. Der Name des Programms lautet: 0-fusion_25033_1_smsub.sas und es war im Dezember 2012 [hier](#)⁴ im Web zu finden.

Tab. 4: Beispiele Aufruf

Bsp 1: Anzahl Personen je höchste Ausbildung x Migrationsstatus %calvarSE(datain=se2010, kateg=3, var=highestCompletedEduAggl*migrationStatusAgg , ratio=, domain=, dataout=estim_ausb_migrat, pfadexp=&pfadexport. , strata=zone, weight=weight2010, weight_glm=weight2010);	
Bsp 2: Mittlere Zeit zur Arbeit nach Geschlecht Variante 1 %calvarSE(datain=se2010, kateg=2, var=timetowork, ratio=, domain=sex, dataout=estim_timetowork_sex2, pfadexp=&pfadexport. , strata=zone, weight=weight2010, weight_glm=weight2010); Variante 2 (dabei ist nenner=1 konstant) %calvarSE(datain=se2010, kateg=4, var=, ratio=timetowork/nenner, domain=sex, dataout=estim_timetowork_sex, pfadexp=&pfadexport. , strata=zone, weight=weight2010, weight_glm=weight2010);	
Bsp 3: Total Personen, die Schweizerdeutsch sprechen nach Geschlecht; Variante 1 %calvarSE(datain=se2010, kateg=3, var=languageSpokenAtHome1*sex, ratio=, domain=, dataout=estim_spracheCH_sex, pfadexp=&pfadexport. , strata=zone, weight=weight2010, weight_glm=weight2010); Variante 2 (dies geht, da var eine 0-1 Variable ist) %calvarSE(datain=se2010, kateg=1, var=languageSpokenAtHome1, ratio=, domain=sex, dataout=estim_spracheCH_sex2, pfadexp=&pfadexport. , strata=zone, weight=weight2010, weight_glm=weight2010);	
Bsp 4: Anteil Personen mit Ausbildung Sekundarschule innerhalb Schweizer/innen ohne Migrationshintergrund a) Vorarbeit data vz2; set se2010; if migrationStatusAgg=1 and highestCompletedEduAggl=1 then ch_omig_sekunar=1; else ch_omig_sekunar=0; if migrationStatusAgg=1 then ch_o_mig=1; else ch_o_mig=0;run; b) Aufruf Schätzung %calvarSE(datain=vz2, kateg=4, var=, ratio=ch_omig_sekunar/ch_o_mig, domain=, dataout=estim_ch_omig_sekunar, pfadexp=&pfadexport. , strata=zone, weight=weight2010, weight_glm=weight2010);	

Bemerkungen:

Rechenzeit In Abhängigkeit der Anzahl Ausprägungen der betrachteten Analysevariablen kann die Rechenzeit recht gross werden. Die Rechenzeit ist natürlich auch von der vorhandenen Kapazität des Rechners (CPU) abhängig.

Output Die Form des %CalVarSE - Outputs entspricht nicht jener der offiziellen BFS-Resultate (s. Abbildung 1). Es existiert aber ein Makro, welches die %CalVarSE-Resultate transponiert und entsprechend darstellen kann. Das Makro war im Dezember 2012 [hier](#)⁵ im Web zu finden.

⁴ <http://support.sas.com/kb/25/033.html>

⁵ <http://www.medicine.mcgill.ca/epidemiology/Joseph/PBelisle/MultiTranspose.html>

E Details zur Schätzung mit %CalVarSE

Das Hochrechnungsmacro %CalVarSE berücksichtigt den Effekt der Kalibrierung im Rahmen der Varianzschätzungen. Im Folgenden werden diese Schätzungen explizit beschrieben. Dabei werden zwei Fälle unterschieden: Die Schätzung eines Totals und die Schätzung eines Quotienten. Zudem wird beschrieben, wie verschiedene Situationen, wie Schätzungen für Untersuchungsbereiche (Domains), auf diese beiden Fälle zurückgeführt werden können.

E.1 Die Schätzung eines Totals

Wir gehen davon aus, dass das Total Y einer Variablen y auf Ebene Schweiz geschätzt wird. Der entsprechende Schätzer ist

$$\hat{Y} = \sum_{i \in r} w_i y_i, \quad (3)$$

wobei r die Menge der antwortenden Personen bezeichnet. Die Varianzschätzung erfolgt stufenweise:

1. Als Erstes betrachten wir das Regressionsmodell

$$y_i = \sum_{j=1}^p \beta_j x_{i,j} + \epsilon_{y,i} ; j = 1, \dots, p. \quad (4)$$

Dabei ist y_i der Wert der Zielvariablen für Individuum i und $x_{i,j}, j = 1, \dots, p$, bezeichnen die Werte der in der Kalibrierung verwendeten Hilfsvariablen für Individuum i . Im Grunde handelt es sich bei den im Abschnitt Kalibrierungstotale umschriebenen Kalibrierungsvariablen allesamt um kategoriale Variablen. Dabei gibt es zwei Fälle:

- Variablen die auf Niveau CH kalibriert wurden
- und Variablen die auf Niveau Region (Zonen plus Städte Genf, Lausanne, Winterthur) kalibriert wurden.

Im mitgelieferten Datensatz sind die entsprechenden Variablen mit

- CHVar1 - CHVar7 (reflektieren Kalibrierung auf Niveau CH)
- KtVar1 - KtVar10 (reflektieren Kalibrierung auf Niveau Region)

bezeichnet. Aufgrund ihrer qualitativen Natur können diese Variablen nicht direkt ins Modell (4) übernommen werden, sondern sie werden zuerst in 1/0 Variablen umcodiert. Innerhalb %CalVarSE wird diese Umcodierung durch die Prozedur GLM automatisch vorgenommen. Dabei wird für jede Ausprägung der Variablen eine separate 1/0 Variable kreiert. Es liegt auf der Hand, dass aus diesem Vorgehen eine grosse Anzahl Indikatorvariablen resultiert. Sie werden mit $x_1, \dots, x_p$ bezeichnet und in (4) als erklärende Variablen verwendet. Die Regressionsparameter sowie die Residuen in (4) werden basierend auf der Nettostichprobe r mittels der Methode der gewichteten kleinsten Quadrate (WLS) geschätzt. Für die Gewichtung wird dabei das Hochrechnungsgewicht w_i (oder optional das Initialgewicht) verwendet. Dies führt zu

$$y_i = \sum_{j=1}^p \hat{\beta}_j x_{i,j} + \hat{\epsilon}_{y,i} ; j = 1, \dots, p ; i \in r. \quad (5)$$

2. Für jedes Individuum $i \in r$ hat man aufgrund von (5) ein geschätztes Residuum $\hat{\epsilon}_{y,i}$, welches durch die verwendeten Hilfsvariablen nicht erklärt werden kann. Dieser unerklärte Teil wird nun für die Varianzschätzung verwendet:

$$\hat{V}(\hat{Y}) = \sum_h \frac{m_h}{m_h - 1} \left(1 - \frac{m_h}{N_h}\right) \sum_{i \in r_h} \left(w_i \hat{\epsilon}_{y,i} - \frac{\sum_{i \in r_h} w_i \hat{\epsilon}_{y,i}}{m_h}\right)^2 \quad (6)$$

Dabei bezeichnen

- N_h das Total Personen in Zone h
- r_h, m_h die Menge, respektive die Anzahl antwortender Personen in Zone h .

Anmerkungen:

- Sowohl die Punktschätzung (3) wie auch das Verfahren für die Varianzschätzung ist für stetige wie auch für kategorielle Zielvariablen gültig. Für kategorielle Variablen werden die verschiedenen Ausprägungen mittels Indikatorvariablen beschrieben, welche dann die Rolle der Zielvariablen y einnehmen. Im betrachteten Beispiel, wo wir uns für die Anzahl Schweizer/innen ohne Migrationshintergrund und Ausbildung Sek2 interessieren, wäre also $y_i = I(i \in CHoMig \text{ und } i \in Sek2)$.
- Die nächste Bemerkung bezieht sich auf den Fall, in welchem nicht das Total auf Ebene Schweiz, sondern ein Total für einen Untersuchungsbereich d (d für Domain) geschätzt werden soll⁶. In so einer Situation wird in sämtlichen Berechnungen statt der Variablen y_i die Variable $y_{d,i} := y_i I(i \in d)$ verwendet. Dabei ist $I(i \in d)$ die Identifikatorfunktion, welche den Wert 1 annimmt, falls Person i in d ist. Andernfalls hat die Funktion den Wert Null. Je Untersuchungsbereich wird also eine spezifische Variable erstellt, welche auf der ganzen Population definiert ist und die dann ganz normal auf das Niveau Schweiz hochgerechnet werden kann, weil ihr Populationstotal dem gesuchten Total des Untersuchungsbereichs entspricht.

E.2 Die Schätzung eines Quotienten

%CalVarSE schätzt den Quotienten $R = \frac{Y}{Q}$ zwischen dem Total der Variablen y und q mittels

$$\hat{R} = \frac{\hat{Y}}{\hat{Q}} = \frac{\sum_{i \in r} w_i y_i}{\sum_{i \in r} w_i q_i}. \quad (7)$$

Dabei sind $\hat{Y}$ und $\hat{Q}$ gewöhnliche Totalschätzungen im Sinne von (3). Die Schritte, welche uns zur Varianzschätzung führen, werden analog zum Fall des Totales aufgelistet.

1. Anhand der Variablen y_i und q_i sowie den Schätzungen von $\hat{Q}$ und $\hat{R}$ berechnen wir für jede Person $i \in r$ die Variable

$$\hat{z}_i = \frac{1}{\hat{Q}}(y_i - \hat{R}q_i) \quad (8)$$

2. Nun wird analog (5) ein Regressionsmodell für die Zielvariable $\hat{z}_i$ geschätzt:

$$\hat{z}_i = \sum_{j=1}^p \hat{\beta}_j x_{i,j} + \hat{\epsilon}_{z,i}; j = 1, \dots, p; i \in r \quad (9)$$

Hieraus resultieren die Residuen $\hat{\epsilon}_{z,i}$, welche im nächsten Schritt für die Varianzschätzung verwendet werden.

3. Varianzschätzung (analog zu (6))

$$\hat{V}(\hat{R}) = \sum_h \frac{m_h}{m_h - 1} \left(1 - \frac{m_h}{N_h}\right) \sum_{i \in r_h} (w_i \hat{\epsilon}_{z,i} - \frac{\sum_{i \in r_h} w_i \hat{\epsilon}_{z,i}}{m_h})^2 \quad (10)$$

Bemerkungen:

- Sowohl die Punktschätzung (7) wie auch die Varianzschätzung (10) sind für stetige wie für kategorielle Variablen gültig. Für kategorielle Variablen müssen die verschiedenen Ausprägungen mittels Indikatorvariablen beschrieben werden, welche dann die Rollen der Zielvariablen y und q einnehmen und man spricht in diesem Fall eher von Anteilen statt von Quotienten. Wenn wir uns beispielsweise für den Anteil Personen mit Ausbildung Sek2 bei Schweizer/innen ohne Migrationshintergrund interessieren, so betrachten wir die Variablen $y_i = I(i \in Sek2 \text{ und } i \in CHoMIG)$ im Zähler und $q_i = I(i \in CHoMIG)$ im Nenner.
- Die nächste Bemerkung bezieht sich auf den Fall, in welchem nicht ein Quotient auf Ebene Schweiz, sondern ein Quotient für eine Subpopulation d (d für Domain) geschätzt werden soll. In so einer Situation werden in sämtlichen Berechnungen statt der Variablen y_i und q_i die Variablen $y_{d,i} := y_i I(i \in d)$ und $q_{d,i} := q_i I(i \in d)$ verwendet. Dabei ist $I(i \in d)$ die Identifikatorfunktion, welche den Wert 1 annimmt, falls Person i in d ist. Andernfalls hat die Funktion den Wert Null. Je Untersuchungsbereich werden also zwei spezifische Variablen erstellt, welche auf der ganzen Population definiert sind und die dann ganz normal auf das Niveau Schweiz hochgerechnet werden können.

⁶ Beispielsweise wäre die Bevölkerung eines gewissen Gebietes eine solche Subpopulation

- Mittelwerte werden in CALVAR als Spezialfall eines Quotienten behandelt. Dabei ist die Variable y im Zähler numerisch und die Variable q im Nenner eine Indikatorvariable für den interessierenden Untersuchungsbereich.

Bemerkung zu Merkmalausfällen: Falls eine Person gewisse Fragen unbeantwortet lässt, resultieren Merkmalausfälle. Dabei ist zwischen kategoriellen und quantitativen Variablen zu unterscheiden:

- Für kategorielle Variablen wurden Merkmalausfälle in speziellen Kategorien zusammengefasst und stellen somit für die Berechnungen kein spezielles Problem dar.
- Für quantitative Variablen geht %CalVarSE wie folgt vor. Beobachtungen mit Merkmalausfall in der betreffenden Variablen werden vom interessierenden Untersuchungsbereich ausgeschlossen und in der Hochrechnung nicht berücksichtigt.
 - Dies führt im Falle einer Totalschätzung zu einer tendenziellen Unterschätzung.
 - Falls Mittelwerte oder Quotienten geschätzt werden, ist der Einfluss der Merkmalausfälle weniger klar.

Der Anteil von Merkmalsausfällen ist generell sehr klein oder gar Null, so dass der Effekt auf die Schätzungen vernachlässigbar sein sollte. Um sicher zu gehen, empfiehlt es sich aber, den Anteil Merkmalsausfälle im interessierenden Untersuchungsbereich im Auge zu behalten.

F Darstellung der Resultate

Ausbildung	Total		Schweizer/in ohne Migrationshintergrund		Schweizer/in mit Migrationshintergrund		Ausländer/in der ersten Generation		Ausländer/in der zweiten Generation		Ausländer/in der dritten und höheren Generation		Migrationshintergrund unbekannt	
	Anzahl	VI ± (in %)	Anzahl	VI ± (in %)	Anzahl	VI ± (in %)	Anzahl	VI ± (in %)	Anzahl	VI ± (in %)	Anzahl	VI ± (in %)	Anzahl	VI ± (in %)
Total	6'519'253	0.1%	4'118'782	0.3%	773'536	1.0%	1'231'511	0.8%	153'239	2.5%	3'555	16.2%	238'630	1.9%
Sekundarstufe 1	1'736'264	0.6%	891'504	0.9%	216'647	1.9%	500'340	1.3%	50'898	4.5%	1'858	22.5%	75'017	3.5%
Sekundarstufe 2	3'081'594	0.4%	2'219'681	0.5%	338'126	1.5%	337'711	1.7%	77'761	3.5%	1'259	27.5%	107'056	2.9%
Tertiärstufe	1'604'334	0.6%	986'776	0.9%	208'287	1.9%	346'253	1.6%	23'206	6.2%	(298)	56.5%	39'514	4.8%
Ausbildung unbekannt	97'061	3.1%	20'820	6.6%	10'476	8.7%	47'207	4.5%	1'373	27.3%	(141)	67.5%	17'044	7.2%

Fig. 1: Ein Auszug aus der Tabelle *ständige Wohnbevölkerung ab 15 Jahren nach Migrationsstatus: Varianzberechnung mit SURVEYFREQ*

Ausbildung	Total		Schweizer/in ohne Migrationshintergrund		Schweizer/in mit Migrationshintergrund		Ausländer/in der ersten Generation		Ausländer/in der zweiten Generation		Ausländer/in der dritten und höheren Generation		Migrationshintergrund unbekannt	
	Anzahl	VI ± (in %)	Anzahl	VI ± (in %)	Anzahl	VI ± (in %)	Anzahl	VI ± (in %)	Anzahl	VI ± (in %)	Anzahl	VI ± (in %)	Anzahl	VI ± (in %)
Total	6'519'253	0.0%	4'118'782	0.2%	773'536	0.9%	1'231'511	0.3%	153'239	1.9%	3'555	15.9%	238'630	1.9%
Sekundarstufe 1	1'736'264	0.5%	891'504	0.8%	216'647	1.9%	500'340	1.0%	50'898	3.9%	1'858	22.2%	75'017	3.5%
Sekundarstufe 2	3'081'594	0.4%	2'219'681	0.5%	338'126	1.5%	337'711	1.5%	77'761	3.1%	1'259	27.3%	107'056	2.9%
Tertiärstufe	1'604'334	0.6%	986'776	0.8%	208'287	1.9%	346'253	1.2%	23'206	6.0%	(298)	56.4%	39'514	4.8%
Ausbildung unbekannt	97'061	3.0%	20'820	6.5%	10'476	8.7%	47'207	4.4%	1'373	27.2%	(141)	67.3%	17'044	7.2%

Fig. 2: Ein Auszug aus der Tabelle *ständige Wohnbevölkerung ab 15 Jahren nach Migrationsstatus: Varianzberechnung mit %CalVarSE*

G Kalibrierungstotale

Auf Ebene Schweiz wurde auf den Totalen folgender Variablen kalibriert (x=Kreuzung):

CHVar1 Geschlecht x Altersklassen fein

CHVar2 Nationalitätengruppen fein

CHVar3 Geschlecht x Altersklassen grob x Nationalitätengruppen grob

CHVar4 Haushaltgrösse

CHVar5 Geschlecht x Altersklassen grob x Zivilstand

CHVar6 Geschlecht x Altersklassen grob x Ausländerkategorie

CHVar7 Region (=Zone, Ausnahmen: In BE: Berner Jura, Stadt Bern, Rest Bern. In ZH: Stadt Zürich, Winterthur, Rest ZH. In GE: Stadt Genf, Rest GE. In VD: Lausanne, Rest VD)

Sofern die Stichprobengrösse es erlaubt ⁷, wurden auf Niveau der bereits erwähnten Zonen sowie für die Grossstädte Genf, Lausanne und Winterthur die folgenden Variablen kalibriert:

KtVar1 Geschlecht

KtVar2 Altersklassen grob

KtVar3 Nationalität

KtVar4 Ausländerkategorie

KtVar5 Nationalitätengruppen grob

KtVar6 Geschlecht x Altersklassen grob

KtVar7 Geschlecht x Nationalität

KtVar8 Geschlecht x Ausländerkategorie

KtVar9 Geschlecht x Nationalitätengruppen grob

KtVar10 Altersklassen grob x Nationalität

Definition Variablen:

- Geschlecht = 1 / 2
- Altersklassen fein = 15-19 / 20-24 / 25-29 / ... / 80-84 / 85+
- Altersklassen grob = 15-24 / 25-44 / 45-64 / 65+
- Nationalitätengruppen fein = Schweiz / Deutschland / Frankreich / Italien / Österreich / Spanien / Portugal / Benelux / nordische Staaten / restliche EU und EFTA (inkl. Liechtenstein) / Türkei / Kosovo / Serbien / Rest Europas / Rest der Welt
- Nationalitätengruppen grob = Schweiz / EU und EFTA / restliches Europa / Rest der Welt
- Nationalität = Schweiz / Ausland
- Ausländerkategorie = Schweiz / C / B und Rest
- Zivilstand ⁸ = 1 / 2 / 3-7
- Haushaltgrösse = 1 / 2 / 3 / 4 / 5 / 6 und mehr

Die Bedeutung der Codierung ist in der Synopsis der Strukturhebung nachzulesen.

⁷ Mindestens 100 Antworten pro Kreuzung

⁸ Aufgrund geringer Anzahl Beobachtungen werden bei den 15-24 Jährigen die Zivilstandesklassen 2 und 3-7 zusammengefasst.