AUSTRIAN JOURNAL OF STATISTICS
Volume 37 (2008), Number 3&4, 319–334

# Varianzschätzung in komplexen Erhebungen

Ralf Münnich
Universität Trier, Deutschland

**Zusammenfassung:** Im Rahmen der Qualitätsberichterstattung europäischer Statistiken sollen im Europäischen Statistischen System neben klassischen Qualitätsangaben auch Aussagen zur Genauigkeit von Statistiken gemacht werden. Neben Nichtstichprobenfehlern spielen im Rahmen der Genauigkeit von Statistiken Stichprobenfehler eine wesentliche Rolle. Im Allgemeinen erfolgt die Quantifizierung dieser Fehler über Angaben zur Varianz der interessierenden Statistik, welche zumeist aus der selben Stichprobe geschätzt werden müssen.

Im Rahmen der vorliegenden Arbeit soll ein Überblick über die aktuell verwendeten Varianzschätzmethoden gegeben werden. Dabei werden deren Vor- und Nachteile diskutiert. An Hand zweier für die Praxis einer Amtlichen Statistik bedeutsamen Beispiele sollen die Verfahren demonstriert werden.

**Abstract:** Within the European Statistical System, national statistical institutes are requested to publish quality reports. One important component of these quality reports is dedicated to accuracy measurement. In addition to non-sampling errors, sampling errors play an important role. The quantification of accuracy measurement is mainly based on variance estimates of the estimators of interest. In general, the latter have also to be estimated from the samples.

This paper presents an overview of variance estimation methods including their advantages and disadvantages. The evaluation of the methodology is supported by two examples from Official Statistics.

**Keywords:** Datenqualität, Resampling-Methoden, Nonresponse.

## 1 Einleitung

Grundlage für die Qualitätsberichterstattung im Europäischen Statistischen System (ESS) ist der 2005 verabschiedete *Code of Practice* (siehe Eurostat, 2005). Mit ihm sind die statistischen Ämter Europas angehalten, neben den Erhebungen bzw. den Auswertungen der Erhebungen auch Metadaten über die jeweilige Erhebung anzugeben und darüber hinaus Aussagen zu bestimmten Qualitätsmerkmalen zu machen. Ein wesentlicher Aspekt, der im Grundsatz 12 festgelegt ist, ist die *Genauigkeit und Zuverlässigkeit* der Statistiken. Die Genauigkeit umfasst dabei sowohl Stichprobenfehler als auch den Komplex von Nichtstichprobenfehlern, zu dem unter anderem Rahmenfehler und Nonresponse gehören.

Streng genommen sollten zu allen *zentralen* Angaben oder Schätzungen der jeweiligen Erhebung solche Genauigkeitsaussagen getroffen und quantifiziert werden.

Eine Quantifizierung von Genauigkeit wird zumeist durch Maße realisiert, die auf der Varianz der Statistiken basieren. Hierzu gehören die Varianz der Schätzung, der Standardfehler, der relative Root Mean Square Error bei (möglicherweise) verzerrten Schätzungen oder auch die Konfidenzintervallüberdeckungsrate. Eine Diskussion solcher Maße kann beispielsweise in Münnich et al. (2003) nachgelesen werden. Da in Anwendungen im Allgemeinen keine zureichenden Informationen über die Grundgesamtheit vorliegen, müssen sowohl die eigentlich interessierenden Populationsparameter als auch die Werte für die Maße zur Beurteilung der Qualität aus der Stichprobe geschätzt werden. Letzteres führt dann unmittelbar zu einer Anwendung geeigneter Varianzschätzmethoden, welche im Folgenden diskutiert werden. Im Falle der Konfidenzintervallüberdeckungsraten müssten sogar Simulationen herangezogen werden, da aus einer Stichprobe lediglich Informationen über ein einziges Konfidenzintervall gewonnen werden können.

Nach einer kurzen Darstellung der relevanten Schätzmethoden werden im nächsten Kapitel sukzessive die verschiedenen Varianzschätzmethoden eingeführt. Hierzu gehören neben den sogenannten direkten Varianzschätzmethoden zwei Gruppen von Verfahren. Das sind zum einen die Linearisierungs- und Approximationsmethoden und zum anderen die Resampling-Verfahren. Abgerundet wird dieser Abschnitt durch Varianzschätzung unter Berücksichtigung von Nonresponse.

Im dritten Kapitel werden die Varianzschätzmethoden an Hand zweier für die Praxis bedeutsamen Beispiele veranschaulicht. Zum einen sind das klassische Tabellenfragestellungen einer Amtlichen Statistik unter Berücksichtigung von Nonresponse. Zum anderen werden nichtlineare Schätzungen analysiert, wie sie etwa bei modernen Armutsmaßen verwendet werden. Neben einer Diskussion von Vor- und Nachteilen der Verfahren interessiert auch, inwieweit Besonderheiten der Verteilung des Untersuchungsmerkmals theoretische Aussagen, wie etwa die Konsistenz von Schätzverfahren, beeinflussen können. Abschließend folgt eine Zusammenfassung nebst Ausblick.

# 2 Varianzschätzmethoden

## 2.1 Schätzung interessierender Größen

In der Amtlichen Statistik interessieren zumeist Totalwerte, wie sie in Tabellen ausgewiesen werden. Hierzu gehört etwa die Schätzung der Anzahl von Erwerbslosen, gegebenenfalls auch untergliedert nach Altersgruppen und Geschlecht. Darüberhinaus werden gelegentlich auch Funktionen von Totalwerten betrachtet, etwa bei Anteilswerten und Raten. Mittelwerte und Anteilswerte lassen sich prinzipiell analog behandeln. Auf eine Darstellung wird hier jedoch verzichtet.

Ausgangspunkt der Betrachtungen ist zunächst die Schätzung eines unbekannten Totalwertes $\tau$ einer interessierenden Variable $Y$ einer endlichen Population vom Umfang $N$. Funktionen von $K$ Totalwerten werden entsprechend durch $\theta = g(\tau_1, \ldots, \tau_K)$ ausgewiesen.

Aus der Grundgesamtheit werden nachfolgend Stichproben vom Umfang $n$ gezogen. Zur Schätzung des Parameters $\tau$ der Grundgesamtheit dient zunächst der Horvitz-

Thompson-Schätzer (HT):

$$\hat{\tau} = \sum_{i=1}^{n} \frac{w_i}{\pi_i} y_i = \sum_{i=1}^{n} w_i d_i y_i \,.$$

Beim klassischen Horvitz-Thompson-Schätzer gilt $w_i \equiv 1$. Die $y_i$ sind die Werte der interessierenden Variablen, die $\pi_i$ die Inklusionswahrscheinlichkeiten 1. Ordnung und $d_i := 1/\pi_i$ die zugehörigen Designgewichte (vgl. beispielsweise Lohr, 1999, oder Särndal et al., 1992). Mit Hilfe der Hilfsvariablenmatrix $\mathbf{X}$ können spezielle Gewichte $w_i := w_i(\mathbf{x}_i)$ beispielsweise durch Regressionen beziehungsweise Kalibrierungen gewonnen werden, die eine Effizienzsteigerung der Schätzung bewirken. Solche Gewichte erfüllen gerade die so genannten Kalibrierungsgleichungen

$$\sum_{i=1}^{n} w_i d_i \mathbf{x}_i = \sum_{i=1}^{N} \mathbf{x}_i \tag{1}$$

für die Hilfsvariablen, welche als Nebenbedingung bei der Optimierung der Gewichte verwendet werden. Durch diese spezielle Gewichtung werden die Verteilungen der Hilfsmerkmale geeignet berücksichtigt. Eine ausführliche Darstellung kann beispielsweise Deville and Särndal (1992) oder Deville (1999) entnommen werden. Mit Hilfe der Designgewichte lassen sich die voranstehenden Schätzer auch auf allgemeinere Stichprobendesigns anwenden. Dazu sei $\mathcal{U}$ die Menge der Indizes von Beobachtungseinheiten aus der Grundgesamtheit und $\mathcal{S}$ die entsprechende Indexmenge der Stichprobe. Im allgemeineren Fall sind nun auch variable Stichprobenumfänge mit abgedeckt, die einen erwarteten Stichprobenumfang von $n$ aufweisen.

Geht man von einem linearen Modell $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ mit den üblichen Annahmen für den Störterm $\boldsymbol{\varepsilon}$ aus, dann kann der verallgemeinerte Regressionsschätzer (generalized regression estimator: GREG) durch

$$\begin{aligned}\hat{\tau}_{\text{GREG}} &= \hat{\tau}_Y + (\tau_X - \hat{\tau}_X)'\hat{\boldsymbol{\beta}} \\ &= \sum_{i\in\mathcal{S}} \underbrace{\Big(1 + \Big(\sum_{k\in\mathcal{U}} \mathbf{x}_k - \sum_{k\in\mathcal{S}} d_k \mathbf{x}_k\Big)' \Big(\sum_{k\in\mathcal{S}} d_k \mathbf{x}_k \mathbf{x}_k\Big)^{-1} \mathbf{x}_i\Big)}_{=w_i} d_i y_i\end{aligned}$$

angegeben werden (vgl. etwa Särndal et al., 1992). Verwendet man statt $y_i$ die Matrix der Hilfsvariablen, dann erhält man nach Umformung die Gültigkeit der Kalibrierungsgleichungen (1).

## 2.2 Grundproblem der Varianzschätzung

Die Varianz des HT-Schätzers kann durch

$$\text{var}_{\text{HT}}(\hat{\tau}) = \sum_{\text{i}\in\mathcal{U}} \pi_{\text{i}}(1-\pi_{\text{i}}) \Big(\frac{\text{y}_{\text{i}}}{\pi_{\text{i}}}\Big)^2 + 2 \sum_{\substack{i,j\in\mathcal{U} \\ i<j}} \sum (\pi_{\text{ij}} - \pi_{\text{i}}\pi_{\text{j}}) \frac{\text{y}_{\text{i}}}{\pi_{\text{i}}} \frac{\text{y}_{\text{j}}}{\pi_{\text{j}}}$$

angegeben werden, wobei die $\pi_{ij}$ die Inklusionswahrscheinlichkeiten 2. Ordnung sind. Diese kann durch

$$\hat{\text{var}}_{\text{HT}}(\hat{\tau}) = \sum_{i\in\mathcal{S}}(1-\pi_i)\Big(\frac{y_i}{\pi_i}\Big)^2 + 2\sum_{\substack{i,j\in\mathcal{S}\\ i<j}}\sum\Big(1-\frac{\pi_i\pi_j}{\pi_{ij}}\Big)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

geschätzt werden (vgl. beispielsweise Hedayat and Sinha, 1991, oder Särndal et al., 1992). Im Falle von Designs mit festem Stichprobenumfang $n$ wird jedoch die Sen-Yates-Grundy-Varianz (SYG) beziehungsweise deren unverzerrte Schätzung

$$\widehat{\text{var}}_{\text{SYG}}(\hat{\tau}) = \sum_{\substack{i,j\in\mathcal{S}\\ i<j}}\sum\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\Big(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\Big)^2$$

vorgezogen, da bei ihr keine negativen Varianzschätzungen resultieren können. Das wesentliche Problem besteht jedoch in der Ermittlung der $\pi_{ij}$, was im Einzelfall nur mit erheblicher Mühe durchgeführt werden kann. Bei großen Stichprobenumfängen muss auch mit Speicherplatzproblemen gerechnet werden, da die Matrix der $\pi_{ij}$ die Größe $n\times n$ aufweist. Im deutschen Mikrozensus würde bei einfach genauer Arithmetik ein Speicherplatz von fast 1 TB benötigt werden.

Insgesamt resultiert daraus die Frage, ob geeignete Approximationen herangezogen werden können, die nur die Inklusionswahrscheinlichkeiten 1. Ordnung verwenden und gleichzeitig eine hohe Approximationsgüte der Varianzschätzung garantieren. Inzwischen existieren zahlreiche derartige Approximationsformeln. Eine sehr geeignete Approximation, welche bei festen Stichprobenumfängen verwendet werden kann, stammt von Deville (1999)

$$\widehat{\text{var}}(\hat{\tau}) = \frac{1}{1-\sum_{i\in\mathcal{S}}a_i^2}\sum_{i\in\mathcal{S}}(1-\pi_i)\Big(\frac{y_i}{\pi_i}-\sum_{j\in\mathcal{S}}a_j\frac{y_j}{\pi_j}\Big)^2 \qquad \text{mit} \quad a_i=\frac{1-\pi_i}{\sum_{j\in\mathcal{S}}(1-\pi_j)}\,. \quad (2)$$

Eine eingehende Übersicht über mögliche weitere Approximationen bietet die Arbeit von Berger and Skinner (2004). Letztendlich ist die Frage der Approximationsgüte dieser Approximationen zumindest bei ausgefalleneren Designs noch nicht hinreichend untersucht.

Im Gegensatz zu den vorgestellten, linearen Schätzern erfordern nichtlineare Schätzverfahren tiefer gehende Betrachtungen, die nachfolgend systematisch vorgestellt werden. In diese Betrachtungen werden auch die Regressions- und Kalibrierungsschätzer mit eingebunden. Im Wesentlichen basieren die Varianzschätzmethoden entweder auf geeigneten Linearisierungsverfahren, also zumeist auf klassischen Taylor-Argumenten, oder auf so genannten Resampling-Methoden. Eine eingehende Übersicht über Varianzschätzverfahren gibt Wolter (2007). Shao and Tu (1995) behandelt Resampling-Verfahren, auch unter Berücksichtigung von Nonresponse. Eine mit Simulationsergebnissen unterstützte Übersicht findet man in den Abschlussberichten des DACSEIS-Projektes (siehe `http://www.dacseis.de`) oder in Münnich (2005).

## 2.3 Linearisierungsmethoden

Ausgangspunkt der Linearisierungsmethoden ist der Satz von Taylor. Für eine nichtlineare Funktion $g$ von $K$ Totalwertschätzern folgt mit Hilfe des Satzes von Taylor

$$\hat{\theta} - \theta = \sum_{k=1}^{K} \frac{\partial g(\hat{\tau}_1, \ldots, \hat{\tau}_K)}{\partial \hat{\tau}_k} (\hat{\tau}_k - \tau_k) + \mathcal{R}(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}) \,,$$

wobei $\mathcal{R}(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau})$ das Taylorsche Restglied 1. Ordnung ist (vgl. Andersson and Nordberg, 1994, S. 396f., oder Wolter, 2007, S. 230ff.). Für eine Varianzschätzung wird dann die Näherung

$$\widehat{\mathrm{var}}(\hat{\theta}) \doteq \mathrm{var}\Big( \sum_{\mathrm{k}=1}^{\mathrm{K}} \frac{\partial \mathrm{g}(\hat{\tau}_1, \ldots, \hat{\tau}_{\mathrm{K}})}{\partial \hat{\tau}_{\mathrm{k}}} (\hat{\tau}_{\mathrm{k}} - \tau_{\mathrm{k}}) \Big) \tag{3}$$

verwendet. Streng genommen beinhaltet die Darstellung bereits die Verwendung großer Stichprobenumfänge, um mögliche Verzerrungen der Schätzfunktion $\hat{\theta}$ vernachlässigen zu können. Dabei wird von asymptotischer Unverzerrtheit ausgegangen, was in der Praxis im Allgemeinen erfüllt ist.

Um eine Verwendung der aufwändigen Kovarianzmatrix, die aus Gleichung (3) resultiert, zu vermeiden, werden vereinfachend auch Richtungsableitungen verwendet (siehe Andersson and Nordberg, 1994). Dies führt dann zu einer vereinfachten Darstellung, bei der Werte $z_i$ gesucht sind, so dass $\hat{\theta} \doteq \sum_{i \in \mathcal{S}} w_i z_i$ ist. Die Werte $z_i$ können nun mit Hilfe der zuvor erwähnten Taylor-Argumente, mit Taylor-basierten Richtungsableitungen oder mit Hilfe von impliziten Schätzgleichungen bestimmt werden. Die verschiedenen Ansätze werden eingehend in Woodruff (1971), Andersson and Nordberg (1994), Binder and Kovačević (1995), Deville (1999), Demnati and Rao (2004), sowie Wolter (2007) vorgestellt. Eine Varianzschätzung der resultierenden linearisierten Summe erfolgt dann unter Verwendung der zuvor besprochenen Techniken, wobei statt der Beobachtungen $y_i$ die *Ersatzwerte* $z_i$ verwendet werden.

Der erste Anwendungsfall ergibt sich bei Verwendung der Residuen des Regressionsschätzers

$$e_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}$$

als Werte $z_i$. Statt der Residuen selber können auch die mit den Design-Gewichten gewichteten Residuen verwendet werden (siehe D'Arrigo and Skinner, 2003), wodurch eine möglicherweise auftretende Unterschätzung der Linearisierung kompensiert werden kann.

Beim Einsatz von impliziten Schätzgleichungen (siehe Binder and Kovačević, 1995) werden ausgehend von der Schätzgleichung $u(y, \theta) = 0$ die so genannten Einflusswerte $u_i^*$ mittels

$$u^*(y) = - \left( \frac{\partial \mathrm{E}(u(y, \theta))}{\partial \theta} \bigg|_{\theta = \theta_0} \right)^{-1} u(y, \theta_0)$$

an der Stelle $\theta_0$ ermittelt. Diese werden dann wiederum für $z_i$ zur Varianzschätzung herangezogen. Dieser Ansatz wird insbesondere bei nichtlinearen Schätzern häufig verwendet.

Bei der Schätzung von Quantilen erhält man für ein allgemeines $p$-Quantil einer Verteilung mit Dichte $f$ ausgehend von $u(y) = \mathbb{1}(y \leq \theta) - p$, wobei $\mathbb{1}(\cdot)$ die Indikatorfunktion ist, schließlich (vgl. beispielsweise Binder and Kovačević, 1995, oder Deville, 1999)

$$u^*(y) = -\frac{1}{f(\theta)}\big(\mathbb{1}(y \leq \theta) - p\big)\,.$$

Damit lässt sich eine Varianzschätzung für eine verallgemeinerte *At-risk-of-poverty ratio* (ARPR) zum $\alpha$-Anteil des $p$-Quantils einer Einkommensverteilung mit geschätzter Verteilungsfunktion $\hat{F}$

$$\widehat{\text{ARPR}}_{\alpha,p} = \hat{F}(\alpha\hat{y}_p)\,,$$

einem bedeutsamen Maß im Rahmen der Laeken-Indikatoren (vgl. Dennis and Guio, 2004), herleiten. Man erhält schließlich

$$u_i^* = \frac{1}{N}\big(\mathbb{1}(y_i \leq \alpha y_p) - \text{ARPR}_{\alpha,p}\big)$$

als Einflusswerte (Berger and Skinner, 2003). Deville (1999) weist hier darauf hin, dass die ARPR ebenso geschätzt wird, und möglicherweise Verzerrung durch die Nichtberücksichtigung der stochastischen Abhängigkeiten auftreten können. Den ebenso zu den Laeken-Indikatoren gehörenden GINI-Koeffizient ermittelt man auf ähnliche Weise.

In der Amtlichen Statistik spielen, wie bereits erwähnt, Funktionen von Totalwerten eine wichtige Rolle. In diesen Fällen werden häufig vereinfachend Taylor-Richtungsableitungen verwendet, die bei der so genannte Woodruff-Linearisierung eingehen (Woodruff, 1971). Hierbei werden die Werte $z_i$ direkt durch

$$z_i = \sum_{k=1}^{K} \frac{\partial}{\partial\tau_i} g(\tau_1, \ldots, \tau_k) y_{k,i} \tag{4}$$

ermittelt, wobei $g$ wiederum die zuvor eingeführte Funktion von $K$ Totalwerten ist und $y_{k,i}$ die $k$-te Komponente der $i$-ten Beobachtung. Eine eingehende Beschreibung dieser Methode inklusive einiger Anwendungen kann in Andersson and Nordberg (1994) nachgelesen werden.

Ein besonders interessantes Beispiel für die Anwendung der Woodruff-Linearisierung ist wiederum durch die Laeken-Indikatoren in Form der *Quintile-share ratio* (QSR) gegeben, welche durch

$$\text{QSR} = \frac{\sum\limits_{i\in\mathcal{S}} y_i \mathbb{1}(y_i > y_{0.8})}{\sum\limits_{i\in\mathcal{S}} y_i \mathbb{1}(y_i \leq y_{0.2})} \frac{\sum\limits_{i\in\mathcal{S}} \mathbb{1}(y_i \leq y_{0.2})}{\sum\limits_{i\in\mathcal{S}} \mathbb{1}(y_i > y_{0.8})}$$

definiert ist. Dabei ist $y_p$ das $p$-Quantil der Verteilung der interessierenden Variablen. Sie gibt das Verhältnis der Durchschnittseinkommen der *oberen* 20% (R: rich) durch die *unteren* 20% (P: poor) an und ist damit besonders sensitiv bezüglich Einkommens-Ausreißern, die alle in der Zählergröße enthalten sind. Bei genauerer Betrachtung erkennt man, dass die QSR als Funktion von vier Totalwerten dargestellt wird:

$$\widehat{\text{QSR}} = \frac{\hat{\mu}_\text{R}}{\hat{\mu}_\text{P}} = \frac{\hat{\tau}_1}{\hat{\tau}_2} \Big/ \frac{\hat{\tau}_3}{\hat{\tau}_4}\,.$$

Für die konkrete Umsetzung der Beziehung (4) ermittelt man zunächst die linearisierten Variablen

$$
\begin{aligned}
u_{1i} &= y_i - \left((y_i - y_{0.8})\mathbb{1}(y_i \leq \hat{y}_{0.8}) + 0.8y_{0.8}\right), \\
u_{3i} &= (y_i - y_{0.2})\mathbb{1}(y_i \leq \hat{y}_{0.2}) + 0.2y_{0.2}\,, \\
u_{5i} &= \left(u_{1i} - \frac{\hat{\tau}_1}{\hat{\tau}_2}u_{2i}\right)\frac{1}{\hat{N}0.2} = \hat{\mu}_{\mathrm{R}}\,, \\
u_{6i} &= \left(u_{3i} - \frac{\hat{\tau}_3}{\hat{\tau}_4}u_{4i}\right)\frac{1}{\hat{N}0.2} = \hat{\mu}_{\mathrm{P}}\,,
\end{aligned}
$$

wobei $\hat{\tau}_2 = u_{2i} = \hat{\tau}_4 = u_{4i} = \hat{N}0.2$ ist, und erhält schließlich

$$
z_i = (u_{5i} - \widehat{\mathrm{QSR}}u_{6i})\frac{\hat{\tau}_4}{\hat{\tau}_3}\,.
$$

Ein wesentlicher Vorteil der Linearisierungsmethoden ist die Einfachheit und Effizienz bei der Anwendung, nachdem man einmalig den Aufwand der Herleitung der linearisierten Terme aufgebracht hat. Varianzschätzungen können nun auch relativ einfach für kompliziertere Designs vorgenommen werden. Allerdings bleibt stets die Frage der Approximationsgüte zu beachten. Diese kann allgemein kaum beantwortet werden, da die Verteilungen der interessierenden Merkmale, und hier insbesondere Symmetrieeigenschaften und Ausreißer, eine wesentliche Bedeutung haben.

## 2.4 Varianzschätzung auf Basis von Resampling-Methoden

Resampling-Verfahren beruhen auf der Idee aus einer gezogenen Stichprobe wiederholt Substichproben zu ziehen mit deren Hilfe die zugrunde liegende Schätzverteilung abgebildet werden soll. Die einzelnen Verfahren unterscheiden sich in der Art der Substichprobenziehung durch die auch Aufwand und Effizienz bestimmt sind. Letztendlich soll die Variation der Schätzwerte, die auf den Substichproben ermittelt werden, Aufschluss über die Varianz der eigentlich interessierenden Schätzfunktion geben. Inzwischen sind, über Stichprobenanwendungen hinaus, die Methoden des delete-1-Jackknife und des Bootstrap sehr weit verbreitet. Eine aktuelle Diskussion dieser Methoden kann in Davison and Sardy (2007) und der darin aufgeführten Literatur nachgelesen werden. Eine sehr ausführliche Darstellung vor allem der Jackknife- und Bootstrap-Verfahren kann Shao and Tu (1995) entnommen werden.

Bei geschichteten Zufallsstichproben werden häufig Balanced Half Samples verwendet. Es wird davon ausgegangen, dass in jeder Schicht zwei Elemente enthalten sind. Die Resamples werden nun durch Wahl jeweils eines Elementes der $H$ Schichten gewonnen, wodurch insgesamt $2^H$ verschiedene Replikationen gebildet werden können. Um den enormen Aufwand der Auswahl aller gezogener Teilstichproben zu vermeiden, werden so genannte ausbalancierte Stichproben verwendet, welche mit Hilfe von Hadamard-Matrizen gewonnen werden. Dadurch lassen sich Varianzen mit oft sehr wenigen Replikationen schätzen, welche durch die nächste durch 4 teilbare Zahl oberhalb von der Anzahl der Schichten $H$ angenommen wird. Eine geringe Anzahl an Schichten sowie die Tatsache, dass in den einzelnen Schichten selten wirklich nur zwei Elemente enthalten

sind, machen Anpassungen des Verfahrens notwendig. Eine Möglichkeit ist die zufällige Einteilung der Stichprobenelemente jeder Schicht in zwei Gruppen, welche wie die zuvor ausgewiesenen Elemente behandelt werden. Diese Methode ist in den meisten modernen Survey-Softwarepaketen enthalten und eignet sich vor allem bei einer größeren Anzahl an Schichten.

Beim delete-1-Jackknife wird jeweils ein Element in jeder Replikation entfernt. Somit ergeben sich durch sukzessives Auslassen aller Stichprobenelemente insgesamt $n$ Resamples, was bei großen Erhebungen durchaus aufwändig sein kann. Die Varianz der Schätzfunktion ergibt sich anschließend als Varianz der $n$ Schätzwerte auf den Resamples, wobei man in der durch das Auslassen betroffenen Schicht jeweils eine Korrektur der Gewichtung vornehmen muss. Daraus ergibt sich für die Varianzschätzung im Falle einer geschichteten Zufallsstichprobe

$$\widehat{\mathrm{var}}_{\mathrm{d1JK,strat}}(\hat{\tau}) = \sum_{h=1}^{H} \frac{(1-f_h)(n_h-1)}{n_h} \sum_{i=1}^{n_h} \left(\hat{\tau}_{h,-i} - \overline{\hat{\tau}}_h\right)^2 .$$

Der Index $-i$ kennzeichnet, dass das $i$-te Element herausgelassen wurde. $f_h$ ist der Auswahlsatz und $\overline{\hat{\tau}}_h$ der Mittelwert der delete-1-Jackkknife-Totalschätzwerte in der $h$-ten Schicht. Wesentliche Voraussetzung zur Anwendung des delete-1-Jackknife sind glatte Statistiken, also insbesondere stetig differenzierbare Statistiken. Damit können so wichtige Statistiken wie der Median nicht korrekt behandelt werden. Eine Erweiterung auf den delete-$d$-Jackknife, bei dem jeweils $d$ Elemente entnommen werden, löst jedoch dieses Problem. Da jedoch nicht alle Kombinationen von $d$ Elementen entnommen werden können, da dies in der Praxis viel zu aufwändig wäre, müssen sowohl $d$ als auch die Anzahl der Replikationen geeignet gewählt werden. Eine geeignete Wahl von $d$ und der Anzahl der Replikationen hängt aber sehr von der interessierenden Statistik ab.

Bei der Bootstrap-Methode werden aus der Stichprobe Substichproben vom Umfang $n$ gezogen, was nur im Modell mit Zurücklegen möglich ist. In der Praxis wird oft der so genannte Monte-Carlo-Bootstrap (Boot, MC) verwendet, bei dem eine angemessene Anzahl von $B$ Stichproben zufällig gezogen wird, aus der schließlich die Bootstrap-Schätzwerte $\hat{\tau}_i^*$ der eigentlich interessierenden Statistik ermittelt werden. Als Varianzschätzung verwendet man

$$\widehat{\mathrm{var}}_{\mathrm{Boot,MC}}(\hat{\tau}) = \frac{1}{B-1} \sum_{i=1}^{B} \left(\hat{\tau}_i^* - \frac{1}{B}\sum_{j=1}^{B} \hat{\tau}_j^*\right)^2 ,$$

was der empirischen Varianz der Bootstrap-Schätzwerte entspricht. Für Varianzschätzungen reichen in vielen Fällen bereits $B = 100$ Resamples (vgl. `http://rpm.dacseis.de`). Mit Hilfe des Monte-Carlo-Bootstrap-Verfahrens lassen sich auch Konfidenzintervalle von Schätzstatistiken angenähert berechnen, wobei im Allgemeinen aber eine wesentlich höhere Anzahl an Replikationen benötigt wird. Ein Vorteil der Bootstrap-Methode liegt in der Einfachheit der Anwendung und in ihrer relativ geringen Anzahl an benötigten Replikationen. Wenn man Stichproben im Modell ohne Zurücklegen bei nicht kleinen Auswahlsätzen hat, was in Anwendungen häufig vorkommt, sollten allerdings besondere Korrekturverfahren verwendet werden.

Die hier vorgestellten Resampling-Methoden sind nicht auf die Schätzung von Totalwerten eingeschränkt. Vielmehr lassen sie sich ohne Änderungen direkt auf komplexe-

re Statistiken übertragen. Ein wesentlicher Vorteil der Resampling-Verfahren gegenüber den Linearisierungsverfahren ist daher die einfache Anwendbarkeit, da keine besonderen Herleitungen mit Hilfe von Taylor-Verfahren benötigt werden. Allerdings ist der Rechenaufwand im Allgemeinen erheblich höher, was bei großen Stichprobenumfängen auch moderne Rechenanlagen überfordern kann. Ebenso lassen sich Resampling-Verfahren im Gegensatz zu Linearisierungsverfahren nicht so einfach auf sehr unterschiedliche komplexe Designs übertragen.

Insgesamt muss festgehalten werden, dass kaum generelle Aussagen über den optimalen Einsatz eines Resampling-Verfahrens gemacht werden können. Vielmehr muss oft im Einzelfall entschieden werden. Bei nichtlinearen Statistiken, insbesondere wenn keine direkten oder approximativen Varianzschätzer vorhanden sind, sind die Resampling-Methoden vorzuziehen. Bei sehr komplexen Stichprobendesigns erweisen sie sich aber im Allgemeinen als problematisch.

# 3 Varianzschätzung bei fehlenden Werten

In den letzten Jahren wurde die korrekte Behandlung von fehlenden Werten (missing data) viel diskutiert. Im Wesentlichen stehen hier drei grundlegende Ideen zur Verfügung, die jedoch auch von der Struktur der fehlenden Werte abhängen:

**Gewichtungsmethoden:** Der auftretende Nonresponse wird durch Responsemodelle dargestellt, welche auch einfach homogene Klassen sein können, die so genannten *Response Homogeneity Groups* (vgl. Särndal and Lundström, 2005). Als Kompensationsgewichte werden die umgekehrten Ausfallraten verwendet.

**Einfache Imputation:** Die fehlenden Werte werden mit Hilfe statistischer Methoden einmalig, entweder deterministisch oder stochastisch, ergänzt. Hierzu gehören die in der Amtlichen Statistik beliebten Hotdeck-Methoden, bei denen geeignete vorhandene Daten aus dem Datensatz als Schätzwerte für die fehlenden Werte verwendet werden.

**Multiple Imputation:** Die fehlenden Werte werden mit Hilfe von statistischen Modellen mehrfach ergänzt. Ziel dieser auf Bayesianischen Methoden basierenden Technik ist eine korrekte Ermittlung der Varianz der Schätzung, in die sowohl die geschätzte Variabilität der einzelnen imputierten Datensätze als auch die Variabilität der Schätzungen zwischen den einzelnen Datensätzen geeignet berücksichtigt wird.

Im Falle von Gewichtungsmethoden können für die Varianzschätzung die Logit-Gewichte direkt als zusätzliche Gewichtungsfaktoren verwendet werden. Streng genommen sollte dabei jedoch beachtet werden, dass eine mögliche stochastische Abhängigkeit zwischen Logit- und Designgewichten existieren kann, deren Nichtberücksichtigung zu verzerrten Varianzschätzungen führen kann. Bei der Verwendung von Kalibrierungsverfahren muss beachtet werden, dass durch das Auftreten von Nonresponse die Designgewichte nicht mehr korrekt sind, da die tatsächlichen Inklusionswahrscheinlichkeiten nicht erreicht wurden. Eine Verwendung dieser nicht korrigierten Gewichte kann zu verzerrten Schätzungen führen. Die Gewichte können jedoch relativ einfach korrigiert werden, indem die Ausfallwahrscheinlichkeiten beziehungsweise der tatsächliche Response zur Korrektur der Designgewichte, etwa durch Quotientenbildung, herangezogen wird. Dies

wird sinnvollerweise nicht insgesamt sondern für die einzelnen Response Homogeneity Groups gemacht. Bei beiden Methoden muss vorausgesetzt werden, dass die Hilfsinformationen der Respondenten, die für die Modellierung herangezogen werden, keine fehlende Werte enthalten.

In der Amtlichen Statistik ist die Verwendung von einfachen Imputationsverfahren verbreitet. Bei der Verwendung von solchen einfachen Imputationen muss jedoch beachtet werden, dass ein durch Imputation vervollständigter Datensatz für die Schätzungen herangezogen wird und damit die durch den Nonresponse erzeugte Stochastik oft nicht berücksichtigt wird. Dadurch erhält man je nach Responserate zum Teil deutlich unterschätzte Varianzen, wie im nächsten Abschnitt zu sehen ist. Mit Hilfe von Resampling-Verfahren lässt sich diese Unterschätzung wieder kompensieren. Idee der Verfahren ist die Verwendung einer eigenen Imputation in jeder Resampling-Stichprobe, die aber der einfachen Imputation der eigentlichen Stichprobe entsprechen muss und damit von dieser abhängig ist. Diese Ansätze werden für den Jackknife-Schätzer in Rao and Shao (1992) und Shao and Tu (1995) sowie für den Bootstrap in Shao and Sitter (1996) beschrieben. Dabei ist jedoch zu beachten, dass in dieser Form der Rechenaufwand extrem hoch ist, da in jedem Resampling-Schritt eine neue Imputation ermittelt werden muss. Gerade beim delete-1-Jackknife-Schätzer eignen sich jedoch spezielle direkte Methoden, bei denen das Imputationsmodell und der Schätzer bereits im Jackknife berücksichtigt werden. Dadurch resultieren zum Teil sehr kompakte und effiziente Varianzschätzformeln (vgl. etwa Shao and Tu, 1995). Der Nachteil dieser Methode liegt in der Erfordernis für jeden Schätzer und jedes Imputationsmodell eine eigene Formel entwickeln zu müssen.

Bei der multiplen Imputation (MI) werden $m$ vollständige Datensätze mit Hilfe von Ausprägungen eines statistischen Modells erzeugt (siehe Rubin, 1978 und Rubin, 1987). Als Punktschätzung bezüglich der multiplen Imputation wird das arithmetische Mittel der $m$ Schätzungen auf den jeweiligen imputierten Datensätzen verwendet. Als Varianzschätzung wird

$$\widehat{\mathrm{var}}_{\mathrm{MI}}\big(\theta\big) = \frac{1}{m}\sum_{j=1}^{m}\widehat{\mathrm{var}}\big(\hat{\theta}^{(j)}\big) + \Big(1+\frac{1}{m}\Big)\frac{1}{m-1}\sum_{j=1}^{m}\big(\hat{\theta}^{(j)}-\hat{\theta}_{\mathrm{MI}}\big)^2$$

verwendet, wobei $\hat{\theta}^{(j)}$ der Schätzwert zur $j$-ten Imputation ist und $\hat{\theta}_{\mathrm{MI}} = \sum_j \hat{\theta}^{(j)}/m$ der MI-Schätzwert. Man erhält also eine Art Binnen- und Zwischenimputationsvarianz, welche einen kleinen Korrekturfaktor in Bezug auf die Freiheitsgrade der imputierten Datensätze enthält. Eine solche Modellierung hat den wesentlichen Vorteil, dass sie nicht von einem speziellen Schätzverfahren abhängt und damit sehr allgemein angewendet werden kann. Allerdings darf keine beliebige Imputationsregel verwendet werden, sondern nur eine solche, die im Rubinschen Sinn *proper* ist. Gerade bei kategorialen Modellen kann das durchaus problematisch sein (vgl. beispielsweise Münnich and Rässler, 2005).

Als neuere Alternative zur Bayesianischen multiplen Imputation wurde in Bjørnstad (2004) eine nicht-Bayesianische multiple Imputation eingeführt. Der wesentliche Unterschied besteht in der Varianzformel, die eine weitere Konstante $k$ enthält, mit deren Hilfe eine Varianzunterschätzung kompensiert werden soll:

$$\widehat{\mathrm{var}}_{\mathrm{MI},k}\big(\hat{\theta}\big) = \frac{1}{m}\sum_{j=1}^{m}\widehat{\mathrm{var}}\big(\hat{\theta}^{(j)}\big) + \Big(k+\frac{1}{m}\Big)\frac{1}{m-1}\sum_{j=1}^{m}\big(\hat{\theta}^{(j)}-\hat{\theta}_{\mathrm{MI}}\big)^2 .$$

Auf Basis dieser multiplen Imputation können prinzipiell auch Hotdeck-Verfahren verwendet werden. Allerdings muss die Konstante $k$ für jeden Schätzer und jede Imputationsmethode gesondert ermittelt werden. Eine eingehende Übersicht inklusive Diskussion hierüber findet der Leser in Bjørnstad (2007).

Über den geeigneten Einsatz der vorgestellten Methoden wurde eine zum Teil kontroverse Diskussion in der Literatur geführt. Tatsächlich existieren für alle Modelle Beispiele, welche die Methoden mehr oder weniger gut abschneiden lassen. Eingehendere aktuelle Übersichten können beispielsweise Rässler and Riphahn (2006) beziehungsweise Davison and Sardy (2007) entnommen werden.

# 4 Anwendungsbeispiele

Exemplarisch sollen zwei Fälle herangezogen werden. Zum einen handelt es sich um eine klassische Tabellenfragestellung der Amtlichen Statistik zur Schätzung von Erwerbslosen. In Deutschland wird dies durch den Mikrozensus bewerkstelligt. Zum anderen werden aktuell immer mehr auch nichtlineare Statistiken verwendet. Ein besonders bedeutsames Beispiel sind die bereits erwähnten Laeken-Indikatoren.

## 4.1 Schätzung von Erwerbslosen im Mikrozensus

Im Rahmen der Schätzung von Erwerbslosen interessieren zwei unterschiedliche Teilpopulationen, erwerbslose Frauen im Alter zwischen 25 und 44 sowie im Alter ab 65. Aus einer synthetischen, aber realitätsnahen Grundgesamtheit für das Bundesland Saarland mit $N = 1.057.915$ Einwohnern, die im Rahmen des Forschungsprojektes DACSEIS generiert wurde, wurden 10.000 Stichproben nach dem Prinzip des Mikrozensus mit einem Stichprobenumfang von etwa 10.000 Personen gezogen (für eine Darstellung der Mikrozensus-Simulationen sei der Leser auf Münnich and Rässler, 2005 beziehungsweise Wiegert and Münnich (2004) und die dort zitierte Literatur verwiesen). In dieser Grundgesamtheit sind $\tau_1 = 7.248$ bzw. $\tau_2 = 128$ erwerbslose Frauen in den zuvor erwähnten Altersklassen. Bei den Schätzungen wurden 5%, 10%, 25% und 40% Nonresponse in der Untersuchungsvariablen berücksichtigt. In Abbildung 1 sind die Punkt- (links) und Varianzschätzverteilung (rechts) eines Kalibrierungsschätzers zur Schätzung der ersten Teilpopulation in Abhängigkeit der vier verschiedenen Nonresponseraten als Boxplots dargestellt. Die gestrichelte Linie kennzeichnet jeweils den zu schätzenden Referenzwert, die durchgezogene den mittleren Schätzwert.

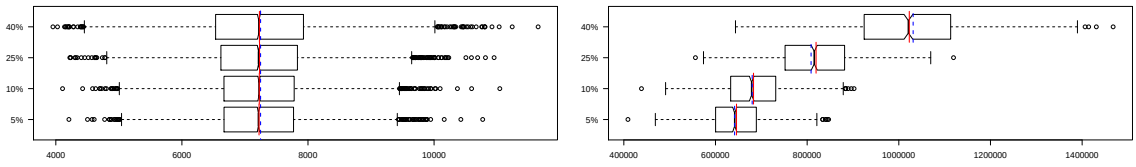


Abbildung 1: Erwerbslose Frauen im Alter von 25 – 44.

Man erkennt, dass die Punkt- und Varianzschätzung hier auch unter Berücksichtigung von Nonresponse sehr gut funktionieren. Ähnliche Graphiken erhält man bei Schätzungen

mit dem Bootstrap-Verfahren nach Shao and Sitter (1996) sowie bei Anwendung der multiplen Imputation.

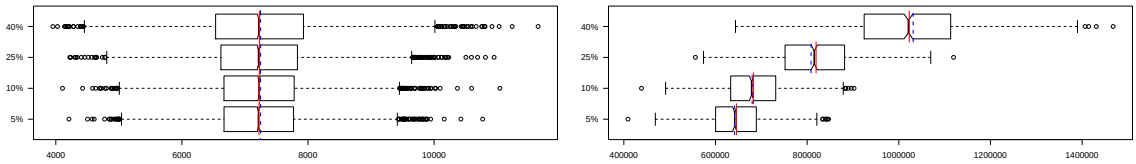


Abbildung 2: Erwerbslose Frauen im Alter ab 65.

Auch bezüglich der zweiten Teilpopulation erhält man im Durchschnitt sehr gute Schätzungen, wie in Abbildung 2 zu erkennen ist. Tatsächlich ergeben jedoch Ausreißerverteilungen, die auf Grund der wenigen Beobachtungen und des Klumpendesigns mit Klumpengrößen von etwa 15 – 20 Personen resultieren. Trotz des relativ großen Stichprobenumfangs und der sehr gut funktionierenden Asymptotik in Bezug auf die durchschnittlichen Schätzwerte erhält man keineswegs approximative Normalverteilungen für die Schätzverteilungen, wie Abbildung 3 zu entnehmen ist. Derartige Beispiele ergeben sich relativ schnell bei Varianzschätzungen, wenn auffällig abweichende Beobachtungen auftreten. Weitere Beispiele sowie Varianzschätzungen bei Auftreten von Ausreißern können in Hulliger and Münnich (2006) nachgelesen werden.

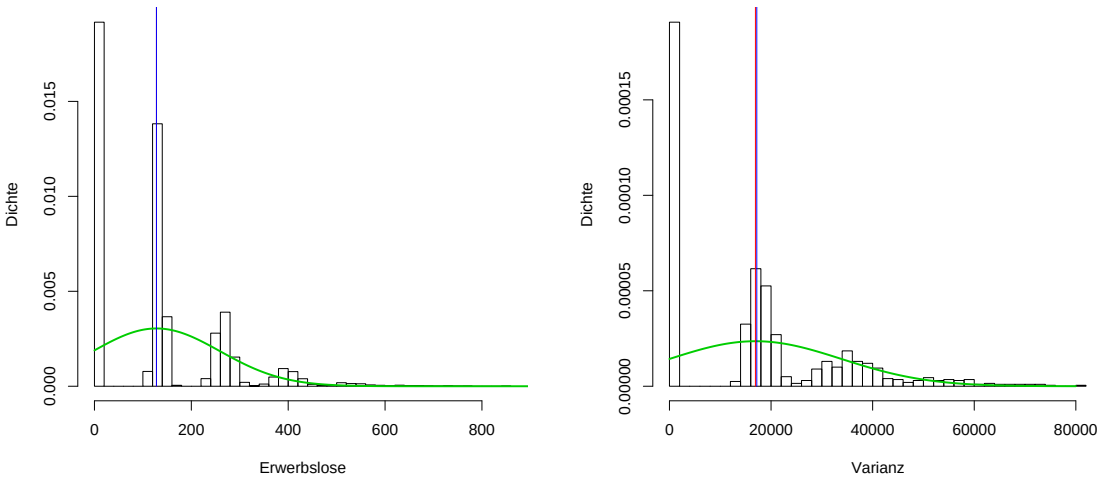


Abbildung 3: Schätzverteilungen erwerbsloser Frauen im Alter ab 65 bei 25% Nonresponse.

## 4.2 Armutsmessung am Beispiel von D-SILC

Im Rahmen einer synthetischen Einkommensverteilung des deutschen Survey on Income and Living Conditions (D-SILC) mit etwa 4 Millionen Beobachtungen wurden in $H = 18$ Schichten 1.000 geschichtete Zufallsstichproben vom Umfang $n_h = (6, 48, 110, 14, 8, 44,$

Tabelle 1: Varianz und mittlere geschätzte Varianz ausgewählter Varianzschätzer zur Schätzung der ARPR und der QSR.

| Schätzer | $\mathrm{var}(\widehat{\mathrm{ARPR}})$ | $\mathrm{E}\widehat{\mathrm{var}}(\widehat{\mathrm{ARPR}})$ | Schätzer | $\mathrm{var}(\widehat{\mathrm{QSR}})$ | $\mathrm{E}\widehat{\mathrm{var}}(\widehat{\mathrm{QSR}})$ |
|---|---|---|---|---|---|
| ARPR.linearized | 0.0002313 | 0.0002108 | QSR.linearized | 1.2058 | 1.6729 |
| ARPR.bhs | 0.0002132 | 0.0002398 | QSR.bhs | 1.2433 | 1.1941 |
| ARPR.d1JK | 0.0002299 | 0.0016697 | QSR.d1JK | 1.2064 | 2.0982 |
| ARPR.boot99 | 0.0002098 | 0.0002253 | QSR.boot99 | 1.2114 | 1.0848 |
| ARPR.boot499 | 0.0002084 | 0.0002264 | QSR.boot499 | 1.2047 | 1.0794 |

$107, 3, 4, 51, 23, 21, 149, 14, 10, 82, 31, 5)$ gezogen, was Auswahlsätzen von 0.76%-10.5% entspricht. Zu schätzen sind die Werte der *At-risk-of-poverty ratio* ARPR $= 0.1966$ beziehungsweise der *Quintile-share ratio* QSR $= 8.621$. Als Punktschätzwerte resultierten $\widehat{\mathrm{ARPR}} = 0.1905$ beziehungsweise $\widehat{\mathrm{QSR}} = 8.636$ den linearisierten Schätzer mit jeweils nur sehr geringen Abweichungen zu den Resampling-Verfahren.

Man erkennt, dass in diesen Beispielen, in denen nicht glatte Statistiken verwendet werden, der Jackknife-Schätzer (d1JK) ungeeignet ist. Die Approximationsgüte der linearisierten Varianzschätzer (linearized), wie sie zuvor eingeführt wurden, sind nicht optimal. Sie weisen aber als Varianzschätzer geringe Streuungen auf. Der Bootstrap-Varianzschätzer (boot) zeigt relativ gute Schätzungen. Eine Erhöhung der Bootstrap-Resamples erweist sich hier jedoch als unnötig. Tatsächlich kommt man bei Varianzschätzungen oft mit 100 Bootstrap-Resamples aus. Der Balanced Half Sample-Varianzschätzer ist hier mit dem Bootstrap-Varianzschätzer vergleichbar, er weist jedoch auf Grund der geringen Anzahl an Replikationen eine relativ hohe Varianz der Varianz auf, die hier nicht explizit dargestellt wurde.

# 5 Zusammenfassung und Ausblick

Bei linearen Statistiken, wie sie beispielsweise bei Tabellenfragestellungen auftreten, eignen sich zumeist alle vorgestellten Methoden. Insofern verlagert sich bei einfacheren Designs die Frage der Auswahl eines geeigneten Verfahren ein wenig hin zur rechnerischen Effizienz der Verfahren. Bei komplizierteren Designs, wenn sich die Matrizen der Inklusionswahrscheinlichkeiten 2. Ordnung nur mit großem Aufwand berechnen lassen oder im Falle sehr großer Erhebungen einen zu großen Speicherbedarf benötigen, ist die Genauigkeit der approximativen Verfahren ein wesentlicher Parameter.

Weniger ausgereift sind Empfehlungen bei nichtlinearen Schätzungen, wie sie am Beispiel der Armutsmessung angedeutet wurden, oder bei Auftreten von Nonresponse und dessen Kompensation. Hier spielen weitere Gründe bei der Beurteilung der Auswahl einer geeigneten Varianzschätz-Methode eine Rolle, die sich kaum in konkreten Standardempfehlungen darstellen lassen. In jedem Falle sollten bei seltenen Ereignissen, bei sehr schiefen Verteilungen sowie bei Existenz von Ausreißern eine besondere Sorgfalt der Auswahl der Verfahren eingeplant werden. Hier scheinen in vielen Fragestellungen Resampling-Verfahren im Vorteil zu sein, solange kein allzu komplexes Stichprobende-

sign vorliegt. Die Diskussionen um ein geeignetes Verfahren beim Auftreten von Non-response werden sicher noch eine Weile andauern. Die Abhängigkeit von der konkreten Datenlage und möglichen verfügbaren Hilfsinformationen wird die Wahl einer geeigneten Methode beeinflussen.

Neuerdings wird die Qualität, insbesondere bei Vergleichen von multinationalen Erhebungen, mit Hilfe von Design-Effekte gemessen, welche als Verhältnis der Varianz der Schätzung bei gegebenem Design zur Varianz bei äquivalenter Schätzung auf Basis der uneingeschränkten Zufallsstichprobe berechnet werden (vgl. beispielsweise Kish, 1995 oder Gabler, Häder, and Lynn, 2003). Hier spielen Varianzschätzmethoden eine besondere Rolle, da Schätzungen sowohl in Zähler als auch im Nenner vorgenommen werden müssen.

## Danksagung

Ein Teil der Arbeiten entstand im Rahmen des EU-Forschungsprojektes DACSEIS (`www.dacseis.de`). Für zahlreiche interessante Diskussionen möchte ich mich bei allen Beteiligten des Projektes bedanken. Herrn Dr. Clæs Andersson, Statistik Schweden, danke ich für einen wertvollen Tipp bei der Herleitung der Formel für die Quintile Share Ratio. Desweiteren gebührt zwei anonymen Referees mein Dank für zahlreiche wertvolle Hinweise.

# Literatur

Andersson, C., and Nordberg, L. (1994). A method for variance estimation of non-linear functions of totals in surveys - theory and software implementation. *Journal of Official Statistics*, *10*, 395-405.

Berger, Y. G., and Skinner, C. J. (2003). Variance estimation for a low income proportion. *Journal of the Royal Statistical Society, Series C*, *52*, 457-468.

Berger, Y. G., and Skinner, C. J. (2004). *Variance estimation for unequal probability designs.* (DACSEIS deliverable D6.1, `http://www.dacseis.de`)

Binder, D. A., and Kovačević, M. S. (1995). Estimating some measures of income inequality from survey data: An application of the estimating equation approach. *Survey Methodology*, *21*, 137-145.

Bjørnstad, J. (2004). Non-Bayesian Multiple Imputation. In *Proceedings of the Q2004 conference.*

Bjørnstad, J. (2007). Non-Bayesian multiple imputation (with discussion). *Journal of Official Statistics*, *23*, 433-491.

D'Arrigo, J., and Skinner, C. J. (2003). *Variance estimation for estimators subject to raking adjustment.* (DACSEIS deliverable D8.1, `http://www.dacseis.de`)

Davison, A. C., and Sardy, S. (2007). Resampling variance estimation in surveys with missing data. *Journal of Official Statistics*, *23*, 371-386.

Demnati, A., and Rao, J. N. K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, *30*, 17-26.

Dennis, I., and Guio, A. (2004). *Poverty and social exclusion in the EU.* (Statistics in Focus 16, Eurostat, Luxembourg, catalogue number: KS-NK-04-016-EN-N. http://epp.eurostat.cec.eu.int/portal/page?_pageid=1073,1135281,1073_1135295&_dad=portal&_schema=PORTAL&p_product_code=KS-NK-04-016)

Deville, J. C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, *25*, 193-203.

Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, 376-382.

Eurostat. (2005). *European Statistics Code of Practice: For the National and Community Statistical Authorities.* (http://epp.eurostat.cec.eu.int/portal/page?_pageid=2273,1,2273_47141302&_dad=portal&_schema=PORTAL)

Gabler, S., Häder, S., and Lynn, P. (2003). Refining the concept and measurement of design effects. In *Bulletin of the International Statistical Institute 54th Session* (Vol. LX, p. 371-372).

Hedayat, A. S., and Sinha, B. K. (1991). *Design and Inference in Finite Population Sampling*. New York: John Wiley & Sons.

Hulliger, B., and Münnich, R. (2006). Variance estimation for complex surveys in the presence of outliers. In *Proceedings of the Survey Research Methods Section* (p. 3153-3161). American Statistical Association. (http://www.amstat.org/Sections/Srms/Proceedings/y2006/Files/JSM2006-000530.pdf)

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, *11*, 55-77.

Lohr, S. (1999). *Sampling: Design and Analysis*. CA: Pacific Grove: Duxbury Press.

Münnich, R. (2005). *Datenqualität in komplexen Stichprobenerhebungen.* (unveröffentlichte Habilitationsschrift, Universität Tübingen)

Münnich, R., Bihler, W., Bjørnstad, J., Zhang, L., Davison, A., Haslinger, A., et al. (2003). *Data Quality in Complex Surveys.* (DACSEIS deliverable D1.1, http://www.dacseis.de)

Münnich, R., and Rässler, S. (2005). PRIMA: A new multiple imputation procedure for binary variables. *Journal of Official Statistics*, *21*, 325-341.

Rao, J. N. K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, *79*, 811-8220.

Rässler, A., and Riphahn, R. T. (2006). Survey item nonresponse and its treatment. *Allgemeines Statistisches Archiv*, *90*, 217-232.

Rubin, D. B. (1978). Multiple Imputation in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse. In *Proceedings of the Survey Research Methods Sections of the American Statistical Association* (p. 20-40).

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Shao, J., and Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, *91*, 1278-1288.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer.

Särndal, C. E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Chichester: John Wiley & Sons.

Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Wiegert, R., and Münnich, R. (2004). German Register Data for Regression Estimation in Survey Sampling – A Study on the German Microcensus Respecting for Data Protection. *Jahrbücher für Nationalökonomie und Statistik*, *224*, 247-259.

Wolter, K. M. (2007). *Introduction to Variance Estimation* (2nd ed.). New York: Springer.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, *66*, 411-414.

Adresse des Autors:

Ralf Münnich
Universität Trier, FB IV, VWL
Wirtschafts- und Sozialstatistik
Universitätsring 15
D-54286 Trier

muennich@uni-trier.de
http://www.statistik.uni-trier.de