

# **METHODENBERICHTE**

**Das Stichprobenverfahren der Testerhebungen  
für einen registergestützten Zensus**

**Wolf Bihler**

**Gruppe Mathematisch-statistische Methoden**

**2002**

**Herausgeber und Vertriebsstelle:** Statistisches Bundesamt, Wiesbaden

**Fachliche Informationen**

zu dieser Veröffentlichung:

Wolf Bihler, Gruppe II A,  
Tel.: 06 11 / 75 28 87  
Fax: 06 11 / 75 39 51  
wolf.bihler@destatis.de

**Allgemeine Informationen**

zum Datenangebot:

Informationsservice,  
Tel.: 06 11 / 75 24 05  
Fax: 06 11 / 75 33 30  
info@destatis.de  
www.destatis.de

**Veröffentlichungskalender  
der Pressestelle:**

[www.destatis.de/presse/deutsch/cal.htm](http://www.destatis.de/presse/deutsch/cal.htm)

Erscheinungsfolge: unregelmäßig

Erschienen im November 2002

Schutzgebühr: EUR 6,75 [D] zzgl. Versandkosten

Bestellnummer: 9219020 - 99900

Recyclingpapier aus 100 % Altpapier.

© Statistisches Bundesamt, Wiesbaden 2002

Für nichtgewerbliche Zwecke sind Vervielfältigung und unentgeltliche Verbreitung, auch auszugsweise, mit Quellenangabe gestattet. Die Verbreitung, auch auszugsweise, über elektronische Systeme/Datenträger bedarf der vorherigen Zustimmung. Alle übrigen Rechte bleiben vorbehalten.

**Verlag:** Metzler-Poeschel, Stuttgart

**Verlagsauslieferung:** SFG – Servicecenter Fachverlage GmbH  
Postfach 43 43  
72774 Reutlingen  
Telefon: 0 70 71 / 93 53 50  
Telefax: 0 70 71 / 93 53 35  
[www.s-f-g.com](http://www.s-f-g.com)  
[destatis@s-f-g.com](mailto:destatis@s-f-g.com)



## Zeitreihenservice



In unserer Datenbank STATIS-BUND sind Daten aus verschiedenen Themenbereichen als Zeitreihen gespeichert und können gegen Entgelt via Internet ([www.destatis.de/zeitreih](http://www.destatis.de/zeitreih)) bezogen werden.

### Schwerpunktt Themen:

- Produzierendes Gewerbe
- Volkswirtschaftliche Gesamtrechnungen
- Preise
- Löhne und Gehälter
- Erwerbstätigkeit
- Bevölkerung
- Binnen- und Außenhandel
- Bautätigkeit

### Nutzungsmöglichkeit:

- Datenrecherche kostenfrei
- Datenabruf als registrierter Kunde  
Anmeldung und Preisregelung über  
[www-zr.destatis.de/cgi-bin/regmeg.pl](http://www-zr.destatis.de/cgi-bin/regmeg.pl)
- Datenbanksegmente ausserdem verfügbar als  
STATIS-CD-ROM (halbjährliche Ausgabe),  
Informationen und **Demo-CD-ROM** über

Das **komplette Datenbestandsverzeichnis** finden Sie als **kostenloses** Download unter:  
[www-zr.destatis.de/dbv/dbv.htm](http://www-zr.destatis.de/dbv/dbv.htm)

Informationen:

Telefon: 06 11 / 75 45 55

E-Mail: [statis@destatis.de](mailto:statis@destatis.de) ←

---

## STATISTIK-SHOP



Über den STATISTIK-SHOP stehen **rund um die Uhr (24 Stunden)** Daten aus verschiedenen Themenbereichen als Download sofort zur Verfügung. Außerdem können diverse Printprodukte, CD-ROMs bzw. Diskettenpakete online bestellt werden unter [www.destatis.de/shop](http://www.destatis.de/shop).

### Downloads-Themenauswahl:



- Gebiet, Bevölkerung, Erwerbstätigkeit, Wahlen
- Bildung, Sozialleistungen, Gesundheit, Rechtspflege
- Wohnen, Umwelt
- Wirtschaftsbereiche
- Außenhandel, Unternehmen, Handwerk
- Preise, Verdienste, Einkommen und Verbrauch
- Öffentliche Finanzen
- Volkswirtschaftliche und umweltökonomische  
Gesamtrechnungen
- Sonderberichte
- Klassifikationen

### Bücher, Fachserien, CD-ROMs:



- Jahrbücher
- Fachserien zu den einzelnen Bereichen
- Schriftreihe „Im Blickpunkt“
- Thematische Veröffentlichungen
- Gutachten des Rates von Sachverständigen  
für Umweltfragen
- Gutachten des Sachverständigenrates zur  
Begutachtung der gesamtwirtschaftlichen  
Entwicklung
- Organisation- und Methodenfragen
- Klassifikationen
- CD-ROMs und Diskettenpakete
- Gesamtkatalog

Informationen:

Telefon: 06 11 / 75 45 55

E-Mail: [shop-produkte@destatis.de](mailto:shop-produkte@destatis.de)



---

# Inhalt

<b>1</b>	<b>Einführung</b> .....	7
<b>2</b>	<b>Register- und Verfahrenstest</b> .....	8
2.1	Das Auswahlverfahren für die Stichprobe Registertest .....	8
2.1.1	Zielsetzung .....	8
2.1.2	Auswahlverfahren .....	8
2.1.3	Auswahlgrundlage und Auswahlgesamtheit .....	9
2.1.4	Schichtung.....	9
2.1.5	Auswahltechnik.....	10
2.1.6	Anordnung .....	11
2.1.7	Stichprobenumfang und Aufteilung des Stichprobenumfangs.....	12
2.1.7.1	Abschätzung der zu erwartenden Zufallsfehler .....	12
2.1.7.2	Erstellung des Auswahlplans .....	17
2.2	Das Auswahlverfahren für die Stichprobe Verfahrenstest.....	23
2.2.1	Zielsetzung .....	23
2.2.2	Auswahlverfahren und Auswahlgrundlage.....	24
2.2.3	Schichtung.....	24
2.2.4	Auswahltechnik.....	24
2.2.5	Anordnung .....	24
2.2.6	Stichprobenumfang und Aufteilung des Stichprobenumfangs.....	24
2.3	Das Hochrechnungsverfahren.....	26
2.3.1	Ziehungswahrscheinlichkeiten .....	27
2.3.2	Behandlung von Antwortausfällen .....	29
2.3.2.1	Behandlung von Antwortausfällen für Auswertungen mit Ausnahme des Registertests.....	30
2.3.2.2	Behandlung von Antwortausfällen beim Registertest .....	32
2.3.3	Anpassung an die Gebäudestruktur einschließlich der Mikrozensus-Adressen .....	37
2.3.4	Anpassung an Rahmendaten .....	38
2.4	Abschätzung der Stichprobenzufallsfehler.....	41
2.4.1	Varianzschätzung ohne Anpassung an Rahmendaten .....	41
2.4.2	Varianzschätzung mit Berücksichtigung der Anpassung .....	46
<b>3</b>	<b>Mehrfachfallprüfung</b> .....	48
3.1	Auswahlverfahren .....	48
3.2	Hochrechnungsverfahren .....	48
3.2.1	Ziehungswahrscheinlichkeiten .....	49
3.2.2	Behandlung der Antwortausfälle.....	49
3.2.3	Anpassung an Rahmendaten .....	51
3.3	Abschätzung der Stichprobenzufallsfehler.....	53
<b>4</b>	<b>Literatur</b> .....	55



---

# Das Stichprobenverfahren der Testerhebungen für einen registergestützten Zensus

## 1 Einführung

Nach der zuletzt im alten Bundesgebiet im Jahr 1987 in „traditioneller Art“ durchgeführten Volkszählung haben in den 90er Jahren Überlegungen in Deutschland eingesetzt, um – wie in anderen Staaten – nach Wegen zu suchen, wie vorhandene Datenquellen für Volkszählungszwecke genutzt werden können, um nicht zuletzt die mit einer herkömmlichen Volkszählung verbundenen Kosten und Belastungen der Bürger zu reduzieren. In Deutschland hat deshalb die Prüfung begonnen, ob sich ein Zensus realisieren lässt, der sich insbesondere auf vorhandene Daten aus den Einwohnermelderegistern der Gemeinden und aus den erwerbsstatistischen Registern der Bundesanstalt für Arbeit sowie auf eine postalische Befragung der Gebäudeeigentümer („GWZ“) stützt und auf eine primärstatistische Erhebung der gesamten Bevölkerung verzichtet. Ein solcher Methodenwechsel setzt umfangreiche methodische, organisatorische und praktische Untersuchungen voraus. Die amtliche Statistik in Deutschland hat deshalb in den letzten Jahren umfangreiche Tests zu diesem neuen Modell konzipiert, mit denen u.a. die Qualität der benutzten Register und der geplanten Verfahren durch Befragungen vor Ort kontrolliert wird. Die Tests umfassen drei Teile, die auf Stichprobenbasis durchgeführt werden:

- *Registertest:* Beim Registertest werden die personenbezogenen Daten der Einwohnermelderegister mit Befragungsergebnissen bei den Bewohnern verglichen, um Erkenntnisse über Register-Übererfassungen („Karteileichen“) und Untererfassungen („Fehlbestände“) zu erhalten.
- *Verfahrenstest:* Da in den Einwohnermelderegistern der Haushaltszusammenhang nicht vorhanden ist, sollen nach dem neuen Zensusmodell Haushalte aus den vorhandenen Informationen (aus dem Melderegister und der Eigentümerbefragung) generiert werden. Da in den Melderegistern außerdem keine Informationen über die Erwerbstätigkeit vorhanden ist, sollen die Daten der Melderegister und der erwerbsstatistischen Register zusammengeführt werden. Diese Verfahren werden getestet und ihre Ergebnisse mit Befragungsergebnissen verglichen.
- *Mehrfachfallprüfung:* Bei einem registergestützten Zensus sollen die Daten der Einwohnermelderegister nicht ungeprüft übernommen werden. Es ist vorgesehen, die Registerdaten aller Gemeinden zusammenzuführen und die Angaben von Personen mit mehr als einem Hauptwohnsitz oder nur einem Nebenwohnsitz durch Nachfragen zu überprüfen. Auch dieses Verfahren wird auf Stichprobenbasis getestet, indem festgestellte oder auch vermeintliche Doppel- bzw. Fehleintragungen durch Befragen der betroffenen Bürger aufgeklärt werden.

Dieser Bericht behandelt die Methodik der eingesetzten Stichprobenverfahren. Ausführungen zum Zensusstest aus fachstatistischer Sicht sind zum Beispiel in Rompel 2001 oder Bierau 2001 enthalten.

Die drei Teile des Zensusstests basieren auf unterschiedlichen Stichproben. Der Register- und der Verfahrenstest werden wegen ähnlicher Methodik – die Stichprobe für den Verfahrenstest ist eine Unterstichprobe der Stichprobe Registertest – im Folgenden gemeinsam in einem Kapitel dargestellt. Zum Zeitpunkt der Erstellung dieses Berichts war die Hochrechnung und die Abschätzung der Zufallsfehler noch nicht durchgeführt. Deswegen sind Analyseergebnisse in diesem Bericht nicht enthalten.

---

## 2 Register- und Verfahrenstest

### 2.1 Das Auswahlverfahren für die Stichprobe Registertest

#### 2.1.1 Zielsetzung

Ziel des Registertests ist es, die Raten der Über- und Untererfassungen<sup>1)</sup> in den Einwohnermelderegistern zuverlässig aus der Stichprobe zu schätzen, und zwar zum einen gegliedert nach 4 Gemeindegrößenklassen (unter 10 000/ 10 000 bis unter 50 000/ 50 000 bis unter 800 000/ 800 000 und mehr Einwohner), zum anderen gegliedert nach Ländern.

Für die gegliederten Schätzwerte der Über- bzw. Untererfassungsraten werden relative Standardfehler von maximal 15% angestrebt („Zielsetzung 1“). Zuverlässige Ergebnisse für die Kombination von Gemeindegrößenklasse und Ländern können nicht erwartet werden.

Ferner soll die Zahl der Gemeinden nach 3 Größenklassen der Über- bzw. Untererfassungsraten zuverlässig aus der Stichprobe geschätzt werden können. Die Ergebnisse sind für 4 Gemeindegrößenklassen (Abgrenzung wie oben) und für Länder zu erstellen. Für auf diese Weise getrennt gegliederte Ergebnisse werden relative Standardfehler von maximal 20% angestrebt („Zielsetzung 2“).

#### 2.1.2 Auswahlverfahren

Die Auswahl erfolgte durch ein zweistufiges geschichtetes Verfahren. In beiden Stufen der Auswahl wurde das Verfahren der „größenproportionalen“ Auswahl verwendet. Bei diesem Auswahlverfahren sind die Ziehungswahrscheinlichkeiten der Einheiten einer Schicht nicht gleich groß, sondern hängen proportional ab von einem Merkmal (dem Bezugsmerkmal), das die Größe der Einheiten misst. Mit diesem Verfahren ist gegenüber einer Auswahl mit konstanten Auswahlwahrscheinlichkeiten eine Steigerung der Präzision der Ergebnisse verbunden.

Auf der ersten Stufe der Auswahl wurden Gemeinden und auf der zweiten Stufe in den ausgewählten Gemeinden Adressen (Hausnummern) ausgewählt. In der Haushaltsbefragung wurden die Daten aller Personen bzw. Haushalte, die unter einer ausgewählten Adresse wohnen, erhoben. In der Eigentümerbefragung waren die Eigentümer zu allen Wohnungen unter der Adresse auskunftspflichtig. Das Bezugsmerkmal für die größenproportionale Auswahl ist auf der ersten Stufe die Einwohnerzahl der Gemeinde zum 31.12.1998 (fortgeschriebene Bevölkerung) und auf der zweiten Stufe die Zahl der unter der Adresse gemeldeten Personen (einschl. Nebenwohnsitzen).

---

1) Übererfassung: Person mit Eintrag im Register, die nicht unter der betreffenden Adresse wohnhaft ist. Untererfassung: Person wohnhaft, aber unter der betreffenden Adresse nicht im Register eingetragen. Die Raten beziehen sich auf den Registerbestand.

---

### 2.1.3 Auswahlgrundlage und Auswahlgesamtheit

Auf der ersten Stufe wurden die Gemeinden (ohne unbewohnte gemeindefreie Gebiete) aus dem Gemeindeverzeichnis zum Stand November 1999 (Gebietsstand) mit den fortgeschriebenen Bevölkerungsdaten zum 31.12.1998 ausgewählt. Die ausgewählten Gemeinden lieferten für jede Adresse die Zahl der mit Haupt- oder Nebenwohnsitz gemeldeten Personen. Aus diesem Adressenverzeichnis wurden in der zweiten Stufe die Stichprobenadressen ausgewählt. Anstalten, militärische Gebäude, Botschaften, Konsulate, Adressen mit Haushalten, die am Mikrozensus 2001 teilgenommen haben („MZ-Adressen“)<sup>2)</sup>, sowie Sammeladressen von nicht wohnhaften Personen wurden vor der Auswahl – soweit bekannt – gekennzeichnet und erhielten keine Auswahlchance. Nicht zur Auswahlgesamtheit gehören also auch Privathaushalte in Anstalten, Kasernen oder Gebäuden, in denen sich eine Botschaft befindet.

Die Auswahlgrundlage enthielt nur Adressen, unter denen Personen gemeldet waren. Adressen ohne gemeldete Personen, in denen aber Personen wohnen, hatten demnach keine Auswahlchance. Die Untererfassungsrate wird deshalb etwas unterschätzt.

### 2.1.4 Schichtung

Auf der ersten Stufe wurden die Gemeinden der Auswahlgrundlage vor der Auswahl nach Bundesländern und nach 4 Größenklassen geschichtet:

- Größenklasse 1: Gemeinden unter 10 000 Einwohner,
- Größenklasse 2: Gemeinden mit 10 000 bis unter 50 000 Einwohnern,
- Größenklasse 3: Gemeinden mit 50 000 bis unter 800 000 Einwohnern und
- Größenklasse 4: Gemeinden mit 800 000 und mehr Einwohnern.

Die oberste Größenklasse (800 000 Einwohner oder mehr), die die 4 Großstädte Berlin, Hamburg, München und Köln umfasst, wird total erfasst. Auch in den Gemeindegrößenklassen 2 und 3 gibt es Gemeinden, die so groß sind, dass ihre Wahrscheinlichkeit, in die Stichprobe zu gelangen, bei einem Auswahlverfahren mit größenproportionalen Auswahlwahrscheinlichkeiten rechnerisch größer oder gleich Eins wäre. Diese Gemeinden wurden aus der größenproportionalen Auswahl herausgenommen und der total zu erfassenden Schicht zugeordnet. Die 2 Gemeinden Bremens wurden ebenfalls der Totalschicht zugeordnet. Insgesamt ergaben sich also je Land folgende Schichten:

- Schicht 1: Gemeinden unter 10 000 Einwohner,
- Schicht 2: Gemeinden mit 10 000 bis unter 50 000 Einwohnern ohne Gemeinden der Totalschicht,
- Schicht 3: Gemeinden mit 50 000 bis unter 800 000 Einwohnern ohne Gemeinden der Totalschicht,
- Schicht 4: Gemeinden der Totalschicht.

Die Gemeinden der Totalschicht sind in der Tabelle 1 (siehe unten, Kapitel 2.1.7.2) aufgelistet.

---

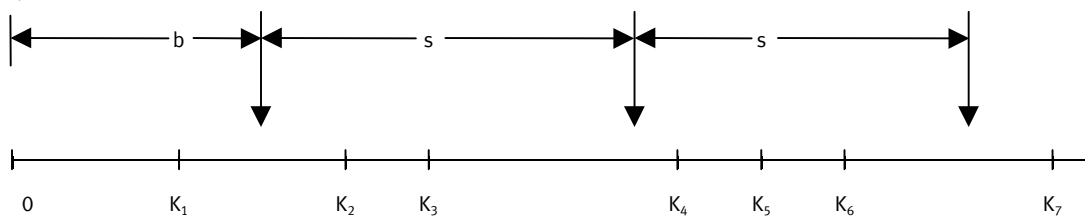
2) „MZ-Adressen“ mit 500 oder mehr gemeldeten Personen erhalten eine Auswahlchance, da ansonsten sehr große Gebäude in der Stichprobe nicht ausreichend vertreten wären. Eine erhöhte Belastung einzelner Haushalte durch zwei Befragungen muss hier zur Wahrung der Repräsentativität in Kauf genommen werden.

Auf der zweiten Auswahlstufe erfolgte keine Schichtung. Sehr sinnvoll wäre hier eine Schichtung nach dem Ausländeranteil der Adressen gewesen. In der Datenlieferung der Gemeinden (auf Adressenebene) war dieses Merkmal aber nicht enthalten. Da die Gemeinden aber die Zahl der gemeldeten Ausländer und Deutschen in aggregierter Form geliefert haben, kann ersatzweise im Rahmen der Hochrechnung eine „Schichtung nach der Auswahl“ (Anpassung) nach der Nationalität vorgenommen werden. Auf eine denkbare Schichtung nach Adressengrößenklassen (Zahl der unter der Adresse gemeldeten Personen) wurde verzichtet. Sie dürfte kaum wirksam sein, da die Adressengröße schon bei der größtenproportionalen Auswahl (siehe 2.1.5) und der Anordnung (siehe 2.1.6) berücksichtigt worden ist.

## 2.1.5 Auswahltechnik

Die größtenproportionale Auswahl erfolgte auf beiden Stufen nach dem Kumulationsverfahren. Je Schicht wurde das Bezugsmerkmal (Einwohnerzahl auf der ersten Stufe bzw. die Zahl der gemeldeten Personen auf der zweiten Stufe) Einheit (Gemeinde bzw. Adresse) für Einheit kumuliert. Anschließend wurde eine systematische Auswahl mit Schrittmaß und Zufallsstart durchgeführt. Auf der ersten Stufe wurde das berechnete Schrittmaß (siehe 2.1.7.2) verwendet, auf der zweiten Stufe war das Schrittmaß gleich der Gesamtzahl der gemeldeten Personen dividiert durch den angestrebten Stichprobenumfang an Adressen (gemäß Auswahlplan). Diejenigen Einheiten, deren Spannen (Intervalle zwischen dem Wert des kumulierten Bezugsmerkmals der Einheit und der vorangegangenen Einheit) von der systematischen Auswahl getroffen werden, gehören zur Stichprobe.

Beispiel:



s: Schrittmaß

b: Zufallsstartzahl (kleiner gleich s, größer als 0)

$K_i$ : kumulierter Merkmalswert für Einheit i

Im Beispiel werden die zweite, die vierte und die siebte Einheit ausgewählt.

Auf der ersten Stufe wurden Gemeinden, deren Einwohnerzahl größer als das Schrittmaß ist, der Totalschicht zugeordnet. Somit war gewährleistet, dass eine Gemeinde höchstens einmal durch die größtenproportionale Auswahl getroffen wurde. Auf der zweiten Stufe wären Adressen, deren Personenzahl größer als das Schrittmaß ist, durch die größtenproportionale Auswahl mehrfach getroffen worden. Diese Adressen wurden auf jeden Fall in die

---

Stichprobe genommen und von der größtenproportionalen Auswahl ausgeschlossen. Für die Auswahl der restlichen Adressen muss das Schrittmaß nachiteriert werden, damit der angestrebte Stichprobenumfang erreicht wird. In kleinen Gemeinden kann dies durchaus auftreten: beispielsweise waren in der Gemeinde Sassen (Mecklenburg-Vorpommern, Schicht 1) mit 661 gemeldeten Personen aus 158 Adressen 40 Adressen auszuwählen. Das Schrittmaß wurde zunächst als  $661/40 = 16,525$  berechnet. Unter 5 Adressen waren jeweils mehr als 16 Personen gemeldet (mit insgesamt 97 Personen). Diese Adressen gehören auf jeden Fall zur Stichprobe. Aus den verbleibenden Adressen wurden 35 größtenproportional mit einem Schrittmaß von  $(661-97)/35 = 16,114$  gezogen. De facto ist dies eine Schichtung innerhalb der Gemeinde nach total und repräsentativ zu erfassenden Adressen.

## 2.1.6 Anordnung

Auf der ersten Stufe wurden die Gemeinden der Auswahlgrundlage (Gemeindeverzeichnis Stand November 1999, ohne unbewohnte gemeindefreie Gebiete) vor der Stichprobenziehung innerhalb jeder Schicht angeordnet nach den Merkmalen

- Universitätsstadt (1 = ja, 0 = nein),
- Verdichtungsraum (nichtadministrative Gebietsgliederung der Ministerkonferenz für Raumordnung; 1 = verdichtet, 0 = nicht verdichtet),
- Einwohnerzahl (fortgeschriebene Bevölkerung am 31.12.1998).

Eine zunächst diskutierte Anordnung nach Raumordnungsregionen wurde nicht realisiert, da die große Anzahl von Ausprägungen (97) die Wirksamkeit der Anordnung nach der Einwohnerzahl zu sehr eingeschränkt hätte. Eine Anordnung nach Ausländeranteilen konnte mangels einer geeigneten Datenbasis nicht durchgeführt werden.

Auf der zweiten Stufe wurden die Adressen der Auswahlgrundlage sortiert nach

- Gemeinde,
- Gemeindeteil/Stadtbezirk,
- Zahl der unter der Adresse gemeldeten Personen (einschl. Nebenwohnsitzen),
- Straße,
- Hausnummer,
- Hausnummernteile.

Damit die Sortierung nach der Personenzahl wirksam bleibt, sollte die Zahl der Gemeindeteile/Stadtbezirke je Gemeinde beschränkt sein. Deshalb wurde folgende Regelung getroffen: Bei mehr als

- 5 Gemeindeteilen/Stadtbezirken in Gemeinden der Schicht 1,
- 10 Gemeindeteilen/Stadtbezirken in Gemeinden der Schicht 2,
- 20 Gemeindeteilen/Stadtbezirken in Gemeinden der Schichten 3 und 4

wurden die Gemeindeteile sinnvoll zusammengefasst oder – falls dies nicht möglich war – nicht berücksichtigt.

---

## 2.1.7 Stichprobenumfang und Aufteilung des Stichprobenumfangs

### 2.1.7.1 Abschätzung der zu erwartenden Zufallsfehler

Zunächst wurde der für die Zielerreichung erforderliche Stichprobenumfang kalkuliert. Dafür muss die Präzision der Ergebnisse abgeschätzt werden. Die Fehlerabschätzung musste hier ohne Simulationsrechnungen mit geeignetem Material durchgeführt werden. Die dargestellten Standardfehler sind deshalb als grobe Schätzungen anzusehen, die auf Basis der folgenden plausiblen, aber wenig gesicherten Annahmen über die „wahren“ Werte für Über- und Untererfassungsraten ermittelt wurden. Gesicherte Aussagen über die Präzision können erst nach Vorliegen der Ergebnisse durch eine Fehlerrechnung getroffen werden.

Für die Kalkulation der Stichprobenumfänge wurden je Land folgende Annahmen über die Über- und Untererfassungsraten und ihre Streuungen getroffen:

1. Für die Abschätzung der Präzision der Ergebnisse werden Raten von 0,5%, 0,6% und 1,6% angenommen. Dabei spielt es für die Stichprobenplanung keine Rolle, ob es sich um Über- oder Untererfassungsraten handelt.
2. Für die Fehlerabschätzung wird unterstellt, dass die durchschnittliche Rate je Gemeinde in einer feinen Gemeindegrößenklassengliederung (15 Größenklassen in der Abgrenzung wie im Statistischen Jahrbuch 1998, Tabelle 3.7) sich abgeschwächt proportional mit Exponent 0,1 zur durchschnittlichen Einwohnerzahl der Gemeindegrößenklasse verhält: Die durchschnittliche Rate etwa in der Größenklasse von 200 000 bis unter 500 000 Einwohnern ist ca. 2,3-mal so hoch wie in der Klasse unter 100 Einwohnern. Unterstellt man eine stärkere Abstufung der Raten, erhält man präzisere Ergebnisse, da dann die größenproportionale Auswahl wirksamer ist. Der Proportionalitätsfaktor wurde so gewählt, dass sich insgesamt die Raten von Annahme 1 ergeben.
3. Für die Varianz zwischen den Gemeinden (1. Stufe der Auswahl) wird die Annahme getroffen, dass die relative Standardabweichung (Variationskoeffizient) für Über- und Untererfassungsraten 100% nicht übersteigt. Es wird ferner angenommen, dass sich gegliedert nach Gemeindegrößenklassen die Variationskoeffizienten nicht reduzieren lassen.
4. Für die Varianz innerhalb der Gemeinden (2. Stufe der Auswahl), wird angenommen, dass sich Über- und Untererfassungsraten im Erwartungswert für die Adressen jeweils proportional zur Zahl ihrer Bewohner verhalten, wobei sich im Durchschnitt die unter der Annahme 2 berechneten Raten je Schicht ergeben. Zugrundegelegt wurde eine feine Gebäudegrößenklassengliederung (nur bundesweit) aus der Gebäude- und Wohnungsstichprobe 1978.
5. Für die Größenklassengliederung der Raten wird unterstellt, dass die Klassengrenzen so gewählt werden, dass in jede der 3 Klassen mindestens 20% der Stichprobengemeinden entfallen. Für die Zielsetzung 2 ist es sehr schwierig, den Varianzbeitrag der 2. Stufe a priori abzuschätzen. Es wird deshalb angenommen, dass er gegenüber dem Beitrag der 1. Stufe vernachlässigbar ist.

Unter diesen Annahmen können die Fehlerformeln hergeleitet werden. Dabei wird eine freie Hochrechnung unterstellt, d.h. der Totalwert  $X$  der Zahl der Über- oder Untererfassungen wird aus der Stichprobe mit den Kehrwerten der Ziehungswahrscheinlichkeiten auf beiden Stufen hochgerechnet:

$$\hat{X} = \sum_{h=1}^4 \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \frac{x_{hij}}{f_{hi}^{(1)} f_{hij}^{(2)}} \quad (2.1.1)$$

mit

$\hat{X}$  : Schätzwert für die Zahl der Über- oder Untererfassungen

$x_{hij}$  : Zahl der Über- bzw. Untererfassungen im Gebäude  $j$  der Gemeinde  $i$  der Schicht  $h$ ,

$f_{hi}^{(1)}$  : Ziehungswahrscheinlichkeit der Gemeinde  $i$  der Schicht  $h$  auf der 1.

$$\text{Stufe, } f_{hi}^{(1)} = n_h \cdot \frac{Z_{hi}}{Z_h}$$

$Z_{hi}$  : Zahl der Einwohner gemäß Auswahlgrundlage (fortgeschriebene Bevölkerung laut Gemeindedatei) in der Gemeinde  $i$  der Schicht  $h$ ,

$Z_h$  : Zahl der Einwohner gemäß Auswahlgrundlage in der Schicht  $h$ ,

$$Z_h = \sum_{i=1}^{N_h} Z_{hi}$$

$N_h$  : Zahl der Gemeinden in der Auswahlgrundlage in der Schicht  $h$ ,

$n_h$  : Zahl der Stichprobengemeinden (Stichprobenumfang 1. Stufe) in der Schicht  $h$ ,

$f_{hij}^{(2)}$  : Ziehungswahrscheinlichkeit des Gebäudes  $j$  in der Gemeinde  $i$  der Schicht  $h$

$$\text{auf der 2. Stufe, } f_{hij}^{(2)} = m_{hi} \cdot \frac{y_{hij}}{y_{hi}}$$

$y_{hij}$  : Zahl der Bewohner (gemäß Register) des Gebäudes  $j$  in der Gemeinde  $i$  der Schicht  $h$ ,

$y_{hi}$  : Zahl der Bewohner (gemäß Register) in der Gemeinde  $i$  der Schicht  $h$ ,

$$y_{hi} = \sum_{j=1}^{M_{hi}} y_{hij}$$

$M_{hi}$  : Zahl der Gebäude in der Gemeinde  $i$  der Schicht  $h$  laut Register,

$m_{hi}$  : Zahl der Stichprobengebäude (Stichprobenumfang 2. Stufe) in der Gemeinde  $i$  der Schicht  $h$ .

Die Varianz für den Schätzwert  $\hat{X}$  beträgt näherungsweise und ohne Berücksichtigung des Anordnungseffekts

$$V(\hat{X}) = \underbrace{\sum_{h=1}^4 \frac{Z_h}{n_h} \left(1 - \frac{n_h}{N_h}\right) \sum_{i=1}^{N_h} Z_{hi} \left(\frac{X_{hi}}{Z_{hi}} - \frac{X_h}{Z_h}\right)^2}_{V_1} + \underbrace{\sum_{h=1}^4 \frac{Z_h}{n_h} \sum_{i=1}^{N_h} \frac{\sigma_{hi}^2}{Z_{hi}}}_{V_2} = V_1 + V_2 \quad (2.1.2)$$

(siehe Des Raj: 1968, Theorem 3.8, S.48 und Theorem 6.3, S.118),

mit

$x_{hi}$  : Zahl der Über- bzw. Untererfassungen in der Gemeinde  $i$  der Schicht  $h$ ,

$X_h$ : Zahl der Über- bzw. Untererfassungen in der Schicht  $h$ ,  $X_h = \sum_{i=1}^{N_h} x_{hi}$

$\sigma_{hi}^2$ : Bedingte Varianz der auf die Gemeinde  $i$  (d.h. von der 2. auf die 1. Stufe) hochgerechneten Zahl der Über- bzw. Untererfassungen.

$$\sigma_{hi}^2 = \frac{y_{hi}}{m_{hi}} \sum_{j=1}^{M_{hi}} y_{hij} \left( \frac{x_{hij}}{y_{hij}} - \frac{x_{hi}}{y_{hi}} \right)^2 \quad (2.1.3)$$

$$Y_h = \sum_{i=1}^{N_h} y_{hi}$$

Auf eine näherungsweise Berücksichtigung der Auswahl ohne Zurücklegen durch die Endlichkeitskorrektur wie auf der 1. Stufe wird verzichtet, da die Auswahlätze in den Gemeinden in der Regel sehr gering sein dürften.

Da keine geeignete Datenbasis für die Kalkulation nach den Formeln (2.1.1) und (2.1.2) zur Verfügung steht, müssen gewisse Annahmen und Vereinfachungen getroffen werden.

*Varianzbeitrag  $V_1$  der 1. Stufe:*

Der Anordnungseffekt wird näherungsweise berücksichtigt, indem für die Varianzberechnung eine feine Schichtung nach 15 Gemeindegrößenklassen (Klassengrenzen wie in der Tabelle 3.7 des Statistischen Jahrbuchs 1998 mit zusätzlicher Grenze bei 800 000 Einwohnern) unterstellt wird, wobei die Schrittmaße für die größenproportionale Auswahl in allen Größenklassen einer Schicht dem Schrittmaß dieser Schicht entsprechen:

$$V_1 = \sum_{k=1}^{15} \frac{Z_k}{n_k} \left( 1 - \frac{n_k}{N_k} \right) \sum_{i=1}^{N_k} z_{ki} \left( \frac{x_{ki}}{z_{ki}} - \frac{X_k}{Z_k} \right)^2 \quad (2.1.4)$$

Bezeichnungen analog zu Formel (2.1.2) mit Index  $k$  für die Klassenzugehörigkeit anstelle des Indexes  $h$  für die Schichtzugehörigkeit.

Analog zur Vorgehensweise bei der Schätzung der Standardabweichung von Einzelwerten durch klassierte Daten werden die Einzelwerte  $z_{ki}$  durch den Klassenmittelwert  $\bar{Z}_k$  ersetzt.

Außerdem kann guten Gewissens angenommen werden, dass die Klassenrate  $\frac{X_k}{Z_k}$  (Anteil der Über- bzw. Unterer-

fassungen in der Klasse  $k$ ) durch den ungewichteten Mittelwert der Gemeinderaten  $\frac{1}{N_k} \sum_{i=1}^{N_k} \frac{x_{ki}}{z_{ki}}$  ersetzt werden

kann:

$$V_1 = \sum_{k=1}^{15} \frac{Z_k}{n_k} \left(1 - \frac{n_k}{N_k}\right) \cdot \bar{Z}_k (N_k - 1) \cdot \left(\frac{X_k}{Z_k}\right)^2 \cdot v_k^2 \quad (2.1.5)$$

mit  $v_k$  : Variationskoeffizient der Gemeinderaten in der Gemeindegrößenklasse  $k$ .

Es wird angenommen, dass die Klassenraten  $R_k = \frac{X_k}{Z_k}$  abgeschwächt proportional (mit Exponent 0,1) zum Klassenmittelwert  $\bar{Z}_k$  der Einwohnerzahl sind:

$$R_k = b \cdot \bar{Z}_k^{0,1} \quad (2.1.6)$$

Ferner wird angenommen, dass die Gesamtrate  $R = \frac{X}{Z}$  0,5%, 0,6% oder 1,6% beträgt. Die Proportionalitätskonstante  $b$  wird nun so gewählt, dass sich insgesamt die angenommene Gesamtrate ergibt:

$$b = \frac{R \cdot Z}{\sum_{k=1}^{15} \left(\frac{Z_k}{N_k}\right)^{0,1} \cdot Z_k} \quad (2.1.7)$$

also wegen  $N_k - 1 \approx N_k$

$$V_1 = \sum_{k=1}^{15} \frac{Z_k}{n_k} \left(1 - \frac{n_k}{N_k}\right) Z_k (a_k \cdot R \cdot Z)^2 v_k^2 \quad (2.1.8)$$

mit

$$a_k = \frac{(Z_k/N_k)^{0,1}}{\sum_{k=1}^{15} (Z_k/N_k)^{0,1} \cdot Z_k} \quad (2.1.9)$$

Der Variationskoeffizient  $v_k$  wurde in allen Klassen als 100 % angenommen. Gemeindezahl  $N_k$  und Einwohnerzahl  $Z_k$  haben den Stand vom 31.12.1998.

*Varianzbeitrag  $V_2$  der 2. Stufe:*

Es wird zunächst die Varianz  $\sigma_{hi}^2$  einer einzelnen Gemeinde  $i$  betrachtet.

Um den Anordnungseffekt berücksichtigen zu können, werden feinstmögliche Gebäudegrößenklassen ( $k = 1, \dots, L$ ) gebildet, d.h. für jede auftretende Gebäudegröße eine eigene Klasse, so dass in jeder Klasse die Einzelwerte  $y_{ikj}$  dem Klassenmittelwert  $\bar{y}_{ik} = \bar{y}_k$  (unabhängig von der Gemeinde) entsprechen und eine Schichtung nach diesen Klassen unterstellt:

$$\sigma_{hi}^2 = \frac{y_{hi}}{m_{hi}} \sum_{k=1}^L \bar{y}_k (M_{ik} - 1) s_{ik}^2 \quad (2.1.10)$$

mit

$s_{ik}^2$  : Varianz der Gebäuderaten zwischen den einzelnen Gebäuden der Gebäudegrößenklasse  $k$  in der Gemeinde  $i$ ,

$M_{ik}$  : Zahl der Gebäude in der Gemeinde  $i$  und der Gebäudegrößenklasse  $k$  laut Register.

Unter der Summe in Formel (2.1.10) und im folgenden wird der Übersichtlichkeit halber der Index für die Schicht  $h$ , zu der die Gemeinde gehört, weggelassen. Um die Varianz  $s_{ik}^2$  abschätzen zu können, wird ein Binominalansatz gewählt:  $p_{ikj}$  sei die Wahrscheinlichkeit, dass ein registrierter Einwohner eine Überfassung ist bzw. dass auf einen registrierten Einwohner eine Unterfassung kommt. Dann ergibt sich als Varianz  $v$  für die Zahl der Überfassung bzw. des Unterfassung im Gebäude  $j$

$$v = y_{ikj} p_{ikj} (1 - p_{ikj}) \quad (2.1.11)$$

Es wird angenommen, dass  $p_{ikj}$  proportional zur Gebäudegröße ist, d.h.

$$p_{ikj} = c_h \cdot y_{ikj} = c_h \cdot \bar{y}_k \quad (2.1.12)$$

wobei die Proportionalitätskonstante  $c_h$  für alle Gemeinden einer Schicht gleich ist. Durch Summation über alle Gebäude und Gemeinden einer Schicht kann man leicht sehen, dass gilt:

$$c_h = R_h / \frac{\sum_k M_{hk} \bar{y}_k^2}{\sum_k M_{hk} \bar{y}_k} \quad (h = 1, \dots, 4) \quad (2.1.13)$$

mit

$M_{hk}$  : Zahl der Gebäude in der Schicht  $h$  und der Gebäudegrößenklasse  $k$ ,

$R_h$  : Anteil der Über- bzw. Unterfassungen in der Schicht  $h$  (aus Formel (2.1.6)).

Das Verhältnis  $R_h/c_h$  entspricht also gerade einer mit der Gebäudegröße gewichteten durchschnittlichen Zahl der Personen je Wohnung. Diese Werte wurden aus einer Tabelle der Gebäude- und Wohnungsstichprobe 1978 geschätzt.

Ausgehend von Formel (2.1.11) ist dann

$$s_{ik}^2 = \frac{c_h \cdot \bar{y}_k^2 (1 - p_{ikj})}{\bar{y}_k^2} \approx c_h \quad (2.1.14)$$

( $1 - p_{ikj}$  vernachlässigt, Nenner  $y_{ikj} = \bar{y}_k$  ist konstant)

und damit ergibt sich aus Formel (2.1.10) und  $M_{ik} - 1 \approx M_{ik}$ :

$$\sigma_{hi}^2 \doteq \frac{y_{hi}^2}{m_{hi}} c_h \quad (2.1.15)$$

und insgesamt für den Varianzbeitrag der 2. Stufe:

$$V_2 = \sum_{h=1}^4 \frac{Z_h c_h}{n_h m_h} \sum_{i=1}^{N_h} \frac{y_{hi}^2}{Z_{hi}} \quad (2.1.16)$$

wobei in diesem Fall die Zahl der Stichprobengebäude in einer Gemeinde  $m_{hi}$  konstant ist für alle Gemeinden einer Schicht ( $= m_h$ ).

Leider standen keine Informationen über die Variable  $Y$  (Einwohnerzahl laut Register) zur Verfügung. Um a priori eine Varianzabschätzung vornehmen zu können, wird sie durch die ähnliche Variable  $Z$  (Einwohnerzahl laut Fortschreibung) ersetzt:

$$V_2 = \sum_{h=1}^4 \frac{Z_h^2 c_h}{n_h m_h} \quad (2.1.17)$$

Eigentlich interessiert mehr die Varianz der geschätzten Rate  $\hat{R} = \frac{\hat{X}}{\hat{Y}}$  als die Varianz des geschätzten Totalwerts  $\hat{X}$ . Wenn die großproportionale Auswahl auf der ersten Stufe bezüglich der Einwohnerzahl laut Register (anstelle der Einwohnerzahl laut Fortschreibung) hätte erfolgen können, wären die relativen Standardfehler für  $\hat{R}$  und  $\hat{X}$  dieselben, da  $\hat{Y} = Y$  konstant wäre. Da die Variablen  $Y$  und  $Z$  hochkorreliert sind, dürfte aber auch mit  $Z$  als Merkmal der großproportionalen Auswahl der relative Standardfehler von  $\hat{R}$  in guter Näherung durch denjenigen von  $\hat{X}$  geschätzt werden können.

## 2.1.7.2 Erstellung des Auswahlplans

Der Auswahlplan wurde an der Zielsetzung 1 (siehe Kapitel 2.1.1) ausgerichtet, da diese in der Regel zu einem höheren Stichprobenumfang als die Zielsetzung 2 führt. Anschließend wurde geprüft, ob mit diesem Auswahlplan die Zielsetzung 2 erreicht werden kann.

Um zuverlässige Länderergebnisse zu erreichen, müssen kleinere Länder mit einem deutlich höheren Auswahl-satz als größere Länder in die Stichprobe einbezogen werden. Um nach Gemeindegrößenklassen gegliederte Ergebnisse zu erhalten, müssen die Schrittmaße für die Auswahl (innerhalb eines Landes) abgestuft werden: Angenommen wurde in der Schicht 2 ein doppelt so großes Schrittmaß wie in der Schicht 1 und in der Schicht 3 ein viermal so großes Schrittmaß.

Auf der zweiten Stufe werden in jeder Stichprobengemeinde der Schicht 1 40, der Schicht 2 80 und der Schicht 3 160 Adressen ausgewählt, so dass sich innerhalb eines Landes gleiche Auswahlchancen für Adressen mit gleich viel gemeldeten Personen ergeben. Die Schrittmaße auf der ersten Stufe (und damit auch die Stichprobenumfänge) wurden dann in einem iterativen Prozess nach den oben stehenden Formeln so bestimmt, dass sich für die Raten der Länder ein relativer Standardfehler von ca. 15% ergibt. Dabei diente die Rate von 0,5% als Orientierungswert, da sie die größten Stichprobenfehler liefert. Für die Stadtstaaten führte die Fehlervorgabe zu einem er-

forderlichen Stichprobenumfang auf der zweiten Stufe von jeweils 320 Adressen. In Bremen war dieser Stichprobenumfang bevölkerungsproportional auf Bremen und Bremerhaven aufzuteilen. Für die in Flächenstaaten liegenden Städte der Schicht 4 wurde die Zahl der auszuwählenden Adressen so festgelegt, dass sich für Adressen gleicher Größe die gleichen Auswahlchancen innerhalb des Landes ergeben. Die Adressen-Stichprobenumfänge der Totalschicht-Gemeinden sind in Tabelle 1 enthalten. Der Auswahlplan war ursprünglich mit Gemeindedaten der fortgeschriebenen Bevölkerung zum 31.12.1996 erstellt worden. Aufgrund dieses Plans wurden für die Stichprobe Obergrenzen von 570 Gemeinden und 38 000 Adressen festgelegt. Der Auswahlplan wurde später mit den Bevölkerungsdaten der Auswahlgrundlage (31.12.1998) aktualisiert. Bei der Aktualisierung hätte in einigen Fällen eine fixe Zielgröße von 15% für den relativen Standardfehler der Raten in einigen Ländern eine zu große Abweichung von den vorläufig berechneten Stichprobenumfängen bedeutet; außerdem hätte die Obergrenze von 38 000 Adressen nicht eingehalten werden können. Geringfügige Abweichungen von der Zielgröße wurden deshalb toleriert. Tabelle 2 enthält die Aufgliederung des Stichprobenumfangs nach Ländern und Tabelle 3 die Aufgliederung nach Gemeindegrößenklassen. Die nach Schichten gegliederten Stichprobenumfänge sowie die für die größtenproportionale Auswahl verwendeten Schrittmaße sind in Tabelle 4 enthalten.

**Tabelle 1: Gemeinden der Totalschicht**

Land	Gemeinde	Einwohner am 31.12.1998 (in 1000)	Adressenstich- probenumfang	
			Registertest	Verfahrenstest
SH	Kiel	237,3	216	148
	Lübeck	214,0	195	134
HH	Hamburg	1 700,1	320	320
NS	Hannover	516,2	224	154
HB	Bremen	543,3	260	126
	Bremerhaven	124,7	60	40
NRW	Köln	962,6	218	150
HE	Frankfurt am Main	643,9	330	227
BW	Stuttgart	582,0	201	138
BY	München	1 188,9	305	210
SL	Saarbrücken	184,9	352	242
	Homburg	45,6	87	60
	Völklingen	43,5	83	57
BE	Berlin	3 398,8	320	320
MV	Rostock	207,4	268	184
SN	Leipzig	491,2	240	165
	Dresden	478,3	233	160
ST	Halle	260,8	227	156
	Magdeburg	239,5	208	143
TH	Erfurt	202,9	214	147

**Tabelle 2:** Stichprobenumfänge und erwartete Stichprobenfehler nach Ländern in der Stichprobe Registertest

Land	Zahl der Gemeinden		Zahl der Stichprobenadressen	Relativer Standardfehler (%) bei einer geschätzten Rate von 0,5 %
	in der Auswahl-gesamtheit (31.12.1998)	in der Stichprobe		
Baden-Württemberg	1 111	52	3 521	15,1
Bayern	2 056	51	3 105	15,2
Berlin	1	1	320	14,9
Brandenburg	1 489	48	2 760	15,2
Bremen	2	2	320	14,9
Hamburg	1	1	320	14,9
Hessen	426	39	2 890	15,1
Mecklenburg-Vorpommern	1 014	39	2 348	15,0
Niedersachsen	1 032	48	3 464	15,5
Nordrhein-Westfalen	396	36	4 138	15,1
Rheinland-Pfalz	2 306	58	3 120	15,1
Saarland	52	24	2 082	14,9
Sachsen	546	32	2 193	15,1
Sachsen-Anhalt	1 289	38	2 315	15,1
Schleswig-Holstein	1 130	41	2 491	15,2
Thüringen	1 018	45	2 574	15,1
Insgesamt	13 869	555	37 961	5,1

**Tabelle 3: Stichprobenumfänge und erwartete Stichprobenfehler  
nach Gemeindegrößenklassen in der Stichprobe Registertest**

Nr.	Gemeinden von ... bis unter ... Einwohnern	Gemeinden		Adressen der Stichprobe	Geschätz- te Rate	Rel. Standard- fehler	Geschätz- te Rate	Rel. Standard- fehler	Geschätz- te Rate	Rel. Standard- fehler
		in der Auswahl- gesamtheit	in der Stichprobe		bei einer Gesamtrate von					
					0,5 %		0,6 %		1,6 %	
	1 000	Anzahl	Anzahl	Anzahl	%	%	%	%	%	%
1	unter 10	12 338	331	13 240	0,43	8,2	0,50	7,9	1,34	6,9
2	10 – 50	1 342	158	12 650	0,47	9,4	0,57	9,3	1,52	8,8
3	50 – 800	185	62	10 908	0,58	10,2	0,70	9,7	1,85	9,5
4	800 und mehr	4	4	1 163	0,56	7,9	0,67	7,2	1,80	4,4
	Insgesamt	13 869	555	37 961	0,50	5,1	0,60	4,9	1,60	4,7

**Tabelle 4:**

*Schicht - und Stichprobenumfänge der Stichproben Registertest und Verfahrenstest nach Ländern und Schichten*

Schicht	Auswahlgesamtheit		Stichprobe Registertest			Stichprobe Verfahrenstest		
	Gemeinden	Einwohner am 31.12.1998 (in 1000)	Gemeinden	Adressen	Schrittmaß der Gemeindeauswahl	Gemeinden	Adressen	Schrittmaß der Unterauswahl der Gemeinden
<i>Schleswig-Holstein</i>								
1	1 078	1 210,0	28	1 120	44 000	7	280	4,0
2	47	868,4	10	800	88 000	5	300	2,0
3	3	236,3	1	160	176 000	1	110	1,0
4	2	451,4	2	411		2	282	1,0
Zusammen	1 130	2 766,1	41	2 491		15	972	
<i>Hamburg</i>								
4	1	1 700,1	1	320		1	320	1,0
Zusammen	1	1 700,1	1	320		1	320	
<i>Niedersachsen</i>								
1	832	2 063,7	23	920	92 000	5	200	4,0
2	181	3 562,5	19	1 520	184 000	10	600	2,0
3	18	1 723,5	5	800	368 000	5	550	1,0
4	1	516,2	1	224		1	154	1,0
Zusammen	1 032	7 865,8	48	3 464		21	1 504	
<i>Bremen</i>								
4	2	668,0	2	320		2	166	1,0
Zusammen	2	668,0	2	320		2	166	
<i>Nordrhein-Westfalen</i>								
1	58	463,3	2	80	177 000	1	40	2,0
2	264	6 168,3	18	1 440	354 000	6	360	3,0
3	73	10 381,3	15	2 400	708 000	10	1 100	1,5
4	1	962,6	1	218		1	150	1,0
Zusammen	396	17 975,5	36	4 138			1 650	
<i>Hessen</i>								
1	255	1 405,8	18	720	78 000	5	200	4,0
2	159	2 802,1	17	1 360	156 000	9	540	2,0
3	11	1 183,4	3	480	312 000	3	330	1,0
4	1	643,9	1	330		1	227	1,0
Zusammen	426	6 035,1	39	2 890		18	1 297	
<i>Rheinland-Pfalz</i>								
1	2 260	2 359,3	46	1 840	51 000	12	480	4,0
2	38	803,9	8	640	102 000	4	240	2,0
3	8	861,8	4	640	204 000	4	440	1,0
Zusammen	2 306	4 025,0	58	3 120		20	1 160	

noch Tabelle 4:

Schicht	Auswahlgesamtheit		Stichprobe Registertest			Stichprobe Verfahrenstest		
	Gemeinden	Einwohner am 31.12.1998 (in 1000)	Gemeinden	Adressen	Schrittmaß der Gemein- deauswahl	Gemeinden	Adressen	Schrittmaß der Unteraus- wahl der Ge- meinden
<i>Baden-Württemberg</i>								
1	869	3 381,2	29	1 160	116 000	7	280	4,0
2	219	4 142,6	17	1 360	232 000	8	480	2,0
3	22	2 320,3	5	800	464 000	5	550	1,0
4	1	582,0	1	201		1	138	1,0
Zusammen	1 111	10 426,0	52	3 521		21	1 448	
<i>Bayern</i>								
1	1 840	5 620,7	36	1 440	156 000	9	360	4,0
2	199	3 416,4	11	880	312 000	5	300	2,0
3	16	1 860,5	3	480	624 000	3	330	1,0
4	1	1 188,9	1	305		1	210	1,0
Zusammen	2 056	12 086,5	51	3 105		18	1 200	
<i>Saarland</i>								
1	12	95,9	5	200	21 000	1	40	4,0
2	36	652,9	15	1 200	42 000	8	480	2,0
3	1	51,5	1	160	84 000	1	110	1,0
4	3	274,0	3	522		3	359	1,0
Zusammen	52	1 074,2	24	2 082		13	989	
<i>Berlin</i>								
4	1	3 398,8	1	320		1	320	1,0
Zusammen	1	3 398,8	1	320		1	320	
<i>Brandenburg</i>								
1	1 438	1 253,9	33	1 320	38 000	8	320	4,0
2	47	935,5	12	960	76 000	6	360	2,0
3	4	400,9	3	480	152 000	3	330	1,0
Zusammen	1 489	2 590,4	48	2 760		17	1 010	
<i>Mecklenburg-Vorpommern</i>								
1	987	926,6	30	1 200	31 000	7	280	4,0
2	22	365,7	5	400	62 000	3	180	2,0
3	4	299,0	3	480	124 000	3	330	1,0
4	1	207,4	1	268		1	184	1,0
Zusammen	1 014	1 798,7	39	2 348		14	974	
<i>Sachsen</i>								
1	474	1 684,7	21	840	82 000	5	200	4,0
2	65	1 272,2	7	560	164 000	4	240	2,0
3	5	563,1	2	320	328 000	2	220	1,0
4	2	969,5	2	473		2	325	1,0
Zusammen	546	4 489,4	32	2 193		13	985	
<i>Sachsen-Anhalt</i>								
1	1 250	1 234,3	27	1 080	46 000	7	280	4,0
2	35	803,0	8	640	92 000	4	240	2,0
3	2	137,0	1	160	136 975	1	110	1,0
4	2	500,3	2	435		2	299	1,0
Zusammen	1 289	2 674,5	38	2 315		14	929	

noch Tabelle 4:

Schicht	Auswahlgesamtheit		Stichprobe Registertest			Stichprobe Verfahrenstest		
	Gemeinden	Einwohner am 31.12.1998 (in 1000)	Gemeinden	Adressen	Schrittmaß der Gemein- deauswahl	Gemeinden	Adressen	Schrittmaß der Unteraus- wahl der Gemein- den
<i>Thüringen</i>								
1	985	1 258,1	33	1 320	38 000	9	360	4,0
2	28	673,8	9	720	76 000	4	240	2,0
3	4	328,0	2	320	152 000	2	220	1,0
4	1	202,9	1	214		1	147	1,0
Zusammen	1 018	2 462,8	45	2 574		16	967	
<i>Insgesamt</i>								
1	12 338	22 957,5	331	13 240		83	3 320	
2	1 340	26 467,1	156	12 480		76	4 560	
3	171	20 346,7	48	7 680		43	4 730	
4	20	12 265,8	20	4 561		20	3 281	
Zusammen	13 869	82 037,0	555	37 961		222	15 891	

Für diesen Auswahlplan gilt hinsichtlich der Zielsetzung 2:

- Die Zielsetzung kann, was nach Gemeindegrößenklassen gegliederte Ergebnisse betrifft, voraussichtlich erreicht werden: Für die Zahl der Gemeinden einer 20%-Untergruppe (gemäß der Annahme 5 über die Größenklassen der Raten) beträgt der relative Standardfehler für Schicht 1 etwa 11%, für Schicht 2 etwa 15% und Schicht 3 etwa 20%.
- Für die Länder beträgt der relative Standardfehler für die Zahl der Gemeinden einer 20%- Untergruppe durchschnittlich etwa 30%. Die Zielsetzung wird für Länderergebnisse also nicht ganz erreicht. Um einen relativen Standardfehler von 20% zu erreichen, hätte der Stichprobenumfang im Saarland auf 34, in Nordrhein-Westfalen auf 80 und in den übrigen Flächenstaaten auf 90 bis 95 Gemeinden erhöht werden müssen (insgesamt ca. 1020 Gemeinden).

## 2.2 Das Auswahlverfahren für die Stichprobe Verfahrenstest

### 2.2.1 Zielsetzung

Beim Verfahrenstest werden nur Bundesergebnisse angestrebt. Eine Gliederung der Ergebnisse nach Ländern oder Gemeindegrößenklassen wie beim Registertest ist nicht erforderlich. Wegen des Verzichts auf Regionalergebnisse ist ein geringerer Stichprobenumfang als für den Registertest ausreichend. Die fachliche Gliederung der Ergebnisse stand zum Zeitpunkt der Erstellung des Auswahlplans noch nicht fest.

---

## 2.2.2 Auswahlverfahren und Auswahlgrundlage

Die Stichprobe Verfahrenstest ist eine Unterstichprobe der Stichprobe Registertest. Sowohl auf der ersten Stufe (Gemeinden) als auch auf der zweiten Stufe (Adressen) erfolgt eine Unterauswahl durch eine geschichtete Zufallsauswahl. Auswahlgrundlage war die Menge der für den Registertest ausgewählten Gemeinden (1. Stufe) bzw. Adressen (2. Stufe).

## 2.2.3 Schichtung

Vor der Auswahl wurden auf der ersten Stufe die Gemeinden der Stichprobe Registertest wie im Registertest geschichtet (Bundesländer und Gemeindegrößenklassen-Schichten). Auf der zweiten Stufe erfolgte keine Schichtung. Damit auch für den Verfahrenstest in einer Gemeinde größenproportionale Auswahlwahrscheinlichkeiten über beide Phasen hinweg (erste Phase: Auswahl Registertest, zweite Phase: Unterauswahl Verfahrenstest) realisiert wären, müsste - wie im Registertest - nach zwei Adressengrößenklassen geschichtet werden, wobei die Schicht der Adressen mit großer Personenzahl total zu erfassen wäre. Da diese Totalschichten vermutlich ohnehin leer wären - die unterste Grenze für total zu erfassende Adressen würde nämlich bei 167 gemeldeten Personen je Adresse liegen (in einer Gemeinde mit 10 000 Einwohnern; bei größeren Gemeinden läge sie höher) -, wurde auf diese Schichtung verzichtet.

## 2.2.4 Auswahltechnik

Die Unterauswahl erfolgte auf beiden Stufen systematisch mit Zufallsstart. Um den angestrebten Stichprobenumfang (gemäß Auswahlplan) möglichst gut zu treffen, wurden auch gebrochene Schrittweiten zugelassen.

## 2.2.5 Anordnung

Auf der ersten Stufe wurden die Gemeinden vor der Auswahl innerhalb jeder Schicht nach der Einwohnerzahl (fortgeschriebene Bevölkerung am 31.12.1998) angeordnet. Die Anordnung der Adressen auf der zweiten Stufe war dieselbe wie bei der Auswahl der Stichprobe Registertest.

## 2.2.6 Stichprobenumfang und Aufteilung des Stichprobenumfangs

Der erforderliche Stichprobenumfang wurde nicht analog zur Vorgehensweise beim Registertest kalkuliert, da die fachliche Gliederung der Ergebnisse zum Zeitpunkt der Erstellung des Auswahlplans noch nicht feststand; zudem hätte eine neue Berechnung wegen der Zielsetzung (nur Bundesergebnisse) dazu geführt, dass in einigen

---

Ländern der Stichprobenumfang höher als beim Registertest wäre und somit der Verfahrenstest nicht als Unterstichprobe des Registertests hätte angelegt werden können.

Für die erste Stufe wurden in allen Ländern mit Ausnahme Nordrhein-Westfalens folgende Schrittmaße für die systematische Auswahl festgelegt:

- Schicht 1: jede vierte Gemeinde,
- Schicht 2: jede zweite Gemeinde,
- Schichten 3 und 4: alle Gemeinden.

Da die Schrittmaße der Schichten beim Registertest im Verhältnis 4 : 2 : 1 abgestuft waren, ergibt sich für den Verfahrenstest über beide Phasen gesehen, dass der Gemeindestichprobenumfang bevölkerungsproportional auf die Schichten aufgeteilt ist.

In Nordrhein-Westfalen hätte sich bei diesen Schrittmaßen der Stichprobenumfang gegenüber dem Registertest wegen der besonderen Größenklassenstruktur nur wenig reduziert. In Nordrhein-Westfalen wurde deshalb die Unterstichprobe so angelegt, dass sie – über beide Phasen der Auswahl gesehen - einer größenproportionalen Auswahl mit einem Schrittmaß von 1 Mill. Einwohnern entspricht. Dies wurde durch die Auswahl jeder zweiten Gemeinde in der Schicht 1, jeder dritten Gemeinde in der Schicht 2 und einem Schrittmaß von 1,5 in der Schicht 3 erreicht. Das Schrittmaß für die Totalschicht war 1.

Der Zielsetzung (nur Bundesergebnisse) wäre es angemessener gewesen, den Stichprobenumfang auch auf die Länder bevölkerungsproportional aufzuteilen. Dies hätte aber für bevölkerungsreiche Länder nur zu einer geringen Reduzierung oder sogar zu einer Erhöhung des Stichprobenumfangs gegenüber dem Registertest geführt.

Je Gemeinde wurden die folgenden Anzahlen von Adressen ausgewählt:

- Schicht 1: 40 Adressen (d.h. also alle Adressen der Stichprobe Registertest),
- Schicht 2: 60 Adressen (75% der Stichprobenadressen der Stichprobe Registertest),
- Schicht 3: 110 Adressen (ca. 69% der Stichprobenadressen der Stichprobe Registertest),
- Schicht 4: unterschiedliche Stichprobenumfänge je Gemeinde (siehe Tabelle 1).

Der ursprünglich für Schicht 3 festgesetzte Stichprobenumfang von 120 Adressen konnte bei der endgültigen Erstellung des Auswahlplans (auf der Basis der Bevölkerungsdaten zum 31.12.1998) nicht beibehalten werden, da die Obergrenze des Gesamtstichprobenumfangs (16 000 Adressen) überschritten worden wäre. Für die Gemeinden der Totalschicht (Schicht 4) in Flächenstaaten wurde der Stichprobenumfang so festgelegt, dass sich für Adressen einer bestimmten Größe die gleichen Auswahlchancen wie in der Schicht 3 des betreffenden Landes ergaben. Dies führt zu einem Auswahlatz von ca. 69% der Stichprobenadressen der Stichprobe Registertest je Gemeinde (Stichprobenumfänge siehe Tabelle 1).

Der Adressenstichprobenumfang für die Stadtstaaten wurde so angesetzt, dass sich für Adressen einer bestimmten Größe die gleichen Auswahlchancen (im Bundesdurchschnitt) wie in der Schicht 3 ergeben. Dies führt rechnerisch zu einem Stichprobenumfang von 790 Adressen in Berlin und 395 Adressen in Hamburg. Da für die Unterstichprobe nicht mehr Adressen ausgewählt werden sollten als für die Hauptstichprobe, wurde der Stichprobenumfang für beide Städte auf 320 Adressen festgelegt. Für Bremerhaven ergab sich rechnerisch ein Umfang von nur 29 Adressen, der auf 40 heraufgesetzt wurde.

Die resultierenden, nach Ländern und Schichten gegliederten Stichprobenumfänge sind in Tabelle 4 enthalten; in der letzten Spalte befinden sich die für die systematische Auswahl auf der ersten Auswahlstufe verwendeten Schrittmaße.

---

Für einen Anteilswert einer Merkmalsausprägung von 0,5% (bezogen auf die Gesamtbevölkerung) ist – unter denselben Annahmen für eine Fehlerabschätzung wie für die Über- und Untererfassungsrate beim Registertest – mit einem relativen Standardfehler von rund 9%, für einen Anteilswert von 0,1% mit rund 15% zu rechnen.

## 2.3 Das Hochrechnungsverfahren

Allgemeines Ziel der Hochrechnung ist es, mit Hilfe geeigneter Schätzfunktionen aus den Stichprobenparametern (beispielsweise Totalwert, Mittelwert, Anteilswert, Varianz) auf die Parameter der Grundgesamtheit zu schließen. Die Hochrechnung erfolgt in folgenden vier Schritten:

- *Berechnung der Ziehungswahrscheinlichkeiten*  
Basis der Hochrechnung sind die Ziehungswahrscheinlichkeiten der Adressen. Wegen der Auswahl mit Größenproportionalen Wahrscheinlichkeiten auf den beiden Auswahlstufen (Gemeinden und Adressen) müssen diese Wahrscheinlichkeiten bei der Hochrechnung auf jeden Fall berücksichtigt werden, um die Überrepräsentation großer Adressen bzw. Gemeinden in der Stichprobe zu korrigieren.
- *Behandlung von Antwortausfällen*  
Trotz Auskunftspflicht ist mit Antwortausfällen zu rechnen. Antwortausfälle treten in der Regel nicht zufällig auf und können eine Verzerrung der geschätzten Ergebnisse bewirken. Da über einen Großteil der Antwortausfälle Hilfsinformationen bekannt sind, können Antwortwahrscheinlichkeiten geschätzt werden, die im Rahmen der Hochrechnung für die Konstruktion von Korrekturfaktoren verwendet werden, um die Verzerrung zu minimieren.
- *Anpassung an die Adressenstruktur einschließlich der Mikrozensus-Adressen*  
Die Tatsache, dass Mikrozensus-Adressen (MZ-Adressen) mit weniger als 500 gemeldeten Personen keine Auswahlchance erhielten, führt zu einer verzerrten Gebäudestruktur, die bei der Hochrechnung berücksichtigt werden muss.
- *Anpassung an Rahmendaten*  
Durch eine Anpassung an Eckwerte der Grundgesamtheit wird die Genauigkeit von Stichprobenergebnissen oftmals verbessert.

Die Hochrechnung kann erst durchgeführt werden, wenn alle erhobenen Daten plausibilisiert und zusammengeführt worden sind. Da diese Arbeitsgänge noch ausstehen, enthalten die folgenden Ausführungen zum Hochrechnungsverfahren keine quantitativen Aussagen, etwa zum Umfang und der Struktur der Antwortausfälle. Die Hochrechnung erfolgt grundsätzlich je Bundesland; in den Formeln wird der Übersichtlichkeit halber auf einen zusätzlichen Index für das Land verzichtet. Sofern nicht anders erwähnt, werden die Stichproben für Register- und Verfahrenstest nach demselben Schema hochgerechnet und die Formeln gelten für beide Stichproben. Auch hierfür wird auf einen zusätzlichen Index verzichtet. Ebenfalls der Übersichtlichkeit halber wurde in den Formeln auf Doppelindizes (Adressen innerhalb einer Gemeinde) verzichtet und die Adressen über die Gemeinden hinweg indiziert.

Die beschriebenen vier Hochrechnungsschritte führen zu folgender allgemeiner Schätzfunktion für den Totalwert  $Y$  eines Erhebungsmerkmals:

$$\hat{Y} = \sum_{k \in S} g_k^{(2)} g_k^{(1)} c_k w_k y_k \quad (2.3.1)$$

mit

- $y_k$  : Merkmalswert der Darstellungseinheit  $k$ ,
- $w_k$  : Design-Faktor (Hochrechnungsfaktor der freien Hochrechnung),
- $c_k$  : Korrekturfaktor für Antwortausfälle,
- $g_k^{(1)}$  : Anpassungsfaktor für fehlende MZ-Adressen,
- $g_k^{(2)}$  : Anpassungsfaktor für Rahmendaten.

Je nach Auswertungsziel kann die Darstellungseinheit  $k$  die Person, der Haushalt oder die Wohnung sein. Die Summation erstreckt sich über alle Darstellungseinheiten in der Stichprobe  $s$ .

Anteile, Mittelwerte oder Verhältniszahlen werden geschätzt, indem die Totalwerte von Zähler und Nenner zunächst getrennt nach Formel (2.3.1) hochgerechnet und anschließend ins Verhältnis gesetzt werden. Für den Zensustest sind neben diesen Parametern auch Maße für die Streuung wichtig: eine bestimmte Fehlerrate (Über- bzw. Untererfassung) wird anders beurteilt, wenn sie in allen Gemeinden in etwa gleich groß ist, als wenn es sehr große Unterschiede zwischen den Gemeinden gibt. Als Maße für die Streuung der Über- und Untererfassungsraten der Gemeinden sind die Standardabweichung und Quantilsabstände (beide mit der Gemeindegröße gewichtet) in der Diskussion. Wegen der Zweistufigkeit der Auswahl sind Schätzungen dieser Maße verzerrt, wenn sie mit von der zweiten auf die erste Stufe hochgerechneten Gemeinderaten durchgeführt werden. Diese Verzerrung kann wegen des relativ kleinen Stichprobenumfangs je Gemeinde nicht vernachlässigt werden. Für die Standardabweichung besteht ein möglicher Ansatz darin, die Verzerrung unter Verwendung der Fehlervarianzen der Raten der einzelnen Gemeinden zu schätzen. Für Gemeinden der Totalschicht werden diese Fehlervarianzen auch für die Abschätzung der Zufallsfehler von Totalwerten benötigt (siehe im Kapitel 2.4 die Formel 2.4.6). Die Überlegungen hierzu sind noch nicht abgeschlossen.

In den folgenden Kapiteln 2.3.1 bis 2.3.4 werden die von den vier Hochrechnungsschritten hergeleiteten Faktoren aus Formel (2.3.1) beschrieben.

### 2.3.1 Ziehungswahrscheinlichkeit

Basis der Hochrechnung sind die Ziehungswahrscheinlichkeiten der Adressen (Wahrscheinlichkeit für eine Adresse, in die Stichprobe zu gelangen). Für die Berechnung dieser Wahrscheinlichkeiten können die Schrittweiten der großenproportionalen Auswahl herangezogen werden. Die Ziehungswahrscheinlichkeit einer Gemeinde bzw. einer Adresse hängt von der Gemeinde- bzw. Adressengröße ab und ergibt sich als Gemeinde- bzw. Adressengröße dividiert durch das Schrittmaß. Für Gemeinden bzw. Adressen größer gleich dem Schrittmaß ist die Ziehungswahrscheinlichkeit gleich Eins, da diese der total zu erfassenden Schicht zugeordnet wurden. Die Gemeindegröße ist hier gemessen mit der fortgeschriebenen Bevölkerung am 31.12.1998 aus dem Gemeindeverzeichnis und die

Adressengröße ist die Zahl der im Register (mit Haupt- und Nebenwohnung) gemeldeten Personen zum Zeitpunkt der Adressenlieferung der Gemeinden.

Über beide Stufen zusammen gesehen ist die Ziehungswahrscheinlichkeit einer Adresse das Produkt der Ziehungswahrscheinlichkeiten der Adressen und der Gemeinde. Der Kehrwert dieser Wahrscheinlichkeit ist dann der Hochrechnungsfaktor der freien Hochrechnung oder der Design-Faktor. Der resultierende Faktor ist adressenspezifisch, d.h. er ist für alle Personen und Haushalte/Wohnungen einer Adresse gleich.

Die Stichprobe Verfahrenstest ist auf beiden Stufen eine Unterstichprobe der Stichprobe Registertest. Bezogen auf die Stichprobe Registertest ist die Stichprobe Verfahrenstest zwar eine systematische Auswahl mit Zufallsstart (gleiche Ziehungswahrscheinlichkeiten innerhalb einer Schicht), kann aber auf die Grundgesamtheit bezogen auch als größenproportionale Auswahl (mit größeren Schrittmaßen als im Registertest) aufgefasst werden. Die Berechnung der Ziehungswahrscheinlichkeiten kann dann für beide Stichproben nach demselben Schema erfolgen; die größeren Schrittmaße in der Stichprobe Verfahrenstest führen lediglich zu höheren Hochrechnungsfaktoren als in der Stichprobe Registertest.

*Formelmäßige Darstellung :*

- $s_h^{(1)}$  : Schrittmaß der größenproportionalen Auswahl auf der ersten Stufe in der Schicht  $h$  ( $h = 1, 2, 3$ )<sup>3)</sup>,
- $n_h$  : Zahl der Stichprobengemeinden in der Schicht  $h$ ,
- $s_i^{(2)}$  : Schrittmaß der größenproportionalen Auswahl auf der zweiten Stufe in der Stichprobengemeinde  $i$  ( $i = 1, \dots, n_h$ ),
- $Z_i$  : Bevölkerung am 31.12.1998 in der Gemeinde  $i$ ,
- $m_{hi}$  : Zahl der Stichprobenadressen in der Gemeinde  $i$  der Schicht  $h$ ,
- $u_j$  : Zahl der gemeldeten Personen unter der Adresse  $j$  ( $j = 1, \dots, m_{hi}$ ) zum Zeitpunkt der Adressenlieferung.

Die Ziehungswahrscheinlichkeit der Gemeinde  $i$  ist

$$f_i^{(1)} = \begin{cases} \frac{Z_i}{s_h^{(1)}}, & \text{falls Gemeinde } i \text{ zur Schicht } h \text{ (} h = 1, 2, 3 \text{) gehört,} \\ 1, & \text{falls Gemeinde } i \text{ zur Totalschicht (} h = 4 \text{) gehört.} \end{cases} \quad (2.3.2)$$

Die Ziehungswahrscheinlichkeit der Adresse  $j$  in der ausgewählten Gemeinde  $i$  ist

$$f_j^{(2)} = \begin{cases} \frac{u_j}{s_i^{(2)}}, & \text{falls } u_j < s_i^{(2)} \\ 1, & \text{sonst} \end{cases} \quad (2.3.3)$$

3) Kein Schrittmaß in Schicht 4, da in Schicht 4 alle Gemeinden ausgewählt wurden.

---

Der Hochrechnungsfaktor der freien Hochrechnung ist also als Kehrwert der Ziehungswahrscheinlichkeit für eine Adresse  $j$  in der Gemeinde  $i$

$$w_j = w_i^{(1)} \cdot w_j^{(2)}, \quad (2.3.4)$$

wobei  $w_i^{(1)} = \frac{1}{f_i^{(1)}}$  der Gemeinde- und  $w_j^{(2)} = \frac{1}{f_j^{(2)}}$  der Adressenhochrechnungsfaktor ist.

Dieser Hochrechnungsfaktor gilt auch für alle Darstellungseinheiten  $k$  unter der Adresse  $j$ . Bei Aussagen auf der Ebene der Darstellungseinheiten wird er mit  $w_k$  bezeichnet.

*Beispiel:*

In der Schicht 2 von Rheinland-Pfalz wurde die Stadt Pirmasens für den Registertest ausgewählt. Die dem Auswahlplan zugrunde liegende Bevölkerungszahl von Pirmasens war 46 400 Einwohner (fortgeschriebene Bevölkerung am 31.12.98). Das Schrittmaß für die Gemeindeauswahl betrug in dieser Schicht 102 000 Einwohner. Die Wahrscheinlichkeit von Pirmasens, in die Stichprobe zu gelangen, betrug also  $46\,400/102\,000 = 45,5\%$ . Der Gemeindehochrechnungsfaktor ist also  $102\,000/46\,400 = 2,198$ . Aus der Adressenlieferung ergaben sich für Pirmasens 46 120 gemeldete Personen in „normalen“ Gebäuden (ohne Anstalten, Kasernen, MZ-Gebäude etc.). Da 80 Adressen ausgewählt werden sollten, wurde das Schrittmaß für die Adressenauswahl auf  $46\,120/80 = 576,5$  festgelegt. Für eine ausgewählte Adresse mit beispielsweise 20 gemeldeten Personen betrug die Ziehungswahrscheinlichkeit  $20/576,5 = 3,5\%$ . Ihr Adressenhochrechnungsfaktor ist also  $576,5/20 = 28,825$ . Insgesamt ist der Hochrechnungsfaktor der freien Hochrechnung für diese Adresse:  $2,198 \cdot 28,825 = 63,365$ .

## 2.3.2 Behandlung von Antwortausfällen

Während bei der Lieferung von Melderegisterdaten aus den Gemeinden und von Daten der Bundesantalt für Arbeit zur Erwerbstätigkeit keine Antwortausfälle auftreten, ist bei den drei Primärerhebungen

- Haushaltsbefragung Registertest
- Haushaltsbefragung Verfahrenstest
- Eigentümerbefragung<sup>4)</sup>

trotz Auskunftspflicht mit Antwortausfällen zu rechnen. Diese Antwortausfälle werden nicht rein zufällig auftreten, sondern sie hängen ab von Merkmalen, die zum Teil beobachtet werden können. Die Antwortausfälle können danach untersucht werden, in welchen Adressen sie auftreten. Für alle Stichprobenadressen liegen nämlich aus der Datenlieferung der Gemeinden Registerinformationen vor, daneben können natürlich auch noch Regionalinformationen und die Gemeindegröße genutzt werden. Es ist methodisch äußerst wichtig, zwischen *Antwortausfällen* (auch *echte Ausfälle* genannt) und sonstigen Ausfällen (sogenannte *stichprobenneutrale Ausfälle* oder *unechte Ausfälle*) zu unterscheiden.

---

4) Eigentümerbefragung und Haushaltsbefragung Verfahrenstest basieren auf derselben Adressen-Stichprobe.

---

*Stichprobenneutrale Ausfälle* sind Gebäude oder Wohnhaushalte, die nicht zur Zielgesamtheit der Testerhebung gehören, z.B.

- Wohnung oder Gebäude wird nur gewerblich genutzt,
- Wohnung oder Gebäude steht leer,
- unter der ausgewählten Adresse ist kein Gebäude vorhanden,
- Anstalten, Sammeladressen, Kasernen, Botschaften<sup>5)</sup>, Privatwohnungen von Diplomaten oder Angehörigen ausländischer Streitkräfte.

Wenn einer dieser Sachverhalte zweifelsfrei festgestellt wurde, zählt dies als Befragungsergebnis und im stichprobenmethodischen Sinne als Antwort.

*Antwortausfälle* sind Gebäude oder Wohnhaushalte, die zur Zielgesamtheit der Testerhebung gehören, für die aber keine Antwort vorliegt, z.B.

- Gebäudeeigentümer nicht ermittelbar und kein stichprobenneutraler Ausfall,
- Wohnungsinhaber nicht erreichbar oder verweigert Auskunft explizit.

Manchmal ist nicht eindeutig zu erkennen, ob es sich um einen Antwort- oder einen stichprobenneutralen Ausfall handelt. Solche Zweifelsfälle sind als Antwortausfälle zu zählen.

Ausfälle einzelner Personen innerhalb einer Wohnung sind in der Regel überhaupt nicht erkennbar und werden hier nicht behandelt. Da man bei einem sehr großen Teil der Antwortausfälle – durch Interviewerbeobachtungen und Vergleich mit dem Melderegister - weiß, ob es sich um mit dem Melderegister paarige Fälle oder um Untererfassungen handelt, kann man diese Information zusätzlich bei der Schätzung der Über- und Untererfassungsrate beim Registertest nutzen. Wegen des etwas komplexeren Ansatzes ist die Behandlung der Antwortausfälle beim Registertest in einem eigenen Unterkapitel dargestellt.

### **2.3.2.1 Behandlung der Antwortausfälle für Auswertungen mit Ausnahme des Registertests**

Zentraler Punkt ist die Schätzung der Antwortwahrscheinlichkeiten der Haushalte. Die Antwortausfälle können danach untersucht werden, unter welchen Adressen sie auftreten. Für alle Stichprobengebäude liegen aus der Datenlieferung der Gemeinden Registerinformationen vor. Merkmale, die mit dem Antwortverhalten zusammenhängen können, sind z.B.

- Adressengröße, d.h. die Anzahl der unter der Adresse gemeldeten Personen,
- Ausländeranteil unter der Adresse,
- Anteil der Personen mit Nebenwohnsitz unter der Adresse.

Die Merkmale werden in Klassen typisiert. Anzahl und Abgrenzung der Klassen können erst festgelegt werden, wenn die Antwortausfälle bekannt sind. Als weiteres Merkmal kann die Gemeindegrößenklasse herangezogen werden. Die Merkmale werden miteinander kombiniert. Die Kombinationen der Klassen dieser Merkmale werden

---

5) Einzelne Fälle in der Stichprobe traten auf, da die Kennzeichnung dieser Adressen in der Auswahlgrundlage manuell erfolgte und die zugrunde liegenden Listen nicht ganz vollständig war.

---

im folgenden als Antwortklassen bezeichnet. Der Anteil der Antwortausfälle in einer Antwortklasse dient nun als Schätzwert  $q_j$  für die Wahrscheinlichkeit, dass ein Haushalt dieser Antwortklasse ausfällt. Der Anteil muss mit den Faktoren der freien Hochrechnung  $w_j$  hochgerechnet werden:

$$q_j = \frac{\sum_{j \in s} w_j b_j^{(a)}}{\sum_{j \in s} w_j b_j} \quad (2.3.5)$$

mit

$q_j$  : Schätzwert für Ausfallwahrscheinlichkeit eines Haushalts unter der Adresse  $j$ ,

$b_j$  : Anzahl der Haushalte unter der Adresse  $j$ ,

$b_j^{(a)}$  : Anzahl der Antwortausfälle (Haushalte) unter der Adresse  $j$ .

Die Summation erstreckt sich über alle Adressen  $j$ , die zur Antwortklasse  $s$  ( $s = 1, \dots, S$ ) gehören.

Der Korrekturfaktor  $c_j$  für die Antwortausfälle ist nun der Kehrwert der Antwortwahrscheinlichkeit  $p_j = 1 - q_j$  :

$$c_j = \frac{1}{p_j} \quad (2.3.6)$$

Die Gebäudemerkmale als Hilfsinformation wurden nur ersatzweise verwendet für fehlende Daten auf der Ebene der Haushalte. Wenn der Ausfallmechanismus im wesentlichen von Haushaltsmerkmalen bestimmt wird, die nicht oder nur sehr gering mit den Gebäudetypen zusammenhängen, werden die Antwortwahrscheinlichkeiten der Haushalte unbefriedigend geschätzt und die von Antwortausfällen möglicherweise verursachte Verzerrung der Ergebnisse kann durch die Hochrechnung nicht verringert werden.

Für jede der drei Primärerhebungen

- Haushaltsbefragung der Stichprobe Registertest (SR)
- Haushaltsbefragung der Stichprobe Verfahrenstest (SV)
- Eigentümerbefragung (mit dem Adressen-Berichtskreis der Stichprobe SV)

gibt es unterschiedliche Antwortausfälle. Deshalb wird die beschriebene Analyse und Behandlung der Antwortausfälle getrennt für jede dieser Erhebungen durchgeführt.

Antwortausfälle bei der Gebäude- und Wohnungszählung (Stichprobe SV) und der Haushaltsbefragung werden ersetzt, vergleichbar der Vorgehensweise in einem herkömmlichen Zensus (sogenannte Ersatzvornahmen). In der GWZ werden ausnahmslos bei allen Antwortausfällen Ersatzvornahmen erzeugt. Dagegen werden in der Haushaltsbefragung (SR und SV) nicht bei allen Antwortausfällen Ersatzvornahmen erzeugt: Nur bei vermeintlich paarigen Fällen, d.h. bei Personen, die wohnhaft und im Register sind, können ersatzweise Registerdaten für fehlende Angaben übernommen werden, aber nicht bei vermeintlichen Untererfassungen.

---

Bei den Auswertungen der Testerhebungen ist zu unterscheiden zwischen Ergebnisvergleichen und Einzeldatenvergleichen: **Einzeldatenvergleiche** sind Auswertungen, bei denen zwei verschiedene Quellen (z.B. Register und Haushaltsbefragung oder Haushaltsbefragung und Eigentümerbefragung) auf Basis der Einzeldaten miteinander verknüpft werden müssen. Bei **Ergebnisvergleichen** werden dagegen Aggregate verglichen; hierfür ist keine Verknüpfung auf Einzeldatenbasis erforderlich. Grundsätzlich werden Ergebnisvergleiche mit Ersatzvornahmen durchgeführt. Für Ergebnisvergleiche werden die Ersatzvornahmen daher für die Berechnung der Korrekturfaktoren also wie Antwortende behandelt und nicht als Antwortausfälle gezählt. Dagegen müssen Einzeldatenvergleiche ohne Ersatzvornahmen durchgeführt werden und dementsprechend werden Ersatzvornahmen bei Einzeldatenvergleichen als Antwortausfälle gezählt. Ausgenommen von dieser Regel sind Auswertungen der generierten Haushalte auf Einzeldatenbasis, bei denen Ersatzvornahmen von Antwortausfällen, die vom Interviewer als zweifelsfrei wohnhaft erkannt wurden, einbezogen werden. Für alle Auswertungen erfolgt die Korrektur der Antwortausfälle nach obigem Schema, lediglich die Menge der Antwortausfälle ist immer eine andere: Bei der Lieferung der Registerdaten durch die Gemeinden und die Bundesanstalt für Arbeit traten keine Antwortausfälle auf. Für Auswertungen aus den Registern für Ergebnisvergleiche wird also kein Korrekturfaktor für Antwortausfälle benötigt. Insgesamt werden für folgende Auswertungsmassen Korrekturfaktoren gebildet:

#### *Ergebnisvergleich*

- Auswertungen einschließlich aller Ersatzvornahmen in der Haushaltsbefragung SR
- Auswertungen einschließlich aller Ersatzvornahmen in der Haushaltsbefragung SV

#### *Einzeldatenvergleich*

- Auswertungen ohne Ersatzvornahmen in der Haushaltsbefragung SR
- Auswertungen ohne Ersatzvornahmen in der Haushaltsbefragung SV
- Auswertungen einschließlich Ersatzvornahmen zweifelsfrei wohnhafter Personen in der Haushaltsbefragung SV
- Auswertungen ohne Ersatzvornahmen in der Haushaltsbefragung SV *und* in der Eigentümerbefragung

Bei der Eigentümerbefragung wird ausnahmslos für alle Antwortausfälle Ersatz vorgenommen. Für Ergebnisvergleiche muss daher bei den Daten dieser Erhebung kein Korrekturfaktor vorgesehen werden.

### **2.3.2.2 Behandlung der Antwortausfälle beim Registertest**

Hauptzweck der Stichprobe Registertest ist die Schätzung der Über- und Unterfassungsrate des Registers. Da man bei einem sehr großen Teil der Antwortausfälle - bei allen mit Ausnahme der ungeklärten Fälle<sup>6)</sup> - weiß, ob sie mit dem Melderegister paarig oder Über- oder Untererfassungen sind, kann man diese zusätzliche Information bei der Korrektur der Antwortausfälle nutzen.

---

6) Ein ungeklärter Fall ist ein Fall, dessen Status – paarig, Über- oder Untererfassung – wegen Antwortausfällen im Gebäude nicht mit der erforderlichen Sicherheit festgestellt werden konnte.

Wie in 2.3.2.1 erfolgt die Hochrechnung gegliedert nach aus Gebäudetypen gebildeten Antwortklassen. Für eine weitere Gliederung nach paarigen Fällen/Übererfassungen/Untererfassungen müssen zunächst die ungeklärten Fälle geschätzt werden. Dies kann durch die Verwendung einer Art "Dual System Estimator"<sup>7)</sup> geschehen.

Kernstück dieser Schätzung ist folgende Vierfeldertafel:

	Zahl der Personen in der Haushaltsbefragung	
	nicht antwortend, aber geklärt Fall und wohnhaft	ungeklärt und wohnhaft
Im Register <sup>8)</sup>		
Nicht im Register		
Zusammen		

Man kann davon ausgehen, dass der Interviewer immer mit ausreichender Sicherheit feststellen kann, ob eine Wohnung bewohnt ist oder nicht, d.h. unter den Antwortausfällen verbergen sich keine unerkannten stichprobenneutralen Ausfälle. Somit handelt es sich bei Antwortausfällen immer um wohnhafte Personen. Auf Personenebene wird die Zahl der Antwortausfälle in einer Wohnung unter Verwendung aller vorhandenen Informationen (Verknüpfung von Befragungsdaten mit Registerdaten, Interviewerbeobachtungen) zum Teil geschätzt.

Bei nicht antwortenden geklärt Fällen bedeutet "im Register" ein paariger Fall, "nicht im Register" eine Untererfassung.

Da auf Gebäudeebene die Tabellenfelder teilweise nicht oder nur schwach besetzt sind, wird die Tabelle je Antwortklasse aggregiert; wegen der disproportionalen Ziehungswahrscheinlichkeiten erfolgt eine Hochrechnung mit den Faktoren der freien Hochrechnung.

	Zahl der Personen in der Haushaltsbefragung in der Antwortklasse $s$ hochgerechnet	
	nicht antwortend, aber geklärt Fall und wohnhaft	ungeklärt und wohnhaft
Im Register	$\hat{N}_{11}^{(s)}$	$\hat{N}_{12}^{(s)}$
Nicht im Register	$\hat{N}_{21}^{(s)}$	$\hat{N}_{22}^{(s)}$
Zusammen	$\hat{N}_{+1}^{(s)}$	$\hat{N}_{+2}^{(s)}$

Die grau unterlegten Felder sind zunächst unbekannt, während  $\hat{N}_{11}^{(s)}$ ,  $\hat{N}_{21}^{(s)}$ ,  $\hat{N}_{+1}^{(s)}$  und  $\hat{N}_{+2}^{(s)}$  aus der Stichprobe hochgerechnet werden können:

7) Ein "Dual System Estimator" verwendet Informationen aus zwei verschiedenen Quellen. Er wird beispielsweise in den USA vom Census Bureau zur Korrektur der Zensus-Untererfassung eingesetzt.

8) Hier und im folgenden bedeutet "im Register" stichtagsrelevant im Melderegister (ermittelt aus der Verknüpfung der Datenlieferung zum Stichtag 5. Dezember 2001 und der zweiten Lieferung zum 31. März 2002).

$$\hat{N}_{11}^{(s)} = \sum_{j \in S} w_j n_{11}^{(j)} \quad (2.3.7 \text{ a})$$

$$\hat{N}_{21}^{(s)} = \sum_{j \in S} w_j n_{21}^{(j)} \quad (2.3.7 \text{ b})$$

$$\hat{N}_{+1}^{(s)} = \hat{N}_{11}^{(s)} + \hat{N}_{21}^{(s)} \quad (2.3.7 \text{ c})$$

$$\hat{N}_{+2}^{(s)} = \sum_{j \in S} w_j n_{+2}^{(j)} \quad (2.3.7 \text{ d})$$

mit

$n_{11}^{(j)}$  : Zahl der Personen im Stichprobengebäude  $j$ , die in der Haushaltsbefragung nicht geantwortet haben, aber dort wohnen *und* die im Register sind,

$n_{21}^{(j)}$  : Zahl der Personen im Stichprobengebäude  $j$ , die in der Haushaltsbefragung nicht geantwortet haben aber dort wohnen und *nicht* im Register sind,

$n_{+2}^{(j)}$  : Zahl der ungeklärten Fälle im Stichprobengebäude  $j$  (Personen).

Für die Schätzung der grau unterlegten Felder der Tabelle wird angenommen, dass in einer Antwortklasse bei den ungeklärten Fällen dieselben Strukturen wie bei den nichtantwortenden geklärten Fällen bestehen, insbesondere was den Anteil der registrierten Personen betrifft:

$$\hat{N}_{12}^{(s)} = \hat{N}_{+2}^{(s)} \frac{\hat{N}_{11}^{(s)}}{\hat{N}_{+1}^{(s)}} \quad (2.3.8 \text{ a})$$

$$\hat{N}_{22}^{(s)} = \hat{N}_{+2}^{(s)} \frac{\hat{N}_{21}^{(s)}}{\hat{N}_{+1}^{(s)}} \quad (2.3.8 \text{ b})$$

Bei zu geringen Fallzahlen der nichtantwortenden geklärten Fällen müsste man auf die Strukturen der geklärten Fälle insgesamt (Antwortende und Nichtantwortende) zurückgreifen. Die in diesem Fall dann zugrundeliegende Annahme wäre allerdings etwas problematisch, da das Antwortverhalten vermutlich nicht unabhängig von dem Merkmal der Registrierung ist: bei nicht gemeldeten Personen ist die Antwortwahrscheinlichkeit vielleicht niedriger als bei gemeldeten Personen.

$\hat{N}_{12}^{(s)}$  ist die geschätzte Zahl der paarigen Fälle unter den ungeklärten Fällen und  $\hat{N}_{22}^{(s)}$  die geschätzte Untererfassung unter den ungeklärten Fällen.

Die Zahl der paarigen Fälle  $Y_1^{(s)}$ , der Übererfassungen  $Y_2^{(s)}$  und der Untererfassungen  $Y_3^{(s)}$  in der Antwortklasse  $s$  insgesamt können nun – hochgerechnet aus der Stichprobe unter Berücksichtigung der ungeklärten Fälle –

ermittelt werden: Die Zahl der paarigen Fälle bzw. der Untererfassungen insgesamt ergibt sich jeweils als Summe aus den frei hochgerechneten geklärten Fällen und den geschätzten ungeklärten Fällen:

$$Y_1^{(s)} = \sum_{j \in S} w_j y_{1j} + \hat{N}_{12}^{(s)} \quad (2.3.9 \text{ a})$$

$$Y_3^{(s)} = \sum_{j \in S} w_j y_{3j} + \hat{N}_{22}^{(s)} \quad (2.3.9 \text{ b})$$

mit

$y_{1j}$ : Zahl der als paarig klassifizierten Personen unter der Stichprobenadresse  $j$ ,

$y_{3j}$ : Zahl der als Fehlbestand klassifizierten Personen unter der Stichprobenadresse  $j$ .

Die Zahl der Übererfassungen  $Y_2^{(2)}$  ist dann die Differenz zwischen dem (frei hochgerechneten) Registerbestand und den paarigen Fällen  $Y_1^{(2)}$ :

$$Y_2^{(s)} = \sum_{j \in S} w_j x_j - Y_1^{(s)} \quad (2.3.9 \text{ c})$$

mit

$x_j$ : Zahl der gemeldeten Personen am Stichtag unter der Stichprobenadresse  $j$ .

Die Auswertung des Registertests erfolgt mit den Antwortenden sowie den Ersatzvornahmen der zweifelsfrei wohnhaften Personen. Die aus dieser Auswertungsmasse frei hochgerechnete Zahlen der paarigen Fälle  $\hat{Y}_1^{(s)}$ , der Übererfassungen  $\hat{Y}_2^{(s)}$  und der Untererfassungen  $\hat{Y}_3^{(s)}$  sind

$$\hat{Y}_1^{(s)} = \sum_{j \in S} w_j y_{1j,A} \quad (2.3.10 \text{ a})$$

$$\hat{Y}_2^{(s)} = \sum_{j \in S} w_j y_{2j,A} \quad (2.3.10 \text{ b})$$

$$\hat{Y}_3^{(s)} = \sum_{j \in S} w_j Y_{3j,A} \quad (2.3.10 \text{ c})$$

mit

- $Y_{1j,A}$  : Zahl der als paarig klassifizierten, zur Auswertungsmasse gehörenden Personen im Stichprobengebäude  $j$ ,
- $Y_{2j,A}$  : Zahl der als Übererfassung klassifizierten, zur Auswertungsmasse gehörenden Personen im Stichprobengebäude  $j$ ,
- $Y_{3j,A}$  : Zahl der als Untererfassung klassifizierten, zur Auswertungsmasse gehörenden Personen unter der Stichprobenadresse  $j$ .

Die Hochrechnungsfaktoren werden nun so korrigiert, dass sich die oben unter Berücksichtigung der Antwortausfälle geschätzten Werte

- $Y_1^{(s)}$  für die paarigen Fälle,
- $Y_2^{(s)}$  für die Übererfassungen und
- $Y_3^{(s)}$  für die Untererfassungen

ergeben (jeweils in der Antwortklasse  $s$ ). Der Korrekturfaktor für Antwortausfälle  $c_{mj}$  für Personen des Typs  $m$  ( $m=1$ : paariger Fall,  $m=2$ : Überfassung,  $m=3$ : Unterfassung) unter der Adresse  $j$  ergibt sich aus dem Verhältnis des unter Berücksichtigung der Antwortausfälle geschätzten Werts zu dem ohne Berücksichtigung der Antwortausfälle geschätzten Wert:

$$c_{mj} = \frac{Y_m^{(s)}}{\hat{Y}_m^{(s)}} \quad (m = 1, \dots, 3) \quad (2.3.11)$$

wobei  $s$  die Antwortklasse ist, zu der die Adresse  $j$  gehört.

Der Korrekturfaktor für Antwortausfälle wird dann den Personen zugeordnet; der Korrekturfaktor  $c_k$  einer Person  $k$  ist gleich  $c_{mj}$ , wobei  $m$  der Typ und  $j$  die Adresse der Person ist. Zu beachten ist, dass die Einsetzung anhand des Typs erfolgen muss, wie er hier definiert wurde, unabhängig davon, wie für die Auswertungen Übererfassungen und Untererfassungen definiert werden<sup>9)</sup>.

9) Beispielsweise ist bei einer am Wohnungsstatus ausgerichteten Definition eine Person, die mit Nebenwohnsitz registriert ist aber in der Haushaltsbefragung angibt, dort hauptsächlich zu wohnen, eine Untererfassung bezüglich des Hauptwohnsitzes. Sie gehört jedoch bei der Hochrechnung zum Typ "paariger Fall".

---

### 2.3.3 Anpassung an die Gebäudestruktur einschließlich der Mikrozensus-Adressen

Um eine Mehrfachbelastung von Haushalten zu vermeiden, wurden Adressen mit Haushalten, die am Mikrozensus 2001 teilgenommen haben („MZ-Adressen“) in der Auswahlgrundlage der Adressenauswahl gekennzeichnet und erhielten keine Auswahlchance, soweit sie weniger als 500 gemeldete Personen hatten. Da die hochgerechneten Ergebnisse die MZ-Adressen – im Gegensatz zu Sondergebäuden wie Anstalten – einschließen sollen, müssen die MZ-Adressen zugeschätzt werden, da sie zur Zielgesamtheit gehören.

Adressen mit vielen gemeldeten Personen enthalten mit größerer Wahrscheinlichkeit einen MZ-Haushalt als Adressen mit wenigen Personen, deshalb ist die Struktur der Adressengrößen in der resultierenden Stichprobe eine andere als in einer Stichprobe, die MZ-Adressen einschließen würde: die Stichprobe ist in Richtung der „kleinen“ Adressen verzerrt. Da in den Stichprobengemeinden die Struktur der Adressengrößen einschließlich der MZ-Haushalte bekannt ist, sollte hieran angepasst werden, um die Verzerrung zu kompensieren. In einer feinen Adressengrößenklassengliederung<sup>10)</sup> werden die Zahl der Adressen einschließlich der MZ-Adressen der Zahl ohne MZ-Adressen gegenübergestellt und daraus ein Anpassungsfaktor berechnet. Die Anzahl und Abgrenzung der Adressengrößenklassen muss noch festgelegt werden. Wegen zu geringer Fallzahlen in einer einzelnen Stichprobengemeinde sollte der Faktor aus nach Gemeindegrößenklassen aggregierten Daten gebildet werden; dafür sind in jeder Adressengrößenklasse die beiden Adressenzahlen zunächst hochzurechnen mit dem Gemeindefaktor (da es sich um Gemeindedaten handelt).

Der Anpassungsfaktor für fehlende MZ-Adressen  $g_j^{(1)}$  für die Stichprobenadresse  $j$ , die in einer Gemeinde der Gemeindegrößenklasse  $d$  zur Anpassungsklasse  $r$  gehört, ist

$$g_j^{(1)} = \frac{\sum_{i \in S_d} w_i^{(1)} A_{ri}}{\sum_{i \in S_d} w_i^{(1)} B_{ri}} \quad (2.3.12)$$

wobei

$A_{ri}$ : Anzahl der Adressen ohne Anstalten, Kasernen, in der Gebäudegrößenklasse  $r$  ( $r = 1, \dots, R$ ) in der Stichprobengemeinde  $i$ ,

$B_{ri}$ : Anzahl der Adressen ohne Anstalten, Kasernen und ohne MZ-Adressen mit weniger als 500 gemeldeten Personen, in der Gebäudegrößenklasse  $r$  in der Stichprobengemeinde  $i$ .

$S_d$ : Menge der Stichprobengemeinden der Gemeindegrößenklasse  $d$  ( $d = 1, \dots, 4$ ).

Die vier Gemeindegrößenklassen werden in Anlehnung an die Schichtung gebildet:

- Gemeinden mit weniger als 10 000 gemeldeten Personen,
- Gemeinden von 10 000 bis unter 50 000 gemeldeten Personen,
- Gemeinden von 50 000 bis unter 800 000 gemeldeten Personen,
- Gemeinden mit 800 000 oder mehr gemeldeten Personen.

---

10) Größe gemessen an der Zahl der mit Haupt- und Nebenwohnsitz gemeldeten Personen.

---

## 2.3.4 Anpassung an Rahmendaten

Durch eine Anpassung an Eckwerte der Grundgesamtheit kann die Qualität der Stichprobenergebnisse oftmals verbessert werden. Die Ziehung der Stichprobengemeinden basierte auf den Daten der Bevölkerungsfortschreibung. Eine Ausrichtung an den Registerdaten wäre – wenn sie verfügbar gewesen wären – entsprechend den Zielsetzungen der Testerhebungen sinnvoller gewesen. Die Adressenauswahl erfolgte zwar anhand der Registerdaten, allerdings zu einem vor dem Stichtag liegenden Zeitpunkt. Aus diesen Gründen ist eine nachträgliche Anpassung der Stichprobe an Registerdaten wünschenswert.<sup>11)</sup>

Als Rahmendaten stehen die Zahl der am Stichtag 5. Dezember 2001 gemeldeten Personen<sup>12)</sup> nach Deutschen und Ausländern sowie nach Haupt- und Nebenwohnung in regionaler Gliederung bis auf Gemeindeebene prinzipiell zur Verfügung. Wir verwenden eine Größenklassengliederung der Gemeinden. Die Abgrenzung der Größenklassen erfolgt nach der Zahl der (mit Haupt- oder Nebenwohnsitz) gemeldeten Personen. Diese Abgrenzung ist der sonst üblichen Abgrenzung nach der fortgeschriebenen Bevölkerung vorzuziehen, da hier eine – wenn auch vielleicht nur geringfügig – höhere Korrelation mit den Erhebungsmerkmalen bestehen dürfte<sup>13)</sup>. Die Gemeinden sind hierfür nach dem gelieferten Gemeinde-Eckwert zu typisieren. In Anlehnung an die Schichtung werden folgende Größenklassen gebildet:

- Gemeinden mit weniger als 10 000 gemeldeten Personen,
- Gemeinden von 10 000 bis unter 50 000 gemeldeten Personen,
- Gemeinden von 50 000 bis unter 800 000 gemeldeten Personen,
- Gemeinden mit 800 000 oder mehr gemeldeten Personen.

Für Personenauswertungen ist darüber hinaus eine Anpassung nach der Nationalität (deutsch/ausländisch) unbedingt sinnvoll, da die Nationalität bei der Auswahl nicht berücksichtigt worden war<sup>14)</sup>. Eine theoretisch denkbare Anpassung an Haupt- und Nebenwohnung ist nicht vorgesehen, da die Zahl der Nebenwohnungen in der Stichprobe voraussichtlich gering ist und eine stabile Hochrechnung nicht zulässt.

Es gibt allerdings ein Abgrenzungsproblem: Die Zielgesamtheit der Testerhebungen ist die wohnberechtigte Bevölkerung ohne Personen in Anstalten<sup>15)</sup>, militärischen Gebäuden, und ohne Adressen, in denen sich eine diplomatische Vertretung befindet<sup>16)</sup>. Ebenfalls nicht zur Zielgesamtheit gehören Sammeladressen (pro forma gemeldete Personen, z.B. unter der Adresse des Sozialamts gemeldete Obdachlose). Anstalten, militärische Gebäude, diplomatische Vertretungen und Sammeladressen werden – soweit bekannt – in der Auswahlgrundlage gekennzeichnet und erhalten keine Auswahlchance. Da die Eckwerte die ausgeschlossenen Personenkreise aber umfassen, würde man bei einer Anpassung an die Eckwerte implizit für diese nicht in die Auswahl einbezogenen Adressen dieselben Strukturen hinsichtlich der Erhebungsmerkmale wie bei den Stichprobenadressen unterstellen. Deshalb werden zunächst aus den Eckwerten die Bevölkerung in diesen Sonderadressen herausgerechnet. Die

---

11) Eine Anpassung an Daten der Bevölkerungsfortschreibung erscheint nicht sinnvoll.

12) Zu beachten ist, dass „gemeldet“ sich hier im Gegensatz zu Kapitel 2.3.2.2 nur auf die erste Datenlieferung zum Stichtag 5.12.2001 bezieht.

13) Auch in den Ergebnistabellen werden bezüglich der Zahl der gemeldeten Personen abgegrenzte Gemeindegroßenklassen verwendet.

14) Bei der Auswahl wäre eine Schichtung oder Sortierung der Gemeinden nach dem Ausländeranteil der Gemeinden (auf der ersten Auswahlstufe) bzw. der Gebäude (auf der zweiten Auswahlstufe) denkbar gewesen. Diese konnte nicht durchgeführt werden, weil keine geeigneten Daten zur Verfügung standen.

15) Wohnheime, z.B. Studentenwohnheime, Schwesternwohnheime, Durchgangslager oder Asylbewerberheime zählen hier nicht zu den Anstalten.

16) Der Ausschluss aus der Zielgesamtheit erfolgt adressenweise, d.h. Privathaushalte unter der Adresse einer Anstalt beispielsweise gehören auch nicht zur Zielgesamtheit.

Anpassung erfolgt also an die Zahl der am Stichtag gemeldeten Personen ohne in Anstalten, militärischen Adressen, diplomatischen Vertretungen und Sammeladressen gemeldete Personen.

Da diese Sonderadressen im Material der gelieferten Adressen gekennzeichnet wurden, kann die Zielgesamtheit mit Hilfe eines aus der Gemeindestichprobe hochgerechneten Anteils dieser Sonderadressen geschätzt werden. Die Schätzung erfolgt gegliedert nach Gemeindegrößenklassen und der Nationalität. Da die Kennzeichnung der Sonderadressen im Adressenmaterial möglicherweise nicht vollständig ist, sind von den Eckwerten nicht nur eine geschätzte Zahl von in gekennzeichneten Sonderadressen gemeldeten Personen, sondern auch von gemeldeten Personen in Sonderadressen, die erst bei der Haushaltsbefragung erkannt werden, zu subtrahieren. Gekennzeichnete Sonderadressen sind nur mit dem Gemeindefaktor, erkannte Sonderadressen mit Gemeinde- und Adressenfaktor und dem Anpassungsfaktor für fehlende MZ-Adressen hochzurechnen. Allerdings liegt die Zahl der gemeldeten Personen in gekennzeichneten Sonderadressen nicht in der Untergliederung nach Nationalität vor. Wir nehmen deshalb in militärischen Adressen nur Deutsche und in diplomatischen Vertretungen nur Ausländer an, in Anstalten und Sammeladressen unterstellen wir dasselbe Verhältnis von Deutschen und Ausländern wie in der Grundgesamtheit. Die als Hochrechnungsrahmen zu verwendenden Eckwerte  $X_{d\ell}^{(ur)}$  ergeben sich dann als

$$X_{d\ell}^{(ur)} = \begin{cases} X_{d\ell}^{(ur)} (1 - P_d^{(1)}) - X_{d+}^{(ur)} P_d^{(2)} - \sum_{j \in S_d} w_j g_j^{(1)} z_{ij} & \text{für } \ell = \text{deutsch} \\ X_{d\ell}^{(ur)} (1 - P_d^{(1)}) - X_{d+}^{(ur)} P_d^{(3)} - \sum_{j \in S_d} w_j g_j^{(1)} z_{ij} & \text{für } \ell = \text{ausländisch} \end{cases} \quad (2.3.13)$$

mit

$z_{ij}$ : Zahl der am Stichtag gemeldeten Personen der Nationalität  $\ell$  unter der Stichprobenadresse  $j$ , das in der Befragung als Sonderadresse erkannt wurde ( $z_{ij} = 0$  für andere Stichprobenadressen)

$X_{d\ell}^{(ur)}$ : Zahl der am Stichtag gemeldeten Deutschen ( $\ell = 1$ ) bzw. Ausländer ( $\ell = 2$ ) in der Gemeindegrößenklasse  $d$  insgesamt

$X_{d+}^{(ur)}$ : Zahl der am Stichtag gemeldeten Personen in der Gemeindegrößenklasse  $d$ ,

$$X_{d+}^{(ur)} = X_{d1}^{(ur)} + X_{d2}^{(ur)}$$

$P_d^{(m)}$ : Aus der Stichprobe Registertest hochgerechneter Anteil der in gekennzeichneten Sonderadressen des Typs  $m$  gemeldeten Personen an allen gemeldeten Personen in der Gemeindegrößenklasse  $d$  zum Zeitpunkt der Adressenlieferung. ( $m=1$ : Anstalten, Sammeladressen,  $m=2$ : Kasernen,  $m=3$ : Botschaften)

$$P_d^{(m)} = \frac{\sum_{i \in S_d} w_i^{(1)} z_i^{(m)}}{\sum_{i \in S_d} w_i^{(1)} v_i} \quad (m = 1, 2, 3) \quad (2.3.14)$$

---

$Z_i^{(m)}$ : Zahl der gemeldeten Personen in Sondergebäuden des Typs  $m$  in der Gemeinde  $i$  zum Zeitpunkt der Adressenlieferung

$V_i$ : Gesamtzahl aller gemeldeten Personen in der Gemeinde  $i$  zum Zeitpunkt der Adressenlieferung

Die Summationen in den Formeln (2.3.13) und (2.3.14) erfolgen über alle Stichprobengemeinden  $i$  bzw. Stichprobengebäude  $j$ , die zur Gemeindegrößenklasse  $d$  gehören.

Die Zahl der gemeldeten Personen am Stichtag kann nun aus der Stichprobe mit den Kehrwerten der Ziehungswahrscheinlichkeiten hochgerechnet und mit dem Eckwert aus Formel (2.3.13) verglichen werden:

Die aus der Stichprobe hochgerechnete Zahl der gemeldeten Personen ohne Personen in Sonderadressen in der Gemeindegrößenklasse  $d$  und der Nationalität  $\ell$  ist

$$\hat{X}_{d\ell}^* = \sum_{k \in S_{d\ell}} g_k^{(1)} c_k w_k x_k \quad (2.3.15)$$

mit

$$x_k = \begin{cases} 1, & \text{falls Person } k \text{ am Stichtag gemeldet und kein Antwortausfall ist,} \\ 0 & \text{sonst} \end{cases}$$

$S_{d\ell}$ : Menge aller Personen der Nationalität  $\ell$  in Gemeinden der Gemeindegrößenklasse  $d$  ohne Sonderadressen

In der Regel wird der Eckwert durch die Hochrechnung nicht getroffen. Aus der Abweichung wird ein Anpassungsfaktor  $g_{ij}^{(2)}$  für Personen der Nationalität  $\ell$  im Gebäude  $j$  konstruiert:

$$g_{ij}^{(2)} = \frac{X_{d\ell}}{\hat{X}_{d\ell}^*} \quad (2.3.16)$$

mit

$d$ : Größenklasse der Gemeinde, zu der die Adresse  $j$  gehört.

Auf Personenebene ist dann der Faktor  $g_k^{(2)}$  für eine Person  $k$  gleich  $g_{ij}^{(2)}$ , wenn die Person die Nationalität  $\ell$  hat und zur Adresse  $j$  gehört.

Bei denjenigen Personensätzen, die im Register vorhanden sind, erfolgt die Zuordnung des Anpassungsfaktors über die Nationalität gemäß den Registerdaten (erste Datenlieferung) zum Stichtag. Der Anpassungsfaktor gilt streng genommen nur für diese Sätze. Er kann unseres Erachtens aber auch auf Personensätze, für die nur Informationen aus der zweiten Datenlieferung der Gemeinden oder aus der Haushaltebefragung oder der Eigentums-

---

befragung vorliegen, angewendet werden. Die Zuordnung erfolgt dann gemäß der in diesen Quellen festgestellten Nationalität.

Für Haushalts- und Wohnungsauswertungen (mit Wohnung oder Haushalt als Darstellungseinheit  $k$ ) erfolgt die Anpassung ohne die Untergliederung nach der Nationalität. Für die Berechnung des Anpassungsfaktors  $g_k^{(2)}$  werden in Formel (2.3.16) in Zähler und Nenner anstelle der Werte für die Nationalität  $\ell$  die entsprechenden Summen über beide Ausprägungen der Nationalität genommen.

## 2.4 Abschätzung der Stichprobenzufallsfehler

Um Aussagen über die Qualität der Ergebnisse treffen zu können, ist insbesondere eine Abschätzung der Präzision der Ergebnisse (Varianzschätzung) erforderlich. Die Präzision der Ergebnisse wird anhand der Stichprobenzufallsfehler beurteilt, deren Größenordnung mit Hilfe der relativen Standardfehler aus der Stichprobe abgeschätzt werden kann.

Für den Schätzer  $Y$  des Totalwerts eines Erhebungsmerkmals (siehe Formel 2.3.1) wird zunächst ein Varianzschätzer hergeleitet, der die Anpassung an Rahmendaten noch nicht berücksichtigt. Dieser ist die Basiskomponente der Varianzschätzung, aus der dann mit Hilfe einer Linearisierung durch Taylorentwicklung die eigentlichen Varianzschätzer abgeleitet werden.

### 2.4.1 Varianzschätzung ohne Anpassung an Rahmendaten

Der Schätzer für den Totalwert  $Y$  ist ohne Berücksichtigung des Anpassungsfaktors

$$\hat{Y}_0 = \sum_{k \in S} g_k^{(1)} c_k w_k y_k \quad (2.4.1)$$

(Bezeichnungen siehe Formel 2.3.1).

Die Varianzschätzung bei größenproportionaler Auswahl ohne Zurücklegen ist – insbesondere bei der verwendeten Auswahltechnik - sehr schwierig, da die Ziehungswahrscheinlichkeiten zweiter Ordnung (die Wahrscheinlichkeit, dass zwei bestimmte Einheiten in die Stichprobe gelangen) nicht berechnet werden können. Als Näherungslösung wird ein Varianzschätzer für eine Auswahl auf der ersten Stufe *mit* Zurücklegen verwendet. Dies führt zu einer konservativen Schätzung, d.h. zu etwas überhöhten Varianzen. Um dies zu kompensieren, kann man die Endlichkeitskorrektur, wie sie von einer geschichteten einfachen Zufallsauswahl ohne Zurücklegen bekannt ist, verwenden (siehe etwa Wolter 1985, Kapitel 7.7, S. 288). Die systematische Auswahl aus sortiertem Material hat näherungsweise den Effekt einer feinen Schichtung. Der Sortierungseffekt wird dadurch berücksichtigt, dass für die Varianzschätzung eine zusätzliche Schichtung innerhalb jeder Gemeindegrößenklassen-Schicht unterstellt wird. Diese zusätzlichen Schichten werden gebildet, indem jeweils eine konstante Anzahl (5 bis 10) von in der Sortierung aufeinanderfolgenden Gemeinden zu einer Schicht zusammengefasst werden.

Dieser Ansatz ist allerdings nur für diejenigen Schichten möglich, die nicht vollständig erfasst wurden („Repräsentativschichten“).

Die Varianz des auf die Repräsentativschichten eingeschränkten Schätzers

$$\hat{Y}_{0,R} = \sum_{k \in S_R} g_k^{(1)} c_k w_k y_k \quad (2.4.2)$$

mit

$S_R$ : Stichprobe in den Repräsentativschichten

ist dann näherungsweise

$$V(\hat{Y}_{0,R}) = \sum_{h \in H_R} \left( 1 - \frac{\sum_{i=1}^{n_h} z_{hi}}{Z_h} \right) \frac{Z_h}{n_n} \sum_{i=1}^{N_h} z_{hi} \left( \frac{y_{hi}}{z_{hi}} - \frac{Y_h}{Z_h} \right)^2 + \sum_{h \in H_R} \frac{Z_h}{n_h} \sum_{i=1}^{n_h} \frac{\sigma_{hi}^2}{z_{hi}} \quad (2.4.3)$$

mit

$H_R$ : Menge der Repräsentativschichten

$y_{hi}$ : Summe des Erhebungsmerkmals  $y$  in der Gemeinde  $i$  der Schicht  $h$  in der Grundgesamtheit

$Y_h$ : Totalwert des Merkmals in der Schicht  $h$ ,  $Y_h = \sum_{i=1}^{N_h} y_{hi}$

$N_h$ : Zahl der Gemeinden in der Schicht  $h$

$n_h$ : Zahl der Stichprobengemeinden in der Schicht  $h$

$z_{hi}$ : Zahl der Einwohner gemäß Auswahlgrundlage (fortgeschriebene Bevölkerung) in der Gemeinde  $i$  der Schicht  $h$

$Z_h$ : Zahl der Einwohner in der Schicht  $h$ ,  $Z_h = \sum_{i=1}^{N_h} z_{hi}$

$\sigma_{hi}^2$ : Bedingte Varianz<sup>17)</sup> des auf die Gemeinde  $i$  (d.h. von der zweiten auf die erste Stufe) hochgerechneten Totalwerts des Erhebungsmerkmals

Diese Varianz kann aus der Stichprobe erwartungstreu geschätzt werden durch

$$\hat{V}(\hat{Y}_{0,R}) = \sum_{h \in H_R} \left( 1 - \frac{\sum_{i=1}^{n_h} z_{hi}}{Z_h} \right) \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} \left( \frac{\hat{Y}_{0,hi} z_{hi}}{z_{hi}} - \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{Y}_{0,hi} z_{hi}}{z_{hi}} \right)^2 \quad (2.4.4)$$

17) Varianz unter der Bedingung, dass die Auswahl der Gemeinden schon erfolgt ist und die ausgewählten Gemeinden fest sind.

---

wobei

$$\hat{Y}_{0,hi} = \sum_{k \in S_{hi}} g_k^{(1)} c_k w_k^{(2)} y_k \quad (2.4.5)$$

der geschätzte Totalwert für die Gemeinde  $i$  ist

mit

$w_k^{(2)}$ : Kehrwert der Ziehungswahrscheinlichkeit auf der zweiten Auswahlstufe

$S_{hi}$ : Menge der Darstellungseinheiten der Stichprobe, die zur Gemeinde  $i$  der Schicht  $h$  gehören.

(siehe Raj 1968, Theorem 6.4, S. 120).

Formel (2.4.4) ist sehr elegant, da sie im Gegensatz zu Formel (2.4.3) nur eine Varianzkomponente enthält und nur die auf die erste Stufe hochgerechneten Totalwerte benötigt. Das Design auf der zweiten Auswahlstufe – systematische größenproportionale Auswahl – muss somit nicht gesondert berücksichtigt werden. Ebenso sind die Antwortausfälle implizit berücksichtigt: Die Gewährung von Antworten kann als Zufallsauswahl mit ungleichen Wahrscheinlichkeiten auf zwei zusätzlichen Auswahlstufen angesehen werden, wobei die Ziehungswahrscheinlichkeit die geschätzte Antwortwahrscheinlichkeit ist bzw. die Wahrscheinlichkeit, dass eine ausgewählte Adresse keine Mikrozensus-Adresse ist. Formel (2.4.4) gibt die aufsummierten, mit einer Endlichkeitskorrektur versehenen schichtweisen Standardabweichungen der Gemeindegewerte

$$u_{hi} = \frac{w_{hi}^{(1)}}{n_h} \hat{Y}_{0,hi}$$

mit  $w_{hi}^{(1)} = \frac{Z_h}{Z_{hi} \cdot n_h}$  als Kehrwert der Ziehungswahrscheinlichkeit der Gemeinde  $i$  auf der ersten Stufe wieder.

Der Varianzbetrag derjenigen Schichten, bei denen auf der ersten Stufe alle Gemeinden ausgewählt wurden („Totalschichten“) kann allerdings nicht auf diese Weise geschätzt werden, da in den Totalschichten keine zweistufige Auswahl, sondern in jeder Totalschicht-Gemeinde eine einstufige größenproportionale Auswahl von Adressen vorliegt. Für die Varianzschätzung wird wie auf der ersten Auswahlstufe eine Endlichkeitskorrektur verwendet und eine Schichtung nach Adressengrößenklassen unterstellt, um den Sortierungseffekt auf der zweiten Stufe zu berücksichtigen.

Für eine Gemeinde  $i$  aus einer Totalschicht führt das zu der Varianz

$$V(\hat{Y}_{0,hi}) = \sum_{\ell} \left( 1 - \frac{\sum_{j=1}^{m_{hi\ell}} z_{hi\ell j}}{Z_{hi\ell}} \right) \frac{Z_{hi\ell}}{m_{hi\ell}} \sum_{j=1}^{m_{hi\ell}} z_{hi\ell j} \left( \frac{y_{hi\ell j}}{z_{hi\ell j}} - \frac{Y_{hi\ell}}{Z_{hi\ell}} \right)^2 \quad (h \in H_T) \quad (2.4.6)$$

mit

$y_{hi\ell j}$ : Summe des Erhebungsmerkmals unter der Adresse  $j$  der Adressengrößenklasse  $\ell$  in der Gemeinde  $i$  der Schicht  $h$

$Y_{hi\ell}$ : Totalwert des Merkmals in der Schicht  $\ell$  der Grundgesamtheit,  $Y_{hi\ell} = \sum_{j=1}^{M_{hi\ell}} y_{hi\ell j}$

$M_{hi\ell}$ : Zahl der Adressen der Adressengrößenklasse  $\ell$  in der Grundgesamtheit

$m_{hi\ell}$ : Zahl der Adressen der Adressengrößenklasse  $\ell$  in der Stichprobe

$z_{hi\ell j}$ : Zahl der gemeldeten Personen zum Zeitpunkt der Adressenlieferung unter der Stichprobenadresse  $j$  der Adressengrößenklasse  $\ell$

$Z_{hi\ell}$ : Zahl der gemeldeten Personen der Adressengrößenklasse  $\ell$  in der Grundgesamtheit,

$$Z_{hi\ell} = \sum_{j=1}^{M_{hi\ell}} z_{hi\ell j}$$

$H_T$ : Menge der Gemeinde-Totalschichten

Diese Varianz kann aus der Stichprobe erwartungstreu geschätzt werden durch

$$\hat{V}(\hat{Y}_{0,hi}) = \sum_{\ell} \left( 1 - \frac{\sum_{j=1}^{m_{hi\ell}} z_{hi\ell j}}{Z_{hi\ell}} \right) \frac{1}{m_{hi\ell} - 1} \sum_{j=1}^{m_{hi\ell}} (y_{hi\ell j}^* - \bar{y}_{hi\ell}^*)^2 \quad (2.4.7)$$

mit

$$y_{hi\ell j}^* = \frac{1}{m_{hi\ell}} \sum_k g_k^{(1)} c_k w_k^{(2)} y_{i\ell jk}$$

$$\bar{y}_{hi\ell}^* = \frac{1}{m_{hi\ell}} \sum_{j=1}^{m_{hi\ell}} y_{hi\ell j}^*$$

(siehe Raj 1968, Theorem 3.8, S.48).

Insgesamt kann die Varianz von  $\hat{Y}_0 = \hat{Y}_{0,R} + \sum_{h \in H_T} \sum_i \hat{Y}_{0,hi}$  also geschätzt werden durch

$$\hat{V}(\hat{Y}_0) = \hat{V}(\hat{Y}_{0,R}) + \sum_{h \in H_T} \sum_i \hat{V}(\hat{Y}_{0,hi}) \quad (2.4.8)$$

Die Varianzschätzung für ein bestimmtes Tabellenfeld  $g$  erfolgt wie üblich dadurch, dass für das Merkmal  $Y$  die Merkmalswerte  $y_k$  außerhalb der Untergruppe  $g$  auf Null gesetzt werden. Die Varianzschätzung für die Unterstichprobe des Verfahrenstests erfolgt nach denselben Formeln: Da für die Unterstichprobenziehung keine zusätzliche Schichtung verwendet wurde, ist die Varianz dieselbe wie die für eine Unterstichprobe, die direkt aus der Auswahlgesamtheit gezogen worden ist.

Einen Sonderfall bilden einige Tabellen des Registertests, die Gemeinden gegliedert nach Größenklassen der Unter- und Übererfassungsraten darstellen. Die Zahl der auf eine bestimmte Untergruppe  $d$  entfallenden Gemeinden  $Y_d$  wird geschätzt durch

$$\hat{Y}_d = \sum_h \sum_{i \in S_{hd}} w_{hi}^{(1)} \quad (2.4.9)$$

wobei  $S_{hd}$  die Menge der Stichprobengemeinden in der Schicht  $h$  ist, die zur Gruppe  $d$  gehören und  $w_{hi}^{(1)}$  der Gemeindehochrechnungsfaktor der Gemeinde  $i$  ist.

Die Varianz von  $\hat{Y}_d$  kann dann näherungsweise geschätzt werden durch

$$\hat{V}(\hat{Y}_d) = \sum_h \left( 1 - \frac{\sum_{i=1}^{n_h} Z_{hi}}{Z_h} \right) \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (u_{hi} - \bar{u}_h)^2 \quad (2.4.10)$$

mit

$$u_{hi} = \begin{cases} w_{hi}^{(1)}, & \text{falls Gemeinde } i \text{ zur Gruppe } d \text{ gehört,} \\ n_h & \\ 0 & \text{sonst} \end{cases}$$

$$\bar{u}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} u_{hi}$$

(siehe Raj 1968, Theorem 3.8, S. 48).

Nicht berücksichtigt ist, dass die Zuordnung einer Gemeinde zur Gruppe  $d$  im Fall von Größenklassen der Fehleraten zufallsabhängig ist, da diese Raten auf der zweiten Stufe geschätzt sind. Da es nur eine sehr grobe Einteil-

lung in 3 Größenklassen gibt, kann diese zusätzliche Varianzkomponente wohl vernachlässigt werden. Für die Schätzung von Gemeindefallzahlen wird keine Anpassung an Bevölkerungseckwerte durchgeführt.

## 2.4.2 Varianzschätzung mit Berücksichtigung der Anpassung

Allgemein kann die Varianz einer Parameterschätzung  $\hat{\theta} = f(\hat{Y}_1, \dots, \hat{Y}_p)$ , die sich über eine Funktion  $f$  aus linearen Schätzungen von Totalwerten von  $p$  verschiedenen Merkmalen der Form

$$\hat{Y}_j = \sum_{k \in S} w_k y_{jk} \quad (j=1, \dots, p) \quad (2.4.11)$$

zusammensetzt, näherungsweise mit Hilfe einer Taylorentwicklung abgeschätzt werden, indem die Varianz eines analog zu Formel (2.4.8) geschätzten neuen Merkmals  $U$  mit

$$u_k = \sum_{j=1}^p \frac{\partial f(\hat{Y}_j)}{\partial y_j} y_{jk} \quad (k \in S) \quad (2.4.12)$$

berechnet wird (siehe Wolter 1985, Kapitel 6.5, S.230).

$\hat{Y}$  aus Formel (2.3.1) kann dargestellt werden als

$$\hat{Y} = \sum_g \frac{X_g}{\hat{X}_{0,g}} \hat{Y}_{0,g} \quad (2.4.13)$$

wobei

$X_g$  : Eckwert des Anpassungsmerkmals (Zahl der gemeldeten Personen) für die Anpassungsklasse  $g$

$\hat{X}_{0,g}$  : Schätzer für  $X_g$ , analog zu Formel (2.4.1)

$$\hat{X}_{0,g} = \sum_{k \in S_g} g_k^{(1)} c_k w_k x_k$$

$\hat{Y}_{0,g}$  : Schätzer für  $Y_g$  (Formel (2.4.1) für die Untergruppe  $g$ )

Dies führt auf

$$u_k = g_k^{(2)} \left( y_k - \frac{\hat{Y}_{0,g}}{\hat{X}_{0,g}} \right) \quad (k \in s_g) \quad (2.4.14)$$

Für eine Schätzung der Varianz von  $\hat{Y}$  können also auch die Formeln (2.4.4) und (2.4.7) verwendet werden. Es muss lediglich  $y_k$  durch  $u_k$  ersetzt werden.

### **Varianzschätzung von Verhältniswerten**

Sehr wichtig ist die Schätzung von Verhältniswerten oder Anteilen, z.B. die Über- und Untererfassungsraten im Registertest.

Ein Verhältniswert  $R = \frac{Y}{Z}$  wird geschätzt durch

$$\hat{R} = \frac{\hat{Y}}{\hat{Z}}$$

wobei Zähler- und Nennermerkmal gemäß Formel (2.3.1) hochgerechnet werden.

Die Schätzung der Varianz kann wieder – gemäß dem Ansatz aus Formel (2.4.12) - nach den Formeln (2.4.4) und (2.4.7) berechnet werden, wobei als Merkmal

$$u_k = \frac{g_k^{(2)}}{\hat{Z}} \left( y_k - \hat{R}z_k + \frac{\hat{R}\hat{Z}_{0,g} - \hat{Y}_{0,g}}{\hat{X}_{0,g}} \right) \quad (k \in s_g) \quad (2.4.15)$$

anstelle von  $y_k$  genommen wird.

### **Varianzschätzung von Differenzen**

Um die verschiedenen im Zensustest angewandten Maßnahmen beurteilen zu können, werden Totalwerte miteinander verglichen (z.B. wohnberechtigte Bevölkerung oder Übererfassungen vor und nach Durchführung der Haushalteinenerierung). Für Aussagen zur Signifikanz von Unterschieden braucht man die Varianz eines Differenzschätzers  $\hat{D} = \hat{Y} - \hat{Z}$ .

Die Varianz von  $\hat{D}$  kann nach den Formeln (2.4.4) und (2.4.7) geschätzt werden mit

$$u_k = g_k^{(2)} \left( y_k - z_k - \frac{\hat{Y}_{0,g} - \hat{Z}_{0,g}}{\hat{X}_{0,g}} \right) \quad (2.4.16)$$

anstelle von  $y_k$ .

---

## 3 Mehrfachfallprüfung

### 3.1 Auswahlverfahren

Die Mehrfachfallprüfung kann als Test auf Stichprobenbasis naturgemäß nicht durch eine Auswahl von Gemeinden wie beim Register- und Verfahrenstest erfolgen. Die Auswahl muss über ein unveränderliches Personenmerkmal erfolgen, damit etwaige mehrfache Registereinträge einer Person zusammen in die Stichprobe gelangen. Hierfür wurde das Verfahren der *Geburtstagsauswahl* eingesetzt: Als ausgewählt gelten alle Personen, die an bestimmten Tagen, nämlich am 1. Januar, 15. Mai oder 1. September geboren sind. Personen mit unvollständigem Geburtsdatum im Register werden alle in die Stichprobe einbezogen. Die Gemeinden lieferten für den Stichtag 5. Dezember 2001 Auszüge aus den Melderegistern für diese Personen (eine zweite Lieferung erfolgte für den Stichtag 31. März 2002). Personen in der Stichprobe mit mehr als einer alleinigen Wohnung oder Hauptwohnung oder nur mit Nebenwohnungen (die sogenannten Prüffälle) werden nach dem tatsächlichen Wohnort befragt. Bei Geburtstagsauswahlen liegt der Stichprobenumfang nicht a priori fest, sondern hängt vom Zufall ab. In die Stichprobe gelangten 971 037 Registersätze, das sind ca. 1,2% der Bevölkerung. Die Zahl der zu befragenden Prüffälle steht zur Zeit noch nicht fest. Gerechnet wird mit ca. 16 000 Personen, die allerdings oftmals über zwei Anschriften kontaktiert werden müssen.

### 3.2 Hochrechnungsverfahren

Analog zum Register- und Verfahrenstest setzt sich die Hochrechnung aus folgenden Komponenten zusammen:

#### *Ziehungswahrscheinlichkeiten*

Da nicht alle Registersätze dieselbe Wahrscheinlichkeit haben, in die Stichprobe zu gelangen (die Sätze mit unvollständigem Geburtsdatum sind in der Stichprobe überrepräsentiert), müssen diese Wahrscheinlichkeiten bei der Hochrechnung auf jeden Fall berücksichtigt werden.

#### *Behandlung von Antwortausfällen*

Bei der Befragung der Prüffälle ist trotz Auskunftspflicht mit Antwortausfällen zu rechnen. Wie beim Register- und Verfahrenstest werden für die Konstruktion von Korrekturfaktoren Antwortwahrscheinlichkeiten geschätzt.

#### *Anpassung an Rahmendaten*

Die Zahl der am Stichtag gemeldeten Personen in der Grundgesamtheit kann für die Hochrechnung genutzt werden.

Die Hochrechnung erfolgt grundsätzlich je Bundesland; in den Formeln wird der Übersichtlichkeit halber auf einen zusätzlichen Index für das Land verzichtet. Der Totalwert  $Y$  eines Erhebungsmerkmals wird geschätzt durch

$$\hat{Y} = \sum_{k \in S} g_k c_k w_k y_k \quad (3.2.1)$$

mit

- 
- $y_k$  : Merkmalswert der Person  $k$   
 $w_k$  : Design-Faktor (Hochrechnungsfaktor der freien Hochrechnung)  
 $c_k$  : Korrekturfaktor für Antwortausfälle  
 $g_k$  : Anpassungsfaktor aus der Verhältnisschätzung

### 3.2.1 Ziehungswahrscheinlichkeiten

Unter den Geburtstagen nimmt der 1. Januar eine Sonderstellung ein. Da in vielen Fällen bei unbekanntem Geburtsdatum der 1. Januar eingetragen wird, unterscheidet er sich in seiner Struktur eventuell von den übrigen Geburtstagen und enthält vermutlich deutlich mehr Mehrfachfälle. Eine Gleichbehandlung der 3 Stichproben-Geburtstage bei der Hochrechnung könnte dann zu verzerrten Ergebnissen führen. Da der 1. Januar bewusst ausgewählt wurde, stellt er eine total erfasste Teilmenge dar. Alle Registersätze mit Geburtsdatum 1. Januar haben somit die Ziehungswahrscheinlichkeit Eins. Dasselbe gilt für die Registersätze mit unvollständigem Geburtsdatum. Die Registersätze der übrigen Geburtstage (15. Mai und 1. September) können jedoch als Zufallsstichprobe aus der Gesamtheit der am Stichtag gemeldeten Personen ohne die am 1. Januar Geborenen und ohne Personen mit unvollständigem Geburtsdatum aufgefasst werden. Für alle diese Sätze ist die Ziehungswahrscheinlichkeit gleich  $2/364,25$ .

Für die Hochrechnung wird also entsprechend den unterschiedlichen Ziehungswahrscheinlichkeiten die Grundgesamtheit und die Stichprobe in 3 Teile gegliedert:

- Teil 1: Personen mit unvollständigem Geburtsdatum,
- Teil 2: Personen mit Geburtstag 1. Januar,
- Teil 3: Personen mit vollständigem Geburtsdatum ohne die am 1. Januar Geborenen.

Der Hochrechnungsfaktor der freien Hochrechnung  $w_k$  einer Person  $k$  ist als Kehrwert der Ziehungswahrscheinlichkeit also

$$w_k = \begin{cases} \frac{364,25}{2}, & \text{falls Person } k \text{ zum Stichprobenteil 3 gehört,} \\ 1, & \text{falls Person } k \text{ zum Stichprobenteil 1 oder 2 gehört.} \end{cases} \quad (3.2.2)$$

### 3.2.2 Behandlung der Antwortausfälle

Bei der Befragung der Prüffälle ist trotz Auskunftspflicht mit Antwortausfällen zu rechnen. Die Antwortausfälle werden nicht zufällig auftreten und können die Struktur der Stichprobe verzerren. Man kann nun versuchen, im

---

Rahmen der Hochrechnung die Verzerrung zu korrigieren. Die zu betrachtende Einheit ist hier der Prüffall (zu einem Prüffall können mehrere Registersätze gehören).

Es ist äußerst wichtig, zwischen Antwortausfällen (auch echte Ausfälle genannt) und sonstigen Ausfällen (sogenannte stichprobenneutrale Ausfälle oder unechte Ausfälle) zu unterscheiden.

*Stichprobenneutrale Ausfälle* sind Personen (Prüffälle), die nicht oder nicht mehr zur Zielgesamtheit der Mehrfachfallprüfung (d.h. zur wohnberechtigten Bevölkerung) gehören, z.B.

- Person verstorben,
- Person ins Ausland verzogen,
- Person definitiv unter keiner ihrer Meldeadressen wohnhaft.

Wenn einer dieser Sachverhalte zweifelsfrei festgestellt wurde, zählt dies als Befragungsergebnis und im stichprobenmethodischen Sinne als Antwort.

*Antwortausfälle* sind Personen (Prüffälle), die zur Zielgesamtheit der Mehrfachfallprüfung gehören, für die aber keine Antwort vorliegt, z.B.

- Person nicht erreichbar,
- Person verweigert Auskunft.

Manchmal ist nicht eindeutig zu erkennen, ob es sich um einen Antwort- oder einen stichprobenneutralen Ausfall handelt. Solche Zweifelsfälle sind als Antwortausfälle zu zählen.

Zentraler Punkt für die Behandlung der Antwortausfälle ist die Schätzung der Antwortwahrscheinlichkeiten der Prüffälle. Für alle Prüffälle – auch die Antwortausfälle - liegen Informationen aus dem Register über Merkmale vor, die mit dem Antwortverhalten zusammenhängen können, z.B.

- Art des Prüffalls: Personen
  - ausschließlich mit Nebenwohnung
  - mit einer oder mehreren alleinigen und/oder Hauptwohnungen
  - mit einer oder mehreren alleinigen und/oder Hauptwohnungen sowie mindestens einer Nebenwohnung
- Nationalität (deutsch/ausländisch),
- Alter (in einigen wenigen Klassen),
- Geschlecht,
- Regionale Verteilung der Wohnsitze (z.B. 19 Kategorien: alle Wohnsitze im Bundesland  $j$ ,  $j=1,\dots,16$  / Wohnsitze in verschiedenen alten Ländern / Wohnsitze in verschiedenen neuen Ländern / Wohnsitze sowohl in den alten als auch den neuen Bundesländern),
- mindestens ein Wohnsitz in einer Großstadt (ja/nein)

Die Antwortausfälle können nach diesen Merkmalen untersucht werden. Die Untersuchung sollte getrennt nach den 3 Teilen der Stichprobe (unvollständiger Geburtstag/Geburtstag 1. Januar/Geburtstag 15. Mai oder 1. Sep-

tember) durchgeführt werden. Diejenigen Merkmale, bei denen sich deutliche Unterschiede in den Antwortquoten der Ausprägungen zeigen, werden für die Behandlung der Antwortausfälle herangezogen. Die Kombination der Klassen dieser Merkmale werden im Folgenden als Antwortklassen bezeichnet. Die Anzahl und Abgrenzung der Klassen kann erst festgelegt werden, wenn die Befragung abgeschlossen und die Antwortausfälle bekannt sind. Der Anteil der Antwortausfälle  $q_{ms}$  in der Antwortklasse  $s$  des Teils  $m$  ( $m=1$ : unvollständiges Geburtsdatum,  $m=2$ : Geburtstag 01.01.,  $m=3$ : Geburtstag 15.05. oder 01.09.) ist

$$q_{ms} = \frac{b_{ms}^{(a)}}{b_{ms}}, \text{ wobei} \quad (3.2.3)$$

$b_{ms}^{(a)}$  : Anzahl der Antwortausfälle in der Antwortklasse  $s$  des Teils  $m$ ,

$b_{ms}$  : Anzahl der Prüffälle in der Antwortklasse  $s$  des Teils  $m$ .

Für einen Registersatz  $k$ , der zur Antwortklasse  $s$  des Teils  $m$  gehört, ist der Korrekturfaktor  $c_k$  für Antwortausfälle somit

$$c_k = \begin{cases} \frac{1}{1 - q_{ms}} = \frac{b_{ms}}{b_{ms} - b_{ms}^{(a)}}, & \text{falls die Person } k \text{ ein Prüffall ist,} \\ 1 & \text{sonst.} \end{cases} \quad (3.2.4)$$

Bei den Personen, die keine Prüffälle sind, treten keine Antwortausfälle auf und ihr Korrekturfaktor ist daher Eins.

### 3.2.3 Anpassung an Rahmendaten

Bei einer Geburtstagsauswahl handelt es sich stichprobenmethodisch um eine Bernoulli-Auswahl (siehe etwa Särndal 1992, Kapitel 3.2), bei der jede Einheit (hier: Registersatz) mit gleicher Wahrscheinlichkeit und unabhängig von anderen Einheiten in die Stichprobe gelangt und die Höhe des Stichprobenumfangs nicht von vornherein feststeht, sondern vom Zufall abhängt. Eine solche Stichprobe wird normalerweise nicht frei, d.h. mit den Kehrwerten der Ziehungswahrscheinlichkeiten hochgerechnet. Eine deutliche Verbesserung der Präzision der Stichprobenergebnisse erreicht man, wenn man eine gebundene Hochrechnung (Verhältnisschätzung) mit der Zahl der Einheiten durchführt, sofern diese Größe für die Grundgesamtheit bekannt ist.

Als Eckwerte stehen die Zahl der am 5. Dezember 2001 gemeldeten Personen nach Deutschen und Ausländern sowie nach Haupt- und Nebenwohnung in regionaler Gliederung bis auf Gemeindeebene zur Verfügung. Anders als bei der Schätzung der Antwortwahrscheinlichkeiten ist die Einheit hier nicht der Prüffall, sondern der Registersatz, d.h. ein Mehrfachfall wird mehrfach gezählt.

---

Gliederungsmerkmale für die Verhältnisschätzung sind

- Bundesland
- Nationalität (deutsch/ausländisch)
- Gemeindegrößenklassen (Anzahl und Grenzen liegen noch nicht fest, Abgrenzung anhand der Zahl der gemeldeten Personen).

Die Anpassungsklassen sind dann die Kombinationen der Ausprägungen aller Gliederungsmerkmale. Die Verhältnisschätzung erfolgt getrennt nach den 3 Stichprobenteilen.

Für jede Anpassungsklasse  $h$  kann nun die Zahl der gemeldeten Personen aus der Stichprobe mit den Faktoren  $w_k$  und  $c_k$  hochgerechnet und mit den Eckwerten verglichen werden.

Die Eckwerte sind:

$N_h$ : Zahl der gemeldeten Personen in der Anpassungsklasse  $h$  insgesamt,

$N_{1h}$ : Zahl der gemeldeten Personen in der Anpassungsklasse  $h$  mit unvollständigem Geburtsdatum,

$N_{2h}$ : Zahl der gemeldeten Personen in der Anpassungsklasse  $h$  mit Geburtstag 1. Januar.

$N_{1h}$  und  $N_{2h}$  ergeben sich durch Auszählung des von den Gemeinden gelieferten Einzelmaterials, während  $N_h$  die aggregierten Gemeindewerte sind.

Die Eckwerte  $N_{3h}$  für den Teil 3 ergeben sich dann als Differenz:

$$N_{3h} = N_h - N_{1h} - N_{2h} \quad (3.2.5)$$

Die aus der Stichprobe unter Berücksichtigung der Antwortausfälle hochgerechnete Zahl der gemeldeten Personen  $\hat{N}_h$  in der Anpassungsklasse  $h$  des Teils  $m$  ( $m=1,2,3$ ) ist

$$\hat{N}_{mh} = \sum_{k \in S_{mh}} c_k w_k, \quad (3.2.6)$$

wobei

$S_{mh}$ : Menge der Personen des Teils  $m$  in der Klasse  $h$  in der Stichprobe.

In der Regel wird der Eckwert durch die Hochrechnung nicht getroffen. Aus der Abweichung wird nun ein Anpassungsfaktor  $g_k$  konstruiert:

$$g_k = \frac{N_{mh}}{\hat{N}_{mh}} \quad (3.2.7)$$

wobei  $h$  die Anpassungsklasse und  $m$  der Teil ist, zu dem der Registersatz  $k$  gehört.

Der endgültige Hochrechnungsfaktor setzt sich dann multiplikativ aus den einzelnen Komponenten  $w_k$ ,  $c_k$ , und  $g_k$  zusammen. Die so berechneten Hochrechnungsfaktoren gelten streng genommen nur für die Registersätze zum Stichtag 5. Dezember 2001. Sie können aber auch für Registersätze von Personen, die nur am 31. März 2002, nicht aber am 5. Dezember gemeldet sind, übertragen werden.

In einem Teil der Auswertungen werden die Antwortausfälle als Fälle gezählt. Für diese Auswertungen müssen die Antwortausfälle nicht korrigiert werden: für alle Personen wird  $c_k = 1$  gesetzt und damit die geschätzte Zahl der Personen  $\hat{N}_{mh}$  gemäß Formel (3.2.6) berechnet. Insgesamt hat man also zwei Hochrechnungsfaktoren: einen für Auswertungen ohne Antwortausfälle und einen für Auswertungen mit Antwortausfällen.

### 3.3 Abschätzung der Stichprobenzufallsfehler

Wie beim Register- und Verfahrenstest muss die Präzision der Ergebnisse als wichtiger Aspekt der Qualität abgeschätzt werden. Diese Varianzschätzung kann aber wesentlich einfacher als beim Register- und Verfahrenstest erfolgen. Wir fassen die Antwortgewährung als eine zweite Stufe des Auswahlprozesses auf, wobei die Auswahl durch eine Poisson-Auswahl<sup>18)</sup> mit den individuellen Ziehungswahrscheinlichkeiten  $1/c_k$  für die Einheit  $k$  erfolgt. Über beide Stufen kann die Auswahl dann für die Varianzschätzung als Poisson-Auswahl mit Ziehungswahrscheinlichkeit  $1/c_k w_k$  gesehen werden. Allgemein kann bei einer Poisson-Auswahl mit Ziehungswahrscheinlichkeit  $\pi_k$  für den Schätzer des Totalwerts eines Merkmals  $Y$

$$\hat{Y} = \frac{N}{\hat{N}} \sum_{k=1}^n \frac{y_k}{\pi_k} \quad (3.3.1)$$

mit

$N$ : Zahl der Einheiten in der Grundgesamtheit

$n$ : Stichprobenumfang

$\hat{N} = \sum_{k=1}^n \frac{1}{\pi_k}$ : geschätzter Umfang der Grundgesamtheit

die Varianz geschätzt werden durch

$$\hat{V}(\hat{Y}) = \left( \frac{N}{\hat{N}} \right)^2 \sum_{i=1}^n \frac{1 - \pi_k}{\pi_k^2} (y_k - \bar{y})^2 \quad (3.3.2)$$

$$\text{mit } \bar{y} = \frac{\hat{Y}}{N}$$

(siehe Särndal 1992, Result 7.4.1)

18) Die Poisson-Auswahl ist eine Verallgemeinerung der Bernoulli-Auswahl: Jede Einheit kann eine andere Ziehungswahrscheinlichkeit besitzen (siehe etwa Särndal 1992, Kapitel 3.5).

---

Formel (3.3.2) angewendet für jede Anpassungsklasse liefert in vorliegendem Fall

$$\hat{V}(\hat{Y}) = \sum_m \sum_h \left( \frac{N_{mh}}{\hat{N}_{mh}} \right)^2 \sum_{k \in S_{mh}} c_k w_k (c_k w_k - 1) (y_k - \bar{y}_{mh})^2 \quad (3.3.3)$$

wobei

$$\bar{y}_{mh} = \sum_{k \in S_{mh}} c_k w_k y_k / \sum_{k \in S_{mh}} c_k w_k$$

der gewichtete Stichprobenmittelwert in der Anpassungsklasse  $h$  des Stichprobenteils  $m$  ist.

---

## 4 Literatur

Bierau, D. (2001): *Neue Methode der Volkszählung. Der Test eines registergestützten Zensus*, in: *Wirtschaft und Statistik* 5/2001, S. 333 – 341, Statistisches Bundesamt, Wiesbaden.

Raj, D. (1968): *Sampling Theory*. McGraw-Hill, New York.

Rompel, H.-K.(2001): *Testerhebungen für einen registergestützten Zensus*, in: *Staat und Wirtschaft in Hessen* 10/2001, S.322 – 330, Hessisches Statistisches Landesamt, Wiesbaden.

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

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