City Research Online

City St George's, University of London

Citation: Schnell, R. & Noack, M. (2016). Stichproben, Nonresponse und Gewichtung für Viktimisierungsstudien. In: Guzy, N., Birkel, C. & Mischkowitz, R. (Eds.), Viktimisierungsbefragungen in Deutschland. (pp. 8-75). Germany: Bundeskriminalamt (BKA).

This is the published version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/14392/

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Stichproben, Nonresponse und Gewichtung für Viktimisierungsstudien

Rainer Schnell und Marcel Noack

1 Vollerhebungen und Stichproben[1]

Nur bei wenigen Projekten stehen genug Ressourcen zur Verfügung, um alle interessierenden Personen (die Population oder Grundgesamtheit) untersuchen oder befragen zu können (Vollerhebung). Daher werden zumeist nur Teilmengen einer Population untersucht. Dies wird als „Stichprobenuntersuchung“ bezeichnet. Der Sinn einer Stichprobenuntersuchung besteht darin, von der Stichprobe auf die Grundgesamtheit schließen zu können. Für diese Verallgemeinerbarkeit einer Stichprobenuntersuchung ist die Art der Stichprobenziehung von zentraler Bedeutung. Wird diese Ziehung nicht über einen berechenbaren Zufallsprozess durchgeführt, dann ist die resultierende Stichprobe nachträglich nur sehr schwierig als verallgemeinerbar zu rechtfertigen. Mathematisch korrekt ist für nahezu alle praktischen Anwendungen lediglich die Verwendung echter Zufallsstichproben (Abschnitt 5). In diesem Beitrag werden wir nur solche Verfahren vorstellen, die für die Ziehung echter Zufallsstichproben prinzipiell geeignet sind. Vor allem für die tatsächliche Konstruktion, Ziehung und Gewichtung bundesweiter Stichproben der allgemeinen Bevölkerung sind umfangreichere praktische Kenntnisse erforderlich, als sie in einer Einführung vermittelt werden können. Der vorliegende Beitrag stellt nur diejenigen Details dar, die uns für die Beurteilung und Planung der Stichproben von Viktimisierungsstudien unverzichtbar erscheinen.[2]

---

[1] Einzelne Teile des Beitrags basieren auf früheren Arbeiten der Autoren, vor allem auf Schnell 2012 und Schnell u. a. 2013. Für kritische Anmerkungen danken wir Sabrina Torregroza und Christian Borgs.

[2] Dementsprechend setzen wir nur Grundkenntnisse der Inferenzstatistik, wie sie jedes einführende Lehrbuch der Statistik jenseits der deskriptiven Statistik vermittelt, voraus.

2 Konsequenzen des Studiendesigns für die Art der Stichprobenziehung

Beim Design einer Untersuchung muss man sich darüber klar sein, ob Aussagen über den Zustand einer Untersuchungsgruppe zu einem bestimmten Zeitpunkt erforderlich sind (Querschnittserhebungen) oder Aussagen über Veränderungen derselben Personen (Panel- oder Kohortenstudien) gefordert werden. Erhebungen zu einem Zeitpunkt erlauben nur sehr begrenzt Aussagen über Veränderungen, da die Verwendung von Fragen über vergangene Zustände (Retrospektivfragen) immer unter den vielfältigen Problemen autobiografischer Erinnerungen leidet (Schwarz 2007). Aufgrund der Dauer und der erheblichen Kosten von Panelstudien wird trotz der prinzipiell unüberwindbaren Probleme von Retrospektivfragen zumeist eine Entscheidung für eine einmalige Querschnittserhebung gefällt.[3] Wir gehen im Folgenden davon aus, dass es sich bei dem Projekt, für das eine Stichprobenziehung erforderlich ist, um eine einmalige Erhebung handelt.[4]

3 Angestrebte Grundgesamtheit, Auswahlgesamtheit und Inferenzpopulation

Zu Beginn eines Forschungsprojekts muss zunächst die Grundgesamtheit festgelegt werden, für die Aussagen beabsichtigt sind. Die Menge dieser Personen wird als angestrebte Grundgesamtheit bezeichnet. So könnte man z. B.

---

3 Übersichten über das Design von Viktimisierungsstudien finden sich bei Lynch 2006, Groves/Cork 2008 und Aebi/Linde 2014. Hinweise für das Design von Viktimisierungsstudien in Deutschland finden sich bei Schnell/Hoffmeyer-Zlotnik 2002.

4 Wird eine wiederholte Erhebung geplant, bleiben alle prinzipiellen Erwägungen gleich; allerdings wird je nach Art der Wiederholung ein höherer Aufwand erforderlich. Handelt es sich um eine unabhängige Wiederholungsstudie (wiederholte Querschnitte) wird nur ein höherer Dokumentationsaufwand notwendig, da das Stichprobenverfahren bei der Wiederholung exakt repliziert werden muss. Ansonsten können Veränderungen zwischen den Erhebungszeitpunkten nicht mehr auf Veränderungen der Personen zurückgeführt werden. Ist eine wiederholte Befragung derselben Personen beabsichtigt (Panel- oder Kohortenstudie), dann wird neben dem erhöhten Dokumentationsaufwand eine in der Regel größere Ausgangsstichprobe erforderlich, da allein schon durch Tod und Wanderung bei den Teilnehmern der ersten Welle mit Verlusten in der zweiten Welle gerechnet werden muss (mindestens 10 % pro Jahr bei einer Zufallsstichprobe aus der Bevölkerung). Dazu kommen weitere Ausfälle durch Verweigerung. Schließlich muss mit erheblichem Aufwand zur erneuten Kontaktierung und Sicherstellung der Befragung derselben Person (Identitätsmanagement) gerechnet werden. Die resultierenden unvermeidlichen Probleme haben Konsequenzen für die Planung und Durchführung schon bei der Stichprobenziehung. Die Details sprengen den Rahmen dieser Übersicht; es muss auf die Literatur zur Planung von Längsschnittstudien verwiesen werden. Einzelheiten finden sich bei Schnell 2012.

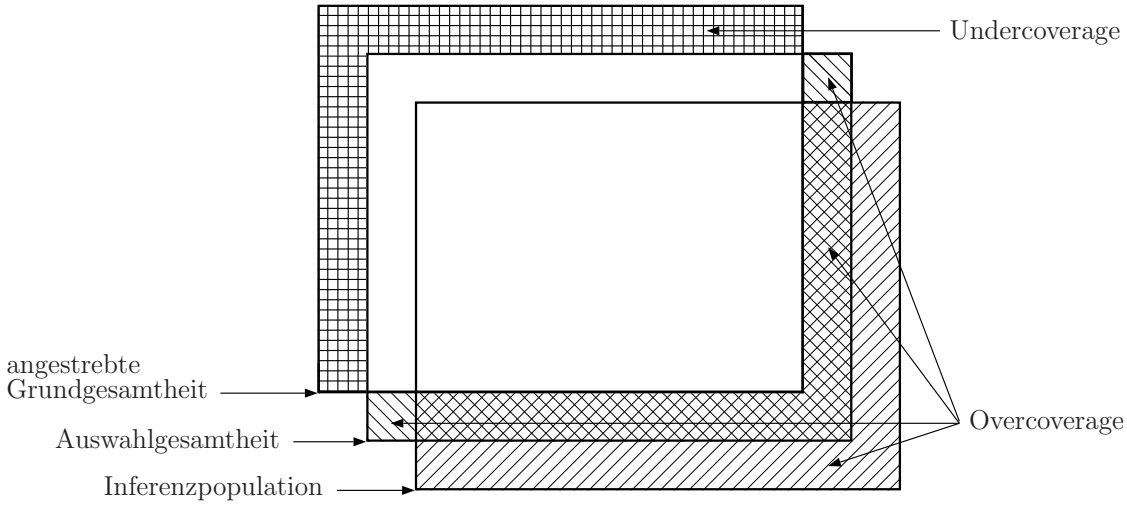
die zu einem Stichtag in Deutschland lebenden Menschen als angestrebte Grundgesamtheit betrachten. Eine solche Definition stößt auf zahlreiche Probleme. Abgesehen von den Ungenauigkeiten der verwendeten Begriffe liegt das Problem vor allem darin, dass es keine vollständigen Listen oder Datenbanken dieser Grundgesamtheit gibt. Es gibt in Deutschland lediglich Annäherungen an solche Listen, z. B. die Gesamtheit aller Einwohnermeldedateien der mehr als 5.600 Meldeämter oder die Datei der mehr als 90 Millionen vergebenen Steueridentifikationsnummern. Man könnte die Menge der Personen, die in diesen Listen enthalten sind (die sogenannte Auswahlgesamtheit oder *frame population*) als Annäherung an die angestrebte Grundgesamtheit betrachten. Zwar ist faktisch weder eine vollständige Einwohnermeldeliste noch die Datei der Steueridentifikationsnummern für Stichprobenziehungen verfügbar, neben den unüberwindlichen Problemen des Zugangs bestehen aber noch prinzipielle Probleme: Derartige Listen sind unvollständig, enthalten Duplikate und überzählige Personen.

Im Zusammenhang mit Stichproben werden solche Probleme als Coverage-Probleme bezeichnet. Es wird in der Regel unterschieden zwischen Undercoverage und Overcoverage (*Abbildung 1*). In den Einwohnermeldedateien fehlen z. B. illegale Migranten, diese werden zum Undercoverage gezählt. Es finden sich aber in der Regel auch mehrfach erfasste Personen (z. B. durch Haupt- und Nebenwohnsitze) sowie Personen, die nicht zur angestrebten Grundgesamtheit gehören (z. B. Verstorbene oder ins Ausland Verzogene), die als Overcoverage gerechnet werden. Ziel der Konstruktionen einer Auswahlgesamtheit (eines *frames*) ist die möglichst große Übereinstimmung zwischen angestrebter Grundgesamtheit und Auswahlgesamtheit, also ein möglichst geringes Ausmaß an Overcoverage und Undercoverage.[5]

[5] Über die fehlerhaften Listen hinaus können auch Fehler, die durch Interviewer oder die befragten Personen entstehen, zu Overcoverage führen. Als Beispiele können kaserniertes militärisches Personal auf Heimaturlaub, das durch den Interviewer fälschlich als Teil des zu befragenden Haushalts angesehen und interviewt wird, oder aufgrund fehlerhafter Angaben befragte Personen (minderjährige Person gibt an, volljährig zu sein) genannt werden.

Abbildung 1:

Verhältnis von angestrebter Grundgesamtheit, Auswahlgesamtheit und Inferenzpopulation (Schnell u. a. 2013, 262)



Durch die tatsächliche Durchführung fallen weitere Personen aus der Auswahlgesamtheit heraus (z. B. durch mangelnde Sprachkenntnisse). Gelegentlich werden auch Personen berücksichtigt, die nicht zur angestrebten Grundgesamtheit gehören, z. B. wenn bei telefonischen Befragungen minderjährige Personen befragt werden, aber nur Erwachsene befragt werden sollen. Die Menge der Personen, aus denen die tatsächlich Befragten eine Zufallsstichprobe darstellen würden, wird als „Inferenzpopulation“ bezeichnet, weil nur über diese Aussagen gemacht werden können. Die Inferenzpopulation sollte also der angestrebten Grundgesamtheit möglichst entsprechen.

4 Aus Bevölkerungserhebungen ausgeschlossene Populationen

In den meisten Projekten erfolgt die Definition der Grundgesamtheit bis zur Ziehung der Stichprobe kaum explizit. Selbst für die Vorbereitung der Stichprobenziehung wird die Grundgesamtheit selten exakt definiert. Üblich sind ungenaue Kurzdefinitionen der Grundgesamtheit wie z. B. „erwachsene Wohnbevölkerung“ oder „in Privathaushalten lebende deutsche Staatsangehörige ab 18 Jahren“. Solche Definitionen sind für eine praktische Umsetzung nicht exakt genug. Die Ungenauigkeiten sind dabei keinesfalls folgenlos, da nicht klar ist, welche Populationen ausgeschlossen sind und welche nicht (Unterkapitel 4.6).

Im Folgenden werden einige der ausgeschlossenen Populationen erwähnt und ihre Größenordnung abgeschätzt.[6] Am bedeutsamsten ist in diesem Zusammenhang die Bevölkerung in Institutionen.

4.1 Bevölkerung in Institutionen

In der älteren deutschen Literatur wurde diese Gruppe als „Anstaltsbevölkerung" bezeichnet, im Zensus 2011 als „Sonderbereiche". Darunter fallen laut § 2 Abs. 5 S. 1–3 des Zensusgesetzes 2011 Gemeinschafts-, Anstalts- und Notunterkünfte, Wohnheime und ähnliche Unterkünfte.[7]

Im Zensus 2011 wurde zwischen sensiblen und nicht sensiblen Sonderbereichen unterschieden. Zu Ersteren gehören Behindertenwohnheime, spezielle Krankenhäuser (wie z. B. Palliativstationen, Hospize, psychiatrische Kliniken), Justizvollzugsanstalten und andere Einrichtungen des Maßregelvollzugs sowie Flüchtlingsunterkünfte und Unterkünfte für Wohnungslose. Sowohl zu sensiblen als auch nicht sensiblen Sonderbereichen können Kinder- und Jugendheime sowie Mutter-Kind-Heime zählen. Als nicht sensible Sonderbereiche galten Studentenwohnheime, Arbeiterheime und sonstige Wohnheime, Alten- und Pflegeheime, Internate, Schulen des Gesundheitswesens sowie Klöster.

Legt man den Anteil der Personen zugrunde, die in Niedersachsen in sensiblen Sonderbereichen wohnen (Mayer 2013), dann wären insgesamt in Deutschland ca. 271.000 Personen in sensiblen Sonderbereichen, für die nicht sensiblen Sonderbereiche entsprechend 1.375.000 Personen zu erwarten. Damit wären insgesamt 1,646 Millionen Personen oder etwas mehr als 2 % der Bevölkerung in Sonderbereichen zu finden. Verglichen mit den USA (2,5 %)[8], erscheint dies als eine plausible Größenordnung.

---

6 Eine ausführliche Übersicht über die Größe faktisch aus der Befragung der „allgemeinen Bevölkerung" ausgeschlossener Populationen in Deutschland findet sich bei Schnell 1991. Trotz des Alters der Studie ist dies bislang die einzige Publikation zu diesem Problem in Deutschland.

7 Die Dokumentation des Zensus 2011 ist auch in diesem Bereich spärlich; einige wenige Einzelheiten zur Erhebung in Sonderbereichen finden sich bei Geiger/Styhler 2012 und Mayer 2013.

8 Für die USA werden in „Group Quarters" insgesamt nahezu 8 Millionen Personen geschätzt, die Hälfte davon in Institutionen (National Research Council 2012, 25).

4.2 Personen mit besonderen Wohnungsbedingungen

In allen Industriegesellschaften lebt ein Teil der Bevölkerung nicht in Wohngebäuden, sondern in anderen Gebäuden oder Unterkünften. Nach den ersten Ergebnissen (Statistische Ämter des Bundes und der Länder 2014) der Gebäude- und Wohnungszählung gab es 2011 3,4 % der Wohnungen in „sonstigen Gebäuden mit Wohnraum“ (Schulen, Gewerbeobjekte etc.). Falls die Stichprobenziehung durch eine Begehung vor Ort erfolgt (z. B. in den sogenannten Random-Route-Verfahren), werden Personen in Nichtwohngebäuden allgemein häufig nicht erfasst, so z. B. Hausmeister, Mönche, Nonnen oder Personen in Bereitschaftsunterkünften der Polizei oder der Bundeswehr (Schnell 1991).

Ein anderes Problem sind „bewohnte Unterkünfte“ wie Wohn- oder Bauwagen, Baracken, Gartenlauben, fest verankerte Wohnschiffe, Schrebergartengebäude und Weinberghütten sowie Dauercampende. Obwohl der größte Teil dieser Population an irgendeiner Stelle administrativ erfasst ist, kann der tatsächliche Wohnort von der erfassten Anschrift abweichen. Je nach Art der verwendeten Auswahlgrundlage, z. B. Listen von Telefonnummern oder Einwohnermelderegister, können solche Populationen enthalten sein oder nicht. Im Einzelfall hängt die Auswahl dieser Personen jedoch zumeist von willkürlichen Entscheidungen ab. Für den Zensus 2011 werden ca. 10.000 bewohnte Unterkünfte mit ca. 15.000 Wohnungen nachgewiesen (Statistische Ämter des Bundes und der Länder 2014).[9] Sollte sich die Haushaltszusammensetzung gegenüber der Volkszählung von 1987 (VZ87) nicht verändert haben, entspräche dies 27.000 Personen. Schließlich sollen noch die Menschen erwähnt werden, die ohne jede Unterkunft auf der Straße leben: Die Bundesarbeitsgemeinschaft Wohnungslosenhilfe gibt diese Zahl mit 24.000 für 2012 an (Bundesarbeitsgemeinschaft Wohnungslosenhilfe 2013).

4.3 Klandestine Populationen

Mit dem Begriff ‚klandestine Populationen‘ werden seit einigen Jahren Subgruppen bezeichnet, die sich willentlich einer amtlichen Erfassung entziehen. Hierzu gehören zunächst einmal Personen ohne Aufenthaltsgenehmigung oder -duldung (ohne Scheinlegale oder registrierte Ausreisepflichtige) (Vogel/Äßner 2011). Dies dürfte den größten Teil dieser Populationen darstellen. Für Deutschland existieren nur höchst unvollkommene Schätzungen in der Grö-

[9] Für die Volkszählung von 1987 wurden 25.400 Haushalte mit 45.600 Personen unter solchen Bedingungen gezählt (Schnell 1991). Es erscheint erstaunlich, dass sich diese Zahl trotz Wiedervereinigung gegenüber der VZ87 mehr als halbiert haben soll.

ßenordnung von 100.000 bis 675.000 Personen (Vogel/Äßner 2011), vereinzelt finden sich auch höhere Angaben, z. B. bei CLANDESTINO Project (2009). In anderen Ländern werden durchaus erhebliche Größenordnungen (für die USA z. B. 8 Millionen Personen) erwartet. Aufgrund der besonderen rechtlichen Situation dieser Personen muss von einer besonders hohen Viktimisierungswahrscheinlichkeit ausgegangen werden. Entsprechend werden Opferbefragungen, die diesen Personenkreis ausschließen, systematisch zu niedrige Viktimisierungsraten ermitteln. Ein ähnliches Argument gilt für andere klandestine Populationen wie z. B. Straftäterinnen und Straftäter. Bei anderen Populationen als illegalen Migranten werden aber in der Regel nur kleine Anteile an der Gesamtpopulation vermutet, sodass kaum ein Effekt auf Viktimisierungsraten oder andere Populationsparameter der Gesamtbevölkerung erwartet werden kann.

4.4 Populationen mit gesundheitlichen Problemen

Personen mit gesundheitlichen Problemen, beispielsweise Patienten in psychiatrischen Kliniken, pflegebedürftige oder demente Personen, stellen ein weiteres Problem dar (Schnell 1991). Wie bei fast allen Gesundheitsstatistiken ist die Datenlage in Deutschland auch für die Prävalenz von Demenz unbefriedigend. Auf der Basis der vorliegenden Analysen (Ziegler/Doblhammer 2009; Doblhammer u. a. 2012) erscheint die Schätzung eines Anteils von Dementen bei der über 65-jährigen Bevölkerung von 6 bis 7 % als realistisch. Dies entspräche einer Gesamtzahl von 1,17 Millionen Erkrankten. Das Robert-Koch-Institut geht davon aus, dass 60 % der Menschen mit Demenz in Privathaushalten gepflegt werden (Weyerer 2005). Das entspräche 709.000 Personen. Diese Personengruppe dürfte aus nahezu allen Befragungen ausfallen.[10]

4.5 Populationen mit Sprachbarrieren

Wenn nicht besondere Maßnahmen ergriffen werden, dann wird faktisch unabhängig von der Definition der Grundgesamtheit immer nur die deutschsprachige Bevölkerung befragt. Man muss sich in Erinnerung rufen, dass bei der Befragung einer „deutschsprachigen“ Bevölkerung letztlich immer das Datenerhebungspersonal über die Zugehörigkeit zur Grundgesamtheit entscheidet; entsprechende Interviewereffekte sind trivialerweise erwartbar.

---

[10] Ein nicht unbeträchtlicher Teil davon dürfte körperlicher Gewalt ausgesetzt sein. Zu den Problemen der Schätzung des Anteils demenziell Erkrankter, die Opfer körperlicher Gewalt werden, siehe Weissenberger-Leduc/Weiberg (2011, 35–38).

Dieser unpräzise Ausschluss der „nicht deutschsprachigen Bevölkerung“ ist für Viktimisierungsstudien kaum zu ignorieren. Legt man die Daten der RAM-Studie 2006/2007 der fünf größten Ausländerpopulationen (Personen aus der Türkei, Ex-Jugoslawien, Italien, Polen, Griechenland; zusammen ca. 57 % der ausländischen Personen in Deutschland) zugrunde, dann ergibt sich ein Anteil von ca. 10 % dieser Personen, die nicht gut genug Deutsch sprechen können, um sich problemlos im Alltag verständigen zu können (Haug 2008). Nimmt man an, dass dieser Anteil auch für andere Ausländergruppen in Deutschland gilt, dann ergäben sich ca. 670.000 Personen, mit denen eine Verständigung auf Deutsch Probleme bereiten würde. Berücksichtigt man nur die über 14-Jährigen, dann handelt es sich grob um mehr als eine halbe Million Personen (ca. 0,7 % der Bevölkerung über 14 Jahren), die allein aufgrund ihrer Deutschkenntnisse ausgeschlossen würden.

4.6 Konsequenzen des Ausschlusses spezieller Populationen

Der Ausschluss zahlreicher und sehr spezieller Populationen bleibt nicht ohne Folgen für die Schätzung von Viktimisierungsraten. In nahezu allen ausgeschlossenen Subgruppen kann von erhöhten Viktimisierungsraten ausgegangen werden, entsprechend führt der Ausschluss dieser Subgruppen zu niedrigeren Viktimisierungsraten (hierzu insbesondere Lynn 1997).[11] Wie fast immer bei Befragungen werden die Ergebnisse sozialpolitisch umso erfreulicher, je schlechter die Erhebungen durchgeführt werden. Daher ist es nicht verwunderlich, dass z. B. der Ausschluss der Bevölkerung in Institutionen aus Erhebungen zumeist undiskutiert bleibt.[12]

Man kann argumentieren, dass Viktimisierungsstudien nicht versuchen, eine unverzerrte Schätzung der Population zu erreichen, sondern lediglich einen Indikator für den Zustand einer Population liefern sollen. Dann muss man aber konstante Verzerrungen unterstellen, wenn Vergleiche über die Zeit oder verschiedene Studien oder mehrere Länder erfolgen sollen. Dies ist ohne exakte Kontrolle der Art und der Größe der ausgeschlossenen Population nicht möglich. Daher muss der Dokumentation der Art und Größe der ausgeschlossenen Populationen gerade bei Viktimisierungsstudien erhöhte Aufmerksamkeit gewidmet werden. Dies bedeutet, dass den Eigenschaften der

---

11 Dies wurde schon für den British Crime Survey (BCS) kritisiert (Smith 2006, 10).

12 Eine der wenigen Ausnahmen findet sich im Hinblick auf Viktimisierungsstudien in den USA in einer neueren Veröffentlichung (National Research Council 2014, 124): „The frame for the ancillary listing of group quarters, which is an important part of the secondary sample for the National Crime Victimization Survey because their residents may be at higher risk for sexual violence, is seriously flawed in terms of both the building and enumeration of this secondary frame.“

für die jeweilige Studie verwendeten Erhebungsinstrumente, den Interviewern und den Auswahlgrundlagen mehr Aufmerksamkeit in Hinsicht auf ausgeschlossene Populationen gewidmet werden muss als z. B. bei einer Wahlabsichtsbefragung.

5 Formen von Auswahlverfahren

Schlüsse von einer Stichprobe auf eine Grundgesamtheit lassen sich ohne schwerlich zu rechtfertigende sonstige Annahmen nur dann mathematisch begründen, wenn die Stichprobe durch einen Zufallsprozess gezogen wird. Dabei ist es von entscheidender Bedeutung, dass die Wahrscheinlichkeit für die Auswahl jedes einzelnen Elements der Grundgesamtheit berechnet werden kann. Es ist dabei nicht wesentlich, ob die Wahrscheinlichkeiten gleich sind oder nicht.[13] Die Auswahlwahrscheinlichkeiten müssen aber berechenbar und größer als Null sein.

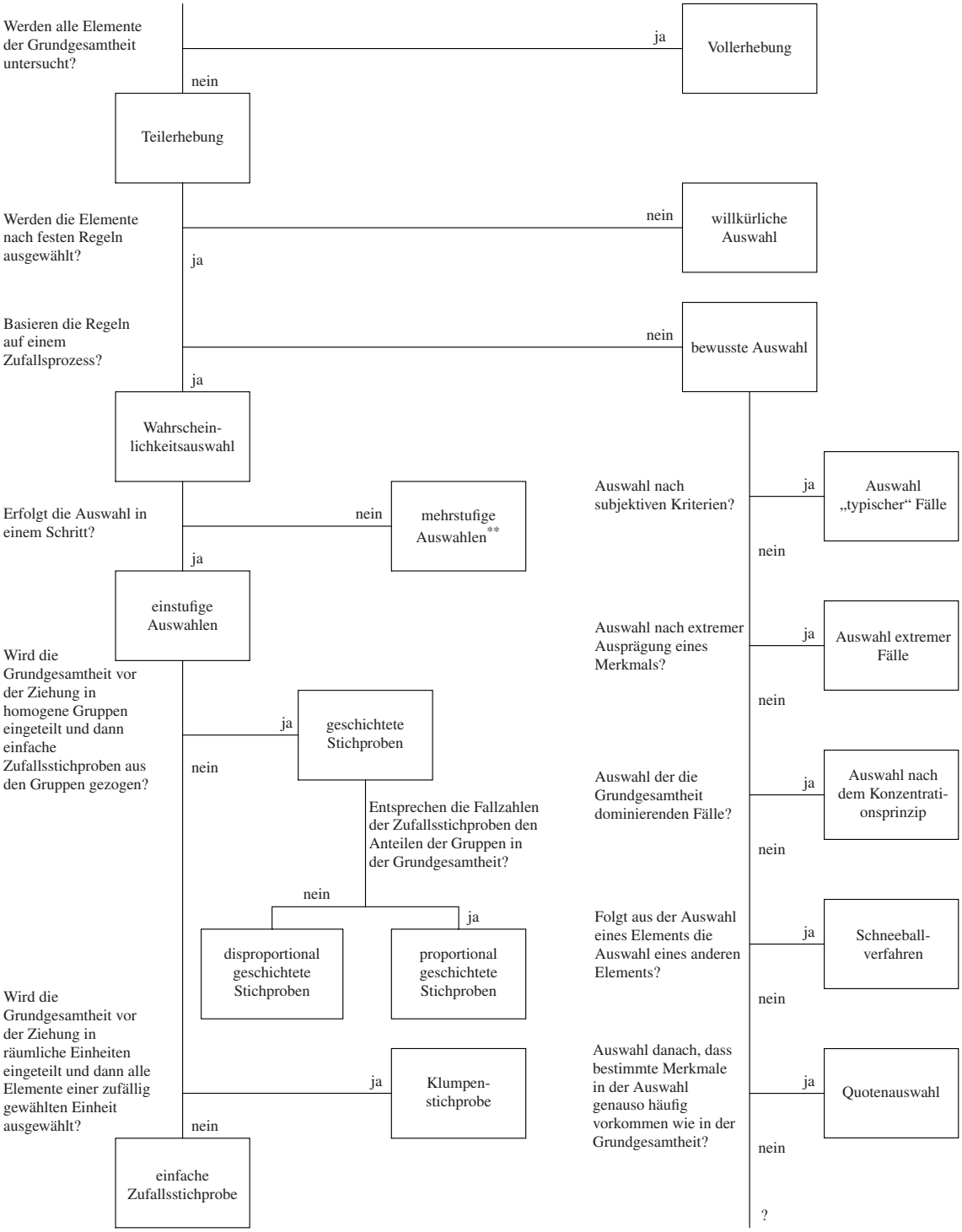
5.1 Willkürliche und bewusste Auswahlverfahren

Bei allen sogenannten willkürlichen oder bewussten Auswahlen sind die Auswahlwahrscheinlichkeiten nicht berechenbar und/oder nicht größer als null (*Abbildung 2*, rechte Seite der Abbildung).

[13] Falls ungleiche Auswahlwahrscheinlichkeiten vorliegen, kann dies durch eine Gewichtung berücksichtigt werden.

Abbildung 2:

Übersicht über Auswahlverfahren (Schnell u. a. 2013, 260)



** Mehrstufige Auswahlen bestehen aus Kombinationen einstufiger Verfahren mit unterschiedlichen Auswahleinheiten.

Willkürliche Auswahlen sind nicht regelgeleitet, sondern man greift auf etwas Verfügbares zurück (sogenannte *convenience samples*). Solche Stichproben sind prinzipiell nicht verallgemeinerbar. Im Gegensatz dazu liegt bewussten Auswahlen zwar eine Regel zugrunde, die Auswahlwahrscheinlichkeiten sind aber trotzdem nicht berechenbar. Daher sind auch bewusste Auswahlen nicht verallgemeinerbar. Dazu gehören z. B. die Auswahl extremer Fälle oder „typischer Fälle“ (häufig irreführend von Laien als *theoretical sampling* bezeichnet). Dazu gehört auch die Auswahl nach dem Konzentrationsprinzip, also z. B. die Auswahl der zehn größten Schulen aus einem Bezirk. Ebenso nicht verallgemeinerbar ist das sogenannte „Schneeballverfahren“: Hat man wie auch immer ein Mitglied einer seltenen Subgruppe gefunden, ist es plausibel, dass dieses Mitglied andere Mitglieder der Subgruppe kennt: Man befragt dann ausgehend von der Indexperson weitere benannte Personen.[14]

Schließlich gehört zu dieser mathematisch nicht zu rechtfertigenden Klasse von Verfahren auch das Quotenverfahren. Dabei werden Interviewern Quoten bestimmter Personenmerkmale oder Kombinationen dieser Merkmale vorgegeben, z. B. fünf Männer im Alter von 30 bis 39 Jahren und fünf Frauen im Alter von 30 bis 39 Jahren, jeweils katholisch und berufstätig. Innerhalb dieser Kombination kann aber der Interviewer die Befragten beliebig auswählen (also: willkürlich). Damit sind die Auswahlwahrscheinlichkeiten nicht berechenbar und das Verfahren mathematisch nicht zu rechtfertigen.[15]

5.2 Auswahlverfahren mit berechenbaren Auswahlwahrscheinlichkeiten

Verfahren mit berechenbaren Auswahlwahrscheinlichkeiten lassen sich mit einem einfachen Modell erläutern: Jedes Element der Grundgesamtheit bekommt eine Losnummer zugeteilt, die Lose werden in einer Trommel gemischt und dann nacheinander gezogen. In diesem Fall handelt es sich um ein einstufiges Auswahlverfahren einer einfachen Zufallsstichprobe ohne Zurücklegen. Heute werden die Personen auf einer Liste einfach nummeriert und

---

14 Eine Variante dieses Verfahrens wird zurzeit als *Respondent Driven Sampling* auch außerhalb der Statistik bekannter. Für das Verfahren müssen einige schwierig zu prüfende Annahmen getroffen werden (z. B. darf die Wahrscheinlichkeit einer Selektion von Merkmalen des letzten Selektionsschritts abhängen, aber nicht mehr von vorherigen Selektionsschritten), daher ist es prinzipiell unklar, wann das Verfahren angewandt werden kann oder nicht (einführend: Schonlau u. a. 2014). Im Zweifelsfall sind auch hier die Auswahlwahrscheinlichkeiten unbestimmbar.

15 In der Politik und in einigen Bereichen der Marktforschung ist das Quotenverfahren aufgrund seiner schnellen und wenig aufwendigen Durchführung immer noch verbreitet, in wissenschaftlichen Anwendungen spielt es kaum noch eine Rolle. Einzelheiten finden sich bei Schnell u. a. 2013.

dann durch Zufallszahlen aus einem Zufallszahlengenerator gezogen. Dies ist zum Beispiel das häufigste Verfahren, wenn für eine Gemeinde aus dem Einwohnermelderegister gezogen wird.[16]

Manchmal verfügt man nicht über eine Liste aller Personen, sondern nur über eine Liste von Ansammlungen („Klumpen“ oder „Cluster“) von Personen. Das klassische Beispiel sind Schulen: Man hat keine Liste der Schülerschaft, aber eine Liste von Schulen. Jede Schülerin bzw. jeder Schüler gehört zu genau einer Schule. In diesem Fall würde man die Klumpen zufällig ziehen und – im einfachsten Fall – die gesamte Schülerschaft in einem Klumpen auswählen. Dies wäre eine Klumpenstichprobe. Ein anderes Beispiel für dieses Auswahlverfahren sind Flächenstichproben: Kann jede Person genau einer Fläche zugeordnet werden, dann kann man Flächen (wie z. B. Häuser) als Klumpen einer Klumpenstichprobe verwenden. Das Problem von Klumpenstichproben besteht darin, dass Personen innerhalb eines Klumpens einander ähnlicher sind als zufällig ausgewählte Personen. Dieses Problem wird als „Klumpeneffekt“ bezeichnet, der letztlich die Präzision der Schätzungen verringert und bei den Analysen berücksichtigt werden muss (Abschnitt 8). Die praktische Konsequenz bei Klumpenstichproben besteht darin, möglichst viele Klumpen und möglichst wenig Personen pro Klumpen zu ziehen.

Häufig gibt es Einteilungen der Grundgesamtheit, deren Vergleich von besonderem Interesse ist. Diese Einteilungen werden als Schichten bezeichnet. Ist die Größe der Schichten in der Grundgesamtheit bekannt, kann die Präzision von Stichprobenschätzungen durch Berücksichtigung dieser Schichtung verbessert werden.[17] Möchte man eine Stichprobe aus Deutschland ziehen und Vergleiche zwischen Bundesländern anstellen, dann wird häufig nach Bundesländern geschichtet. Das bedeutet nichts anderes, als dass pro Bundesland eine unabhängige Stichprobe gezogen wird. Die gesamte Stichprobe ist dann eine geschichtete Stichprobe aus der gesamten Bundesrepublik.

Bei geschichteten Stichproben wird danach unterschieden, ob die Größe der Schichten ihrem Anteil in der Grundgesamtheit entspricht oder nicht. Sind die Anteile der Schichten in der Stichprobe genauso groß wie in der Grund-

---

16 In der Praxis findet man häufig eine Annäherung an dieses Verfahren, das als systematische Stichprobe bezeichnet wird. Dabei wird z. B. jede zehnte Person auf der Liste ausgewählt. Das Problem solcher Verfahren besteht darin, dass die Liste selbst systematisch geordnet sein kann und daher systematische Fehler entstehen. Ähnliches gilt für Buchstaben oder Geburtstagsverfahren, bei denen Personen mit bestimmtem Geburtsdatum oder Anfangsbuchstaben von Namen ausgewählt werden. Da echte Zufallsverfahren heute keine zusätzlichen Kosten verursachen, sollte im Regelfall immer eine echte Zufallsstichprobe gezogen werden, um jede unerwünschte Systematik der Auswahl zu verhindern.

17 Diese Verbesserung ist dann möglich, wenn die Schichten unterschiedliche Kennwerte und Varianzen aufweisen. Dies ist bei den üblichen Schichtungen nach Bundesländern und Ortsgrößen meistens der Fall.

gesamtheit, dann handelt es sich um eine proportionale Schichtung, andernfalls um eine disproportionale Schichtung.[18] Seit der Wende wird bei Bevölkerungsstichproben häufig disproportional nach alten und neuen Bundesländern geschichtet: Neue Bundesländer werden in höherem Ausmaß berücksichtigt als es ihrem Bevölkerungsanteil entspricht.

Die meisten bundesweiten Bevölkerungsstichproben (oder *national samples* in anderen Ländern) verwenden Kombinationen der bisher erläuterten Auswahlverfahren, in der Regel also geschichtete Stichproben von Klumpen, aus denen dann einfache Zufallsstichproben gezogen werden. Solche Kombinationen sind technisch mehrstufige Auswahlen, die häufig als „komplexe Stichproben" bezeichnet werden. Wichtig dabei ist, dass die Auswahlwahrscheinlichkeit auf jeder einzelnen Stufe berechenbar ist: Wird z. B. auf der letzten Stufe nicht berechenbar ausgewählt, dann ist die Stichprobe nicht „komplex", sondern willkürlich und damit unbrauchbar. Einzelheiten komplexer Stichproben hängen vom Erhebungsmodus der Befragung ab und werden in Kapitel 9 diskutiert.

6 Das Total-Survey-Error-Modell

Einführende oder rein mathematische Lehrbücher konzentrieren sich bei der Diskussion um Auswahlverfahren häufig ausschließlich auf Stichprobenschwankungen, also die Varianz der Schätzungen bei wiederholten Ziehungen aus einer stabilen Grundgesamtheit. Die Wurzel aus dieser Varianz ist der Standardfehler, das üblicherweise verwendete Maß für die Präzision einer Schätzung. In einführenden Lehrbüchern wird dies in der Regel weiter vereinfacht auf die ausschließliche Schätzung des Standardfehlers einfacher Zufallsstichproben.

Das Resultat solcher Vereinfachungen sind dann z. B. irreführende Aussagen wie:

> Von 1.250 Befragten entscheiden sich auf die Frage, wen sie am nächsten Sonntag wählen wollen, 40 Prozent für eine Partei. Die Fehlertoleranz liegt hier bei

---

[18] Eine disproportionale Schichtung kann zum Beispiel verwendet werden, um die Schätzungen in jeder Schicht *h* mit einer ausreichenden Präzision durchführen zu können. Falls die resultierenden Stichprobengrößen n_h innerhalb kleiner Schichten einer proportional geschichteten Stichprobe nicht ausreichen, um die interessierenden Parameter mit der gewünschten Präzision zu schätzen, dann könnte eine disproportional geschichtete Stichprobe zur Lösung dieses Problems verwendet werden. In solch einer disproportionalen Stichprobe sind dann die kleinen Schichten überproportional vertreten, die großen Schichten hingegen unterproportional, sodass die gewünschte Präzision in allen Schichten erreicht werden kann (Kalton 1983, 24 f.). Für spezielle Allokationskriterien für disproportionale Schichtungen wie die varianzoptimale Allokation nach Neyman siehe Lohr 2010.

> rund +/– 3 Prozentpunkten. Das heißt, der Anteil dieser Partei bei allen Wahlberechtigten liegt zwischen 37 und 43 Prozent.[19]

Das ist natürlich in mehrfacher Hinsicht falsch; die tatsächlichen Fehlertoleranzen sind sehr viel größer (Schnell und Noack 2014). Das Problem besteht nicht darin, dass die verwendeten Formeln falsch sind oder falsch gerechnet wurde: Die Voraussetzungen für die Anwendungen der einfachen Modelle sind aber nicht gegeben. Im Beispiel des ZDF-Politbarometers dürften die tatsächlichen „Fehlertoleranzen" nicht bei +/– 3 Prozent, sondern bei +/– 9 Prozent liegen.[20]

Um eines deutlich zu machen: Die Tatsache, dass die tatsächlichen „Fehlertoleranzen" komplexer Stichproben in der Forschungspraxis sehr viel größer sind, als die naiven Berechnungen aus Einführungslehrbüchern glauben lassen, ist in der Statistik vollkommen unumstritten. Die Probleme der korrekten Berechnung basieren zum einem darauf, dass man für eine korrekte Berechnung sehr viel mehr über die Erhebung wissen muss, als man üblicherweise einer Pressemitteilung entnehmen kann (z. B. Intraklassenkorrelationen, Klumpengrößen, Schichtung, Gewichtungsfaktoren etc.)[21]. Wir werden dies in Kapitel 8 genauer darstellen. Zum anderen sind die resultierenden Intervalle bei den in der Praxis üblichen kleinen Stichproben so groß, dass man den Nutzen einer solchen Erhebung kaum begründen kann: Wer würde das ZDF-Politbarometer schauen, wenn die Fehlergrenzen korrekt angegeben würden? Dann müsste das Ergebnis lauten: „Die CDU würde zwischen 31 und 49 % der Stimmen erhalten". Aus diesen beiden Gründen ist die Berechnung korrekter Intervalle in der Praxis – vor allem in der Meinungsforschung – nicht weit verbreitet.

Der Unterschied zwischen der naiven Schätzung und den korrekten empirischen Ergebnissen lässt sich am folgenden Beispiel zweier Viktimisierungsstudien zeigen. Schnell und Kreuter (2000) verglichen die Ergebnisse zweier im Auftrag des Bundesministeriums für Justiz 1997 durchgeführter Studien (eine Mehrthemenbefragung, eine Befragung im Rahmen des Sozialwissenschaften-Bus III/97), die nahezu zeitgleich vom selben Institut mit den gleichen Fragen in Deutschland erhoben wurden. Das Ergebnis illustriert *Abbildung 3*.

---

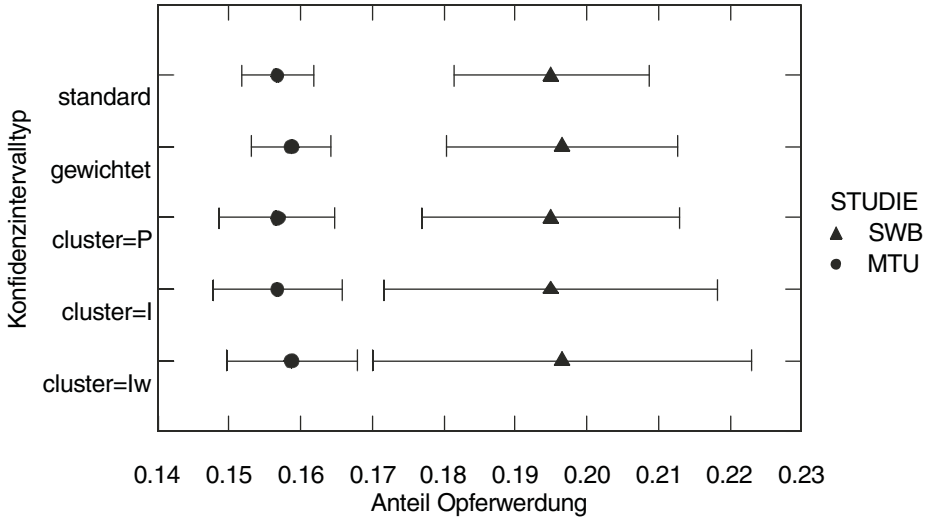
19 Homepage des ZDF-Politbarometers am 04. 11. 2013.

20 Berechnet aus allen Wahlprognosen für die CDU im Datensatz bei Schnell und Noack (2014). Die CDU erzielte bei den Prognosen im Mittel 40,3 %, die mittlere Stichprobengröße liegt bei 1.540 Fällen. Für diese etwas günstigeren Ausgangsdaten wurden die Konfidenzintervalle so vergrößert, dass 95 % der Konfidenzintervalle die Ergebnisse für die CDU bei der höchstens einen Monat nach der Prognose stattfindenden Bundestagswahl enthalten. Dies ist erst dann der Fall, wenn die Konfidenzintervalle +/– 9 % umfassen.

21 Der Intraklassenkorrelationskoeffizient stellt ein Maß für die Homogenität der Cluster dar.

Abbildung 3:

Konfidenzintervalle des Anteils der Opferwerdung zweier nahezu zeitgleicher deutschlandweiter Erhebungen desselben Instituts mit identischen Fragen (Schnell/Kreuter 2000, 102)



Offensichtlich sind die Ergebnisse unvereinbar: Die Konfidenzintervalle überlappen sich nicht, unabhängig davon ob man ein naives Standardkonfidenzintervall („standard“) berechnet, zusätzlich die Gewichtung in das Modell („gewichtet“) aufnimmt, den Erhebungsort („cluster = P“) oder den Interviewer („cluster = I“) als Klumpeneffekt einschließt oder alle Probleme simultan berücksichtigt (Gewichtung und Interviewer als Klumpen, „cluster = Iw“). Solche signifikanten Unterschiede bei einer unabhängigen Replikation bei einer unveränderten Grundgesamtheit sollten selten (in weniger als 0,6 % der Studien) auftreten. Die vermutliche Ursache für solche Ergebnisse liegt darin, das bei einem Schätzergebnis einer Befragung nicht nur die Unsicherheit durch die Stichprobe (*Sampling Error*) berücksichtigt werden muss, sondern auch alle anderen Fehlerquellen.[22]

In der Statistik werden diese anderen Fehlerquellen zusammenfassend als *non-sampling errors* bezeichnet.[23] Die Größe dieser Fehler kann die der Standardfehler deutlich übersteigen.

---

[22] Ein Vergleich der pro Land und Erhebungsjahr geschätzten Anteile von Personen mit „Furcht vor Kriminalität“ für die europaweiten Surveys „ESS“, „ICVS“ und „Eurobarometer“ ergab ebenfalls teilweise unvereinbare Ergebnisse. Analog werden auch hier *non-sampling errors* als vermutliche Ursache dafür angesehen (Noack 2015, 94–123).

[23] Eine Übersicht über die Literatur zu nahezu allen dieser Fehlerquellen gibt Weisberg 2005.

In der Diskussion um die Qualität eines Surveys spielt daher in der wissenschaftlichen Literatur zunehmend ein erweitertes statistisches Fehlermodell eine zentrale Rolle. Dieses Fehlermodell wird als Total-Survey-Error-Modell bezeichnet.[24] Definiert man den Fehler der Schätzung einer Statistik $\hat{\mu}$ eines Parameters μ für einen Survey als

$$Fehler = \hat{\mu} - \mu, \tag{1}$$

dann ist das in der Surveystatistik übliche Gütemaß für die Schätzung der sogenannte *mean-squared error* (MSE).[25] Der MSE ist eine Kombination des Ausmaßes der Abweichung der Schätzungen vom Populationswert (Bias) und des Ausmaßes der Streuung der Schätzungen vom Populationswert (Varianz der Schätzungen):

$$MSE(\hat{\mu}) = B^2 + Var(\hat{\mu}), \tag{2}$$

wobei $B = E(\hat{\mu} - \mu)$ den Bias, E den Erwartungswert und $Var(\mu)$ die Varianz der Schätzungen darstellt. Beim Design und der Durchführung eines Surveys versucht man den MSE für die interessierenden Schätzungen zu minimieren. Üblicherweise führt man in der Gleichung für die Schätzung des MSE die Quellen des Bias eines Surveys einzeln auf:

$$MSE = \left(B_{spez} + B_{nr} + B_{cover} + B_{mess} + B_{da}\right)^2 + Var_{sampling} + Var_{mess} + Var_{da}, \tag{3}$$

wobei

$B_{spez} = \text{Spezifikationsfehler}$
$B_{nr} = \text{Nonresponsebias}$
$B_{cover} = \text{Coveragebias}$
$B_{mess} = \text{Messfehler}$
$B_{da} = \text{Datenaufbereitungsbias}$
$Var_{sampling} = \text{Varianz der Kennwerteverteilung}$
$Var_{mess} = \text{Messfehlervarianz}$
$Var_{da} = \text{Datenaufbereitungsvarianz}$
ist (Biemer/Lyberg 2003, 59).

---

24 Dieser Abschnitt wurde Schnell 2012 entnommen.

25 Die Darstellung folgt hier Biemer 2010.

Mit Ausnahme von Spezifikationsfehler, Messfehler und Datenaufbereitungsfehler werden alle Fehlerquellen in diesem Kapitel behandelt. Die hier nicht behandelten Fehlerquellen betreffen Operationalisierungs- und Messfehler (durch Interviewer, Befragte, Erhebungsinstrument und Erhebungsmodus) sowie Datenaufbereitungs- und Datenanalysefehler – diesbezüglich muss auf die jeweilige Spezialliteratur verwiesen werden.[26]

Im Prinzip ist die Schätzung aller einzelnen Bestandteile des MSE zumindest mit vereinfachenden Annahmen möglich, wenngleich auch außerordentlich aufwendig. Das Modell des Total-Survey-Errors wird daher fast immer nur als regulative Idee verwendet. Bisher wurde das Modell für Erhebungen in Deutschland kaum thematisiert. Einen empirischen Versuch am Beispiel der Ungenauigkeiten der Wahlprognosen in Deutschland findet man bei Schnell und Noack (2014).

7 Standardfehler und Konfidenzintervalle: Ermittlung der benötigten Stichprobengröße bei einfachen Zufallsstichproben und vereinfachten Annahmen

Es gibt keine absolute Mindestgröße einer Zufallsstichprobe. Die notwendige Größe einer Stichprobe hängt nahezu ausschließlich davon ab, mit welcher Genauigkeit man eine Aussage treffen möchte. Die Genauigkeit einer Schätzung auf der Basis einer Stichprobe wird durch die Größe der Konfidenzintervalle ausgedrückt.

Die Breite eines Konfidenzintervalls (selten auch Vertrauensintervall genannt) wird zunächst durch den Standardfehler bestimmt. Der Standardfehler ist dabei nicht zu verwechseln mit der Standardabweichung, also der Streuung der Messungen um ihren Mittelwert. Im Gegensatz dazu ist der Standardfehler definiert als die Standardabweichung der Stichprobenkennwertverteilung, also der Verteilung der geschätzten Stichprobenstatistiken (z. B.

[26] Spezifikationsfehler sind Unterschiede zwischen den tatsächlich gemessenen Variablen und dem eigentlichen Messziel, wobei es sich nicht um Messfehler, sondern um Probleme einer für das Ziel des Surveys unangemessenen Operationalisierung handelt. Zu den Datenaufbereitungsfehlern gehören Fehler durch die Dateneingabe, die Codierung der Antworten, in der Gewichtung und der Datenanalyse. Für Fehler in diesen Stufen einer Erhebung muss auf die entsprechende Literatur verwiesen werden (z. B. Schnell u. a. 2013, 420–429).

Anteilswerte oder Mittelwerte), die für alle möglichen Stichproben der Größe *n* aus einer Population der Größe *N* berechnet werden.[27]

Für den Anteilswert ergibt sich der Standardfehler aus dem Anteilswert und der Stichprobengröße:

$$se(p) = \sqrt{\frac{p(1-p)}{n}} \qquad (4)$$

Für die Präzision der Schätzung spielt die Größe der Population (*N*) keine Rolle, sondern lediglich die Größe der Stichprobe (*n*) und die Größe des Anteilswerts (*p*). Es ist demnach also für die Präzision der Schätzung unerheblich, ob beispielsweise die Viktimisierungsrate für ein bestimmtes Delikt für die Bundesrepublik Deutschland oder eine einzelne Großstadt geschätzt werden soll.[28] Diese Tatsache scheint für Laien schwer akzeptabel zu sein: Eine Stichprobe für eine Großstadt darf nicht kleiner sein als eine Stichprobe für ein gesamtes Land. Diese unangenehme Konsequenz ist mathematisch ebenso unzweifelhaft wie der Politik nur schwierig zu vermitteln.

Die Breite eines Konfidenzintervalls wird weiterhin durch die Festlegung der Irrtumswahrscheinlichkeit bestimmt. Wird diese zu groß gewählt (z. B. 50 %), ist das Konfidenzintervall zwar sehr schmal, wird aber den Populationsparameter für die Hälfte der realisierten Konfidenzintervalle nicht enthalten. Wird die Irrtumswahrscheinlichkeit hingegen zu klein gewählt (0,001 %), so wird zwar nahezu jedes realisierte Konfidenzintervall den Populationsparameter enthalten, die Konfidenzintervalle werden aber wegen ihrer großen Breite faktisch unbrauchbar. Aus diesem Grund wird die Irrtumswahrscheinlichkeit üblicherweise auf 5 % festgelegt. Dieser Wert gibt die Wahrscheinlichkeit an, einen Alpha-Fehler (auch „Fehler erster Art“) zu begehen (siehe hierzu Fahrmeir u. a. 2007, 415–417). In der Regel werden also 95-%-Konfidenzintervalle verwendet.[29]

---

27 In Veröffentlichungen außerhalb der Statistik ist die Verwendung von Groß- und Kleinbuchstaben häufig irreführend (dies wird durch die Verwendung von Textprogrammen wie Word, welche die Großschreibung von Einzelbuchstaben häufig fälschlich erzwingen, befördert). Eindeutig und korrekt hingegen ist die Verwendung von Großbuchstaben für Kennzahlen der Grundgesamtheit und von Kleinbuchstaben für Kennzahlen einer Stichprobe.

28 Dies gilt mit einer unbedeutenden Einschränkung: Ist die Population sehr klein oder die Stichprobe sehr groß ($n/N \geq 0.05$), werden die Ergebnisse *präziser* als in Formel (4) angegeben. In diesen Fällen verkleinert sich das Konfidenzintervall (für den Mittelwert und bei einfachen Zufallsstichproben) dann um $\sqrt{1 - n/N}$. Einzelheiten zu dieser „finiten Populationskorrektur“ (fpc) können in mathematischen Lehrbüchern zur Stichprobentheorie (z. B. bei Lohr 2010) nachgelesen werden.

29 Da die Summe aus Konfidenzniveau (Irrtumswahrscheinlichkeit) und Signifikanzniveau 100 % beträgt, ergibt ein Signifikanzniveau von 5 % somit ein Konfidenzniveau von 100 % – 5 % = 95 %.

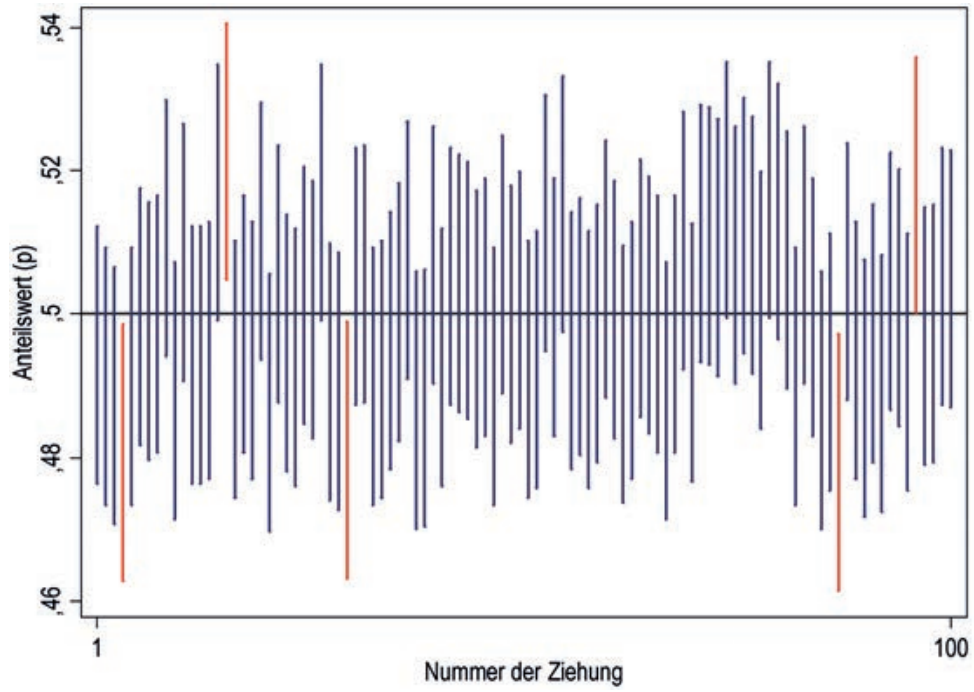
Zur Verdeutlichung: Die Formel zur Schätzung eines 95-%-Konfidenzintervalls des Anteilswerts ist über

$$\left[p - 1.96\sqrt{\frac{p(1-p)}{n}}, p + 1.96\sqrt{\frac{p(1-p)}{n}}\right] \qquad (5)$$

gegeben. Werden nun 100 Stichproben der Größe n = 3.000 aus einer Population der Größe N = 100.000 gezogen, werden die meisten der 100 berechneten Konfidenzintervalle (ca. 95) den Populationsparameter enthalten, einige wenige (ca. 5) aber auch nicht (*Abbildung 4*).[30]

Abbildung 4:

95-%-Konfidenzintervalle für Anteilswerte aus 100 verschiedenen einfachen Zufallsstichproben (*n* = 3.000) aus der gleichen Grundgesamtheit mit π = 0,05. Fünf Konfidenzintervalle enthalten (zufällig) den Populationsmittelwert π = 0,05 nicht. In der Abbildung sind dies die zwei Intervalle mit den höchsten und die drei Intervalle mit den niedrigsten Intervallgrenzen.

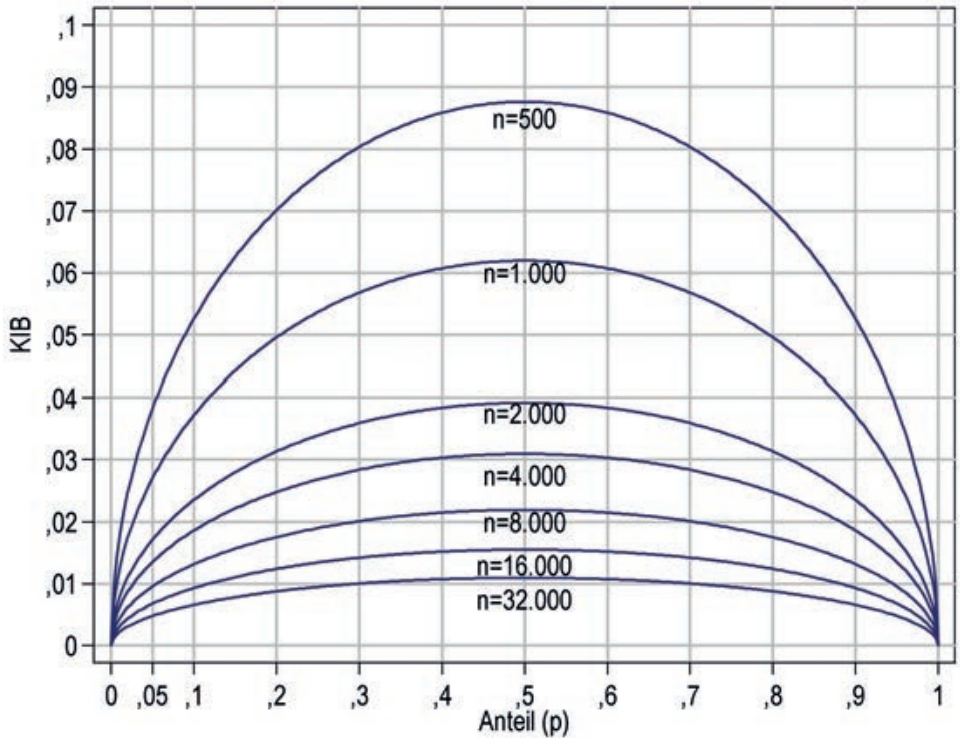


[30] In diesem Zusammenhang ist es sinnlos, davon zu sprechen, dass ein realisiertes Konfidenzintervall den Populationsparameter mit 95 % Wahrscheinlichkeit enthält. Der Populationsparameter liegt entweder innerhalb der Grenzen des Konfidenzintervalls oder nicht.

Die Breite der Konfidenzintervalle ist dabei davon unabhängig, ob die Population 10.000, 100.000 oder 80.000.000 Elemente umfasst, sie ist nicht von *N*, sondern von der Stichprobengröße *n* abhängig. Der Zusammenhang zwischen Konfidenzintervallbreite und Größe der Stichprobe kann in einem Nomogramm dargestellt werden (*Abbildung 5*).

Abbildung 5:

Breite des Konfidenzintervalls (KIB) für gegebene Anteilswerte (p) und verschiedene Stichprobengrößen *n* (in Anlehnung an Schnell/Hoffmeyer-Zlotnik 2002)



Die Breite eines Konfidenzintervalls für Anteilswerte ist maximal für $p = 0{,}50$. Für eine Stichprobe mit n = 1.000 Fällen ergibt sich demnach eine Breite des 95-%-Konfidenzintervalls für Anteilswerte von über 6 %. Um die Breite auf 1 % zu reduzieren, wäre bereits eine Stichprobengröße von über 38.000 Fällen notwendig.[31]

Um einen Anteilswert mit einer bestimmten Genauigkeit schätzen zu können, werden also zwei Dinge benötigt: erstens eine möglichst genaue Vorstellung über die Größe des Anteilswerts sowie zweitens die gewünschte Irrtumswahr-

[31] Generell gilt die Faustregel, dass eine Halbierung der Breite eines Konfidenzintervalls eine Vervierfachung der Stichprobengröße erfordert.

scheinlichkeit. Wenn hierfür Zahlen festgelegt wurden, kann die Breite des Konfidenzintervalls über

$$KIB = 2 * z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \quad (6)$$

geschätzt werden, wobei $z_{\alpha/2}$ den Z-Wert der inversen Standardnormalverteilung für die gegebene Irrtumswahrscheinlichkeit α (für gewöhnlich 5 %) darstellt (Bortz 2005, 104, Formel 3.24). Stellt man diese Formel um, lässt sich der benötigte Stichprobenumfang mit

$$n = \frac{4*z_{\alpha/2}^2 p(1-p)}{KIB^2} \quad (7)$$

schätzen, wobei *KIB* hier die gewünschte Breite des Konfidenzintervalls (z. B. 0,01 für eine Breite von 1 % oder 0,05 für eine Breite von 5 %) bezeichnet (Bortz 2005, 104, Formel 3.26).

8 Designeffekte

Den bisherigen Überlegungen liegt die Annahme zugrunde, dass es sich um einfache Zufallsstichproben handelt. Dies ist jedoch faktisch für keinen bundesweiten Survey der Fall, mit dem Aussagen über die allgemeine Bevölkerung getroffen werden sollen. Das Design solcher Surveys umfasst üblicherweise die Schichtung, Klumpung oder Auswahl der Populationselemente in mehreren Stufen.

Werden beispielsweise natürlich vorkommende räumliche Einheiten als Klumpen für die Stichprobenziehung verwendet, resultieren aufgrund des sogenannten Klumpeneffekts im Normalfall weniger präzise Ergebnisse, als bei Verwendung einer einfachen Stichprobe gleicher Größe (Kish 1965, 164). Die Ursache liegt darin, dass Personen mit einem ähnlichen soziodemografischen Hintergrund dazu tendieren, in der gleichen Nachbarschaft zu leben. Dies führt zu einer größeren klumpeninternen Homogenität als bei rein zufälligem Siedlungsverhalten.

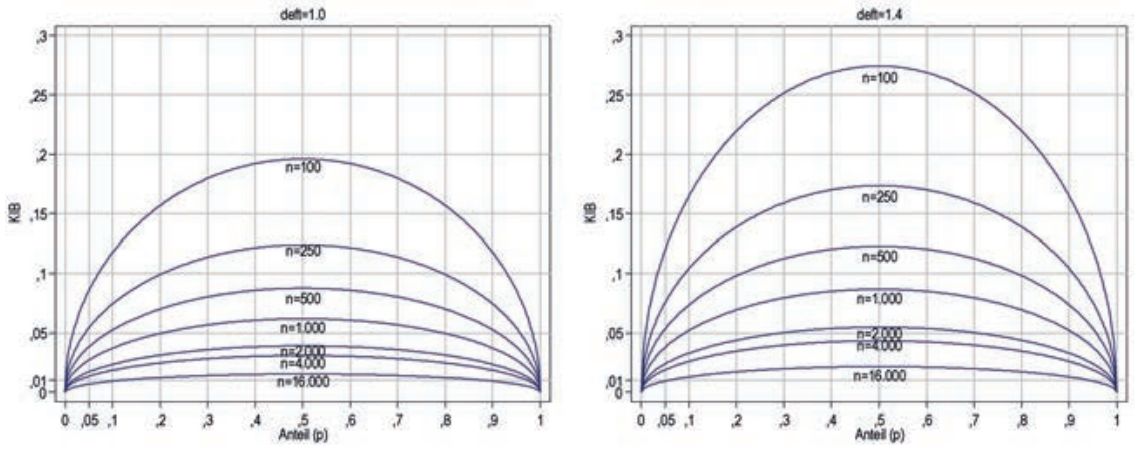
Als Beispiele können das Familieneinkommen (Converse/Traugott 1986, 1095) oder die Fragen nach vorhandenen „Incivilities“ in der jeweiligen Nachbarschaft (Schnell/Kreuter 2005, 401) angeführt werden. Generell bezeichnet man die Vergrößerung der Konfidenzintervalle bei komplexen Stichproben durch Klumpung, Schichtung und Gewichtung sowie einige andere Faktoren als Designeffekt.

Liegen solche Designeffekte vor, dann ist die Berechnung von Konfidenzintervallen über Formel (5) sowie die naive Berechnung statistischer Tests nicht mehr korrekt. In nahezu allen Fällen führen Designeffekte zu größeren Standardfehlern und damit auch zu konservativeren statistischen Tests, also weniger fälschlich signifikanten Ergebnissen.[32]

Die Auswirkungen solcher Designeffekte werden in den Abbildungen 6 und 7 dargestellt.[33] *Abbildung 6* gibt die Breite der 95-%-Konfidenzintervalle für Anteilswerte in Abhängigkeit von der Stichprobengröße sowie des Anteilswerts p an. Hier zeigt sich, dass die Konfidenzintervalle umso breiter werden, je kleiner die Stichprobe ausfällt und je näher der Anteilswert an $p = 0{,}5$ liegt. Ergänzend ist deutlich zu erkennen, dass sich die Breite der jeweiligen Konfidenzintervalle deutlich erhöht, wenn nicht von einer einfachen Zufallsstichprobe ausgegangen wird (linke Abbildung), sondern ein komplexes Stichprobendesign mit einem Designeffekt von $deft = \sqrt{deff} = 1{,}4$ angenommen wird (rechte Abbildung).[34]

Abbildung 6:

Breite des Konfidenzintervalls (KIB) für gegebene Anteilswerte (p) und verschiedene Stichprobengrößen n bei unterschiedlichen Designeffekten (deft=1.0 und deft=1.4)



---

32 Aus diesem Grund sind die naiven Varianzschätzungen wie z. B. in Ahlborn u. a. 1993 nicht vertretbar. Dort wird argumentiert, dass der Klumpeneffekt durch Schichtung kompensiert werden könne. Dies ist mathematisch zwar denkbar, dürfte aber bei kaum einer Anwendung möglich sein. Bei kriminologischen Fragestellungen kann eine solche Kompensation für die allgemeine Bevölkerung nahezu ausgeschlossen werden. Die Formeln und Ergebnisse bei Ahlborn u. a. 1993 sollten daher nicht für die Planung von Viktimisierungsstudien verwendet werden.

33 Die Beispiele sind an Schnell/Hoffmeyer-Zlotnik 2002 angelehnt.

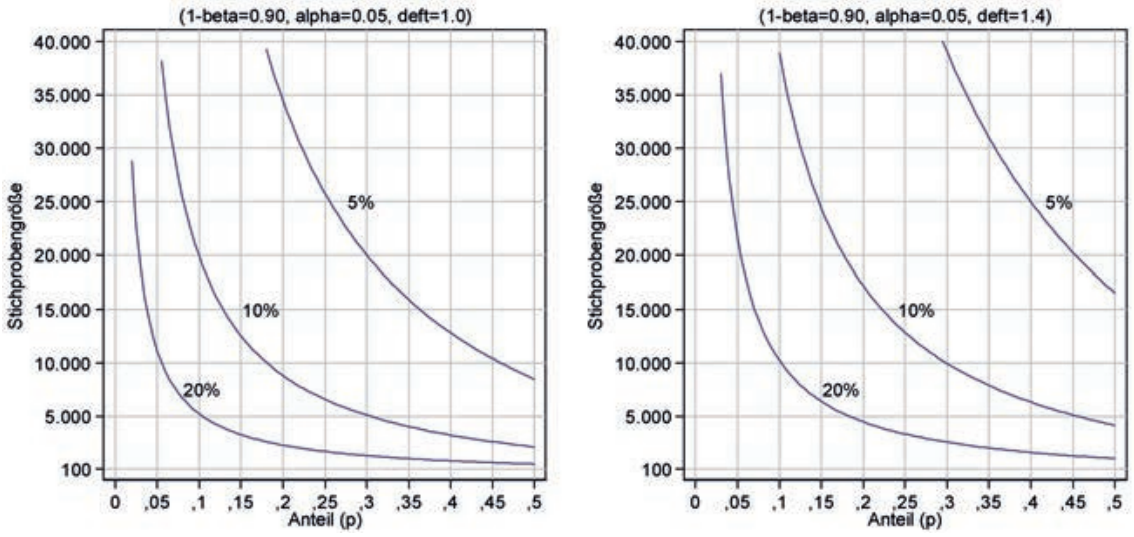
34 Der Wert $deft = 1{,}4$ stellt in der Survey-Literatur die gängige „rule of thumb" dar (Schnell/Kreuter 2005, 390). Zur Berechnung von $deff$ siehe die Formeln (8) und (10).

Wird demnach für ein Merkmal mit einer Prävalenzrate von 5 % in einem Survey mit einem Designeffekt $deft$ = 1,4 ein 95-%-Konfidenzintervall berechnet, so liegt die tatsächliche Breite dieses Konfidenzintervalls bei einer Stichprobengröße von n = 2.000 nicht bei 1,91 % (linke Abbildung), sondern bei 2,67 % (rechte Abbildung). Um unter diesen Bedingungen ein 1-%-Punkt breites Konfidenzintervall zu erhalten, ist eine Stichprobengröße von n = 14.307 erforderlich. Bei einer einfachen Zufallsstichprobe ohne Designeffekt wären hingegen bereits n = 7.300 Fälle ausreichend.

Ist nicht die einmalige Schätzung einer Prävalenzrate, sondern deren Veränderung von Interesse, kann die zu einer Entdeckung dieser Differenz notwendige Stichprobengröße in *Abbildung 7* abgelesen werden. *Abbildung 7* zeigt die benötigte Fallzahl in Abhängigkeit vom Anteilswert, um eine relative Veränderung des Anteilswerts um x % mit einer Wahrscheinlichkeit von $1 - \beta = 0{,}9$ auch zu entdecken.[35]

Abbildung 7:

Benötigte Fallzahl in Abhängigkeit des Anteilswerts und der relativen Veränderung in Prozent (alpha = 0,05; 1-beta = 0,9; deft = 1,0/1,4; Schnell/ Hoffmeyer-Zlotnik 2002)



35 Die Wahrscheinlichkeit, einen tatsächlich vorhandenen Effekt in einer Stichprobe auch zu entdecken, wird in der Statistik als *Power* bezeichnet und ist als $1 - \beta$ definiert, wobei β die Wahrscheinlichkeit für einen Fehler der zweiten Art ist. Ein Fehler der zweiten Art ist die Beibehaltung der Nullhypothese, obwohl sie falsch ist. Die Berechnung der *Power* ist relativ aufwendig, da Stichprobengröße, Effektstärke und Irrtumswahrscheinlichkeit bekannt sein müssen. In der Praxis wird in der Regel eine *Power* von über 0,9 angestrebt. Wie man vor allem anhand der *Abbildung 7* sieht, erfordert eine *Power* von 0,9 in der Regel deutlich größere Stichproben als gemeinhin unter Laien vermutet.

Um die Veränderung eines Anteilswerts von ρ = 5 % um 20 % (also von 5 % auf 5 % × 1,20 auf 6 %) mit einer Wahrscheinlichkeit von 90 % (1-beta) entdecken zu können, ist bei einem Designeffekt von deft = 1,0 eine Stichprobengröße von *n*= 11.120 erforderlich. Liegt ein Designeffekt von 1,4 vor, so erhöht sich die notwendige Stichprobengröße auf 21.684.[36] Sollen kleinere Veränderungen entdeckt werden, so ist dies nur durch eine deutliche Vergrößerung der Stichprobe zu erreichen.

Bislang haben wir keine Möglichkeit vorgestellt, die Größe des Designeffekts zu berechnen. Exakt ist ein Designeffekt ($deff$) definiert als Quotient des Standardfehlers $\widehat{\sigma^2}_{\theta,SRS}$ einer einfachen Zufallsstichprobe (SRS) und des Standardfehlers der gegebenen komplexen Stichprobe $\widehat{\sigma^2}_{\theta,Komplex}$, wobei θ für einen beliebigen Parameter (z. B. μ oder π) steht. Der Designeffekt $deff$ ist also gleich

$$deff = \frac{\widehat{\sigma^2}_{\theta,Komplex}}{\widehat{\sigma^2}_{\theta,SRS}} \qquad (8)$$

Werte größer als 1 zeigen eine geringere Präzision des komplexen Designs im Vergleich zu einer einfachen Zufallsstichprobe gleicher Größe. Häufig ist es einfacher, mit der Wurzel aus dem Designeffekt $deff$ zu rechnen, die als $deft$ bezeichnet wird.

Dies kann am Beispiel der angesprochenen Konfidenzintervalle verdeutlicht werden. Die korrekt berechneten Konfidenzintervalle für komplexe Designs verbreitern sich im Vergleich zu den naiv berechneten Konfidenzintervallen (siehe Formel (5)) um den Faktor $\sqrt{deff}$.

$$\left[\bar{x} - z_{1-\alpha/2}\sqrt{deff}\,\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\sqrt{deff}\,\frac{s}{\sqrt{n}}\right] \qquad (9)$$

Die Schätzung des Standardfehlers einer gegebenen komplexen Stichprobe $\widehat{\sigma^2}_{\theta,Komplex}$ kann auf verschiedene Arten erfolgen.[37] Die einfachste Art der Schätzung basiert auf dem sogenannten Intraklassenkorrelationskoeffizienten (ρ, auch ICC). ρ gibt die Homogenität des untersuchten Merkmals innerhalb der verwendeten Klumpen an. Je ähnlicher sich die Elemente innerhalb der

---

36 Siehe hierzu Schnell/Hoffmeyer-Zlotnik 2002, 10-13.

37 Die korrekte Schätzung über die verbreitetsten Ansätze wie Resampling-Verfahren (Jackknife oder Bootstrap) oder Taylor-Linearisation ist beispielsweise in Stata über die „svy"-Kommandos oder in R über das Paket „survey" möglich. Für Details zu diesen und weiteren Verfahren siehe Wolter 2007.

Klumpen sind, desto größer fällt ρ aus.[38] Neben ρ wird für die Berechnung des Designeffekts ebenfalls die durchschnittlichen Anzahl der Interviews innerhalb der gezogenen Klumpen $\bar{b}$ benötigt. Sind beide Größen bekannt, kann $deff$ über

$$deff = 1 + \rho(\bar{b} - 1) \tag{10}$$

berechnet werden (Kish 1965, 162; Lohr 2010, 174). Diese Größe des Designeffekts hängt also nicht nur von der Homogenität der Klumpen, sondern auch ihrer Größe ab. Die Verwendung großer Klumpen kann also ebenfalls zu einem deutlichen Präzisionsverlust führen.

Da nicht nur räumliche Einheiten als Klumpen angesehen werden können, sondern auch die in einer Studie beteiligten Interviewer, ist auch die Zahl der zu bearbeitenden Fälle pro Interviewer (*Workload*) für den Designeffekt von Interesse. Dies betrifft insbesondere CATI-Studien (*Computer Assisted Telephone Interview*), bei denen hohe Interviewer-Workloads in der Praxis häufig vorkommen.

Daher kann sich auch dann ein großer Wert für $deff$ ergeben, wenn die Klumpen intern zwar heterogen sind ($\rho \approx 0$), die durchschnittliche Anzahl der Interviews pro Interviewer aber ausreichend groß ist, um die geringen Werte für ρ auszugleichen (Schnell/Kreuter 2005, 394). Für CATI-Studien stellen ρ = 0,001 und $\bar{b}$ = 70 übliche Werte dar (Tucker 1983; Groves/Magilavy 1986; Groves 1989). Für diese Werte ergibt sich ein Designeffekt von $deff$ = 1 + 0,01 * (70 – 1) = 1,69 (Schnell/Kreuter 2005, 394). Die effektive Stichprobengröße

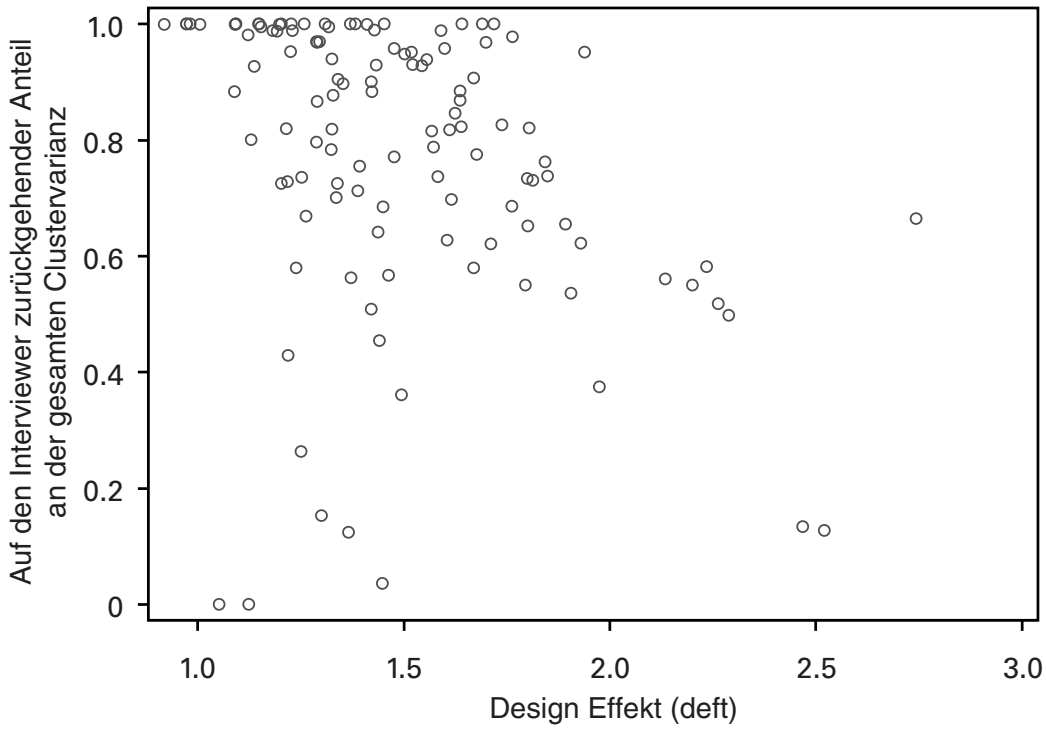
$$n' = \frac{n}{deff} \tag{11}$$

für diese Werte ergibt, dass eine komplexe Stichprobe der Größe n = 1.000 mit einem Designeffekt von 1,69 zu einer effektiven Stichprobengröße von $n' = \frac{1.000}{1,69} = 591,7$ führt. Das bedeutet, dass die Stichprobe, obwohl 1.000 Personen befragt wurden, effektiv so ungenau ist, als wären lediglich 591 Personen befragt worden.

[38] Der Intraklassenkorrelationskoeffizient kann beispielsweise über Varianzanalysen oder Mehrebenenmodelle berechnet werden. Weitere Details finden sich bei Lohr 2010, 174.

Abbildung 8:

Anteil des Interviewers am Designeffekt und die Größe des Designeffekts für 118 Fragen aus dem Defect-Projekt (Schnell/Kreuter 2005, 402)



Überdies muss beachtet werden, dass Designeffekte in von Interviewern durchgeführten Studien mit räumlicher Klumpung sowohl auf die räumlichen als auch die interviewerbedingten homogenisierenden Effekte zurückgehen. In diesem Zusammenhang konnten Schnell und Kreuter (2005) unter Nutzung eines zur Trennung dieser beiden Effekte notwendigen speziellen Stichprobendesigns (interpenetrierende Stichproben) zeigen, dass der Designeffekt für viele Merkmale stärker auf die Interviewer zurückgeht als auf die räumliche Klumpung (*Abbildung 8*). Das Verhalten der Interviewer wirkt sich also auf die tatsächliche Größe der Konfidenzintervalle aus. Diese Tatsache wird in der Forschungspraxis und in der mathematischen Statistik häufig ignoriert. Die Konsequenz ist eine Überschätzung der Genauigkeit statistischer Schätzungen auf der Basis realisierter Stichproben.

9 Mehrstufige Auswahlverfahren für verschiedene Erhebungsmodi

In den meisten Projekten der empirischen Sozialforschung erfolgt die Wahl des Erhebungsmodus nahezu ausschließlich anhand der erwarteten Kosten,

sodass telefonische oder postalische Erhebungen die Regel sind. Stehen die Mittel für die erheblich teureren persönlichen (Face-to-Face-)Befragungen zur Verfügung, dann kann dieser Erhebungsmodus dann eingesetzt werden, wenn telefonische oder postalische Erhebung aus methodischer Sicht ausscheiden (z. B. falls Dokumente herangezogen werden sollen oder die Befragten weder telefonisch noch schriftlich erreicht oder befragt werden können). Für diese Erhebungsmodi stehen geeignete Auswahlverfahren zur Verfügung, die wir im Folgenden detaillierter darstellen werden. Für Web-Befragungen gilt das nicht. Weder existieren geeignete Listen, aus denen ausgewählt werden kann, noch können Auswahlwahrscheinlichkeiten berechnet werden. Entsprechend besteht keine Möglichkeit, auf der Basis eines Web-Surveys auf eine „allgemeine Bevölkerung“ zu schließen. Daher raten wir auf absehbare Zeit von der Verwendung von Web-Befragungen für Bevölkerungserhebungen ab.[39]

9.1 Telefonische Befragungen

Die Stichprobenziehung für telefonische Befragungen (CATI) ist nicht trivial. Üblicherweise muss die Ziehung in mehreren Stufen erfolgen, da im Allgemeinen keine vollständige Liste aller Telefonanschlüsse vorliegt. Aus diesem Grund muss mit einer Technik gearbeitet, die keine vollständigen Listen voraussetzt.

9.1.1 Random Digit Dialing

Die Grundidee der Random-Digit-Dialing-Technik (RDD) lässt sich am einfachsten am Beispiel der USA erläutern: US-amerikanische Telefonnummern sind zehnstellig, wobei die ersten drei Ziffern einer Region und die nächsten drei Ziffern einer Vermittlungsstelle entsprechen. Die letzten vier Ziffern bilden zusammen mit den ersten sechs Nummern die Teilnehmernummer. Da zwar keine vollständigen Listen von Telefonnummern, aber vollständige Listen „aktiver“ Region-Vermittlungsstelle-Kombinationen (die ersten sechs Stellen) vorliegen, können diese zur Stichprobenziehung verwendet werden. Die einfachste Variante von RDD sieht dann vor, eine Zufallsstichprobe aus der Liste der Region-Vermittlungsstelle-Kombinationen zu ziehen und die Teilnehmernummern durch das Anhängen einer vierstelligen Zufallszahl zu

[39] Einzelheiten finden sich bei Schnell 2012.

generieren. Dieser Ansatz ist allerdings recht ineffizient, da hier viele Nummern generiert werden, die keinem Privathaushalt zugeordnet sind. Diese nicht zielführenden Nummern machen ca. 80 % der generierten Nummern aus (Schnell u. a. 2013, 280–281).

Eine effizientere, zweistufige Methode besteht in der 1970 von Mitofsky vorgeschlagenen und 1978 von Waksberg weiterentwickelten sogenannten Mitofsky-Waksberg-Methode. Diese besteht in einem ersten Schritt aus der Einteilung der letzten vier Ziffern in 100er-Blöcke (andere Einteilungen sind ebenfalls möglich), zum Beispiel 678-560-0000 bis 678-560-0099 (Link/Fahimi 2008). Handelt es sich bei der ersten antelefonierten Nummer aus einem der zufällig ausgewählten 100er-Blöcke (z. B. 678-560-0054) um einen Privatanschluss, so werden in einem zweiten Schritt weitere zufällig gewählte Nummern innerhalb des Blocks antelefoniert, ansonsten scheidet der gesamte Block aus (Schnell u. a. 2013, 281).

Aufgrund der technischen Details bei der Vergabe der Telefonnummern durch die Telekom ist eine Stichprobenziehung über das RDD-Verfahren in Deutschland in dieser Weise nicht möglich. Alternativ finden deshalb Telefonbücher oder Telefon-CDs als Auswahlgrundlage Verwendung. Als erste Auswahlstufe werden hier die Ortsnetze der Telekom verwendet, als zweite Auswahlstufe dann eine Zufallsstichprobe aus den in den Telefon-CDs eingetragenen Nummern gezogen. In einem dritten Schritt wird die zu befragende Person ausgewählt. Dies kann über eine vereinfachte Version einer als „Schwedenschlüssel“ bekannten Zufallsauswahl oder über die Frage, welche Person als letzte Geburtstag hatte (*last birthday method*), geschehen (Schnell u. a. 2013, 281–282).

Ein überaus gewichtiger Punkt ist bei dieser Vorgehensweise die Auswahl der Telefonnummern im zweiten Auswahlschritt. Beschränkt sich die Auswahl auf die in den Telefon-CDs gelisteten Nummern, so führt dies zum Verlust aller nicht eingetragenen Telefonnummern. Neben Personen ohne Festnetzanschluss (kein Telefon oder nur Mobiltelefon, hierzu Abschnitt 9.1.2) betrifft dies insbesondere Personen ohne eingetragene Nummer. Dieses Vorgehen ist nicht vertretbar. Somit empfiehlt sich ein Verfahren, in dem auch die nicht in den Telefonbüchern oder Telefon-CDs vorhandenen Nummern berücksichtigt werden. Dies geschieht durch über bestimmte Verfahren generierte zusätzliche Telefonnummern.

Eine Möglichkeit besteht in der Addition einer Zufallszahl zu einer existierenden, zufällig aus der Telefon-CD ausgewählten Telefonnummer (*randomized last digit*, RLD), eine weitere darin, die letzten beiden Ziffern einer exis-

tierenden Nummer durch zufällig generierte Nummern zu ersetzen (Schnell u. a. 2013, 281–282).[40]

Sowohl der sinkende Anteil in den Telefonbüchern erfasster Festnetznummern als auch die steigende Anzahl von Personen, die nur noch über ein Mobiltelefon verfügen (*Cell-Phone-Only*, CPO),[41] stellen ein Problem für telefonische Befragungen dar. Aus diesem Grund werden in der Bundesrepublik vermehrt andere Auswahlverfahren für telefonische Befragungen (Dual-Frame-Verfahren) als die bisher beschriebenen eingesetzt (Schnell u. a. 2013, 282).

9.1.2 Dual-Frame-Stichproben zur Berücksichtigung von Mobiltelefonen

Die steigenden Teilnehmerzahlen im Bereich des Mobilfunks sowie die Auflösung des Telekom-Monopols hatten erhebliche Auswirkungen auf die Stichprobenziehung für telefonische Befragungen. Um die daraus resultierenden Probleme zu lösen, werden in Deutschland seit 2007 die erst seit wenigen Jahren verfügbaren Daten der Bundesnetzagentur über an die Telekommunikationsanbieter vergebene Nummern zur Stichprobenziehung verwendet.[42] Der aus diesen Daten gebildete Nummernraum umfasst alle überhaupt nutzbaren Telefonnummern, nicht nur die tatsächlich verwendeten Nummern sowohl für Festnetz als auch Mobilfunk.

---

40 Dieses auf die letzten vier Ziffern angewendete Verfahren stellte von 1999 bis 2007 die Basis der vom „Arbeitskreis deutscher Marktforschungsinstitute“ (ADM) verwendeten Telefonstichproben dar. Bei dieser Variante werden allerdings viele Nummern generiert, die entweder nicht vergeben („kein Anschluss unter dieser Nummer“) oder keine Privatanschlüsse sind. Obwohl solche Varianten des Mitofsky-Waksberg-Designs bereits 1988 von Lepkowski ausführlich diskutiert wurden, wird dieses Verfahren im deutschsprachigen Raum häufig irreführend nach den ersten Anwendern dieses Verfahrens in der BRD als Gabler-Häder-Design bezeichnet (Schnell u. a. 2013: 282). Seit 2007 basieren die ADM-Telefonstichproben auf den von der Bundesnetzagentur vergebenen Rufnummernblöcken (Heckel u. a. 2014, 142–144) und sind damit nicht mehr als modifiziertes Mitofsky-Waksberg-Design zu verstehen.

41 Nach Berechnungen von Hunsicker/Schroth (2014, 9) mit Daten der Forschungsgruppe Wahlen und des Politbarometers hat sich der Anteil Wahlberechtigter, die nur noch über ein Mobiltelefon verfügen, von 8 % im Jahr 2006 auf 14 % in den Jahren 2012/2013 nahezu verdoppelt. Weitere dort angeführte vergleichbare Dual-Frame-Studien weisen für unterschiedliche Grundgesamtheiten Werte zwischen 11 % und 12 % aus (INFAS 12 %, ADM 12,4 %, Cella 2 11 %). Weitere Hinweise auf eine Zunahme der CPO-Problematik sowie deutliche Unterschiede zwischen den Ländern in Europa für 2006 und 2010 auf Basis verschiedener „Special Eurobarometer: E-Communications Household Surveys“ geben Heckel/Wiese 2012, 111. Aufgrund der spärlichen Informationen hinsichtlich der Datenerhebung und der somit unklaren methodischen Güte dieser Surveys sind diese Aussagen allerdings nicht als endgültig gesichert anzusehen.

42 Für Details hinsichtlich des Vorgehens des ADM siehe Glemser u. a. (2014), zu Details der Erstellung der Auswahlgrundlage siehe Heckel u. a. (2014).

Damit liegt es nahe, jeweils eine Auswahlgrundlage für das Festnetz und eine Auswahlgrundlage für die Mobilfunknetze zu konstruieren und diese dann zu kombinieren. Solche Kombinationen zweier Auswahlgrundlagen werden in der Stichprobentheorie allgemein als *Dual-Frames* bezeichnet. Da eine Person aber sowohl mehrere Festnetz- als auch Mobilnummern besitzen kann, müssen Dual-Frame-Stichproben erst derart gewichtet werden, dass jede Person die gleiche Auswahlwahrscheinlichkeit besitzt (Schnell u. a. 2013, 282–283). In der Praxis in Deutschland basieren diese Gewichtungen auf einer Reihe plausibler Annahmen, so z. B. dass ein Mobiltelefon exakt einer Person sowie ein Festnetzanschluss jeder Person im Haushalt zugeordnet ist. Weiter wird angenommen, dass die Wahrscheinlichkeit, parallel über den Festnetz- und den Mobiltelefonframe gleichzeitig ausgewählt zu werden, gleich null ist. Unter diesen Annahmen lässt sich die kombinierte Auswahlwahrscheinlichkeit als

$$\pi_i = k_i^F \frac{n^F}{N^F} \frac{1}{z_i} + k_i^C \frac{n^C}{N^C} \qquad (12)$$

berechnen (Häder u. a. 2009, 29). Dabei stellt k_i^F die Zahl der Nummern dar, unter der ein Haushalt über das Festnetz erreicht werden kann, k_i^C bezeichnet die Zahl der Nummern, unter der eine Person über Mobiltelefone erreicht werden kann, *n* ist die Zahl der Telefonnummern in der Stichprobe (F oder C), *N* die Zahl der Telefonnummern in der Grundgesamtheit (F oder C) und z_i die Zahl der Personen im Haushalt.

Entscheidend für die Validität der derart berechneten Auswahlwahrscheinlichkeit ist die Gültigkeit der Angaben k_i^F, k_i^C und z_i, die von den subjektiven Wahrnehmungen der Befragten abhängen und somit Raum für Fehler lassen. Damit sind diese Werte prinzipiell schon aufgrund der potenziellen Unkenntnis der Befragten als fehleranfällig anzusehen (Schnell 2012, 273). Die daraus resultierenden Probleme sind nicht abschließend geklärt.

9.2 Schriftliche und postalische Befragung

Gelegentlich werden schriftliche und postalische Befragungen miteinander verwechselt, obwohl es sich um klar abgrenzbare Erhebungsformen handelt.

Schriftliche Befragungen erfolgen innerhalb von Organisationen, bei denen vollständige Listen der zu befragenden Personen vorliegen.[43] Dabei handelt

[43] Schriftliche Befragungen ohne solche Listen entsprechen einer Stichprobe ohne bekannte Auswahlregel, sind also wie die entsprechenden Web-Surveys willkürliche oder bewusste Stichproben. Damit sind sie keine geeignete Basis für verallgemeinerbare Aussagen.

es sich um Organisationen mit festen Mitgliedern (Universitäten, Verwaltungen, Schulen). Werden die Personen innerhalb der Organisation befragt (z. B. im Klassenraum, bei Vorlesungen oder gemeinsamen Veranstaltungen), dann handelt es sich um eine individuelle schriftliche Befragung in einer Gruppe. Das Standardbeispiel wäre eine Befragung im Klassenverband oder während einer Vorlesung oder Personalversammlung. In solchen Fällen handelt es sich um Klumpenstichproben, d. h., die Klumpen werden zufällig ausgewählt und innerhalb der Klumpen werden alle schriftlich befragt. Das Auswahlverfahren ist vergleichsweise trivial durchzuführen. Die zu beachtende Regel wurde schon erwähnt: Möglichst viele Klumpen bei gleicher Fallzahl sind besser als wenige große Klumpen. Allerdings müssen Nonresponse-Probleme bei der Auswahl der Klumpen (Beispiel: Alle Mitglieder einer Vorlesung fallen aus, falls der Dozent nicht kooperiert) und der einzelnen Mitglieder bedacht werden. Weiterhin müssen die Klumpeneffekte bei der Analyse beachtet werden: Eine Schülerbefragung von 2.000 einzelnen Personen hat einen kleineren Standardfehler als eine Befragung von 50 Klassen mit je 40 Personen. Schließlich muss beachtet werden, dass das Antwortverhalten in Gruppen besonderen Dynamiken unterliegen kann: In vielen Fällen ist es kaum möglich, eine Zusammenarbeit mehrerer Personen in einer Gruppe zu unterbinden. Bei Studierenden ist dies z. B. in Vorlesungen nur unter Klausurbedingungen möglich, andere Gruppen sind nicht unbedingt disziplinierter. Sollte als Erhebungsform die Befragung in Gruppen gewählt werden, sollte dies daher entsprechend dokumentiert und die Zugehörigkeit jeder Person zu einer Gruppe Bestandteil des Datensatzes werden, da ansonsten keine korrekte Berechnung der Standardfehler möglich ist.

Bei postalischen Befragungen liegt fast immer eine Liste der zu befragenden Personen vor.[44] Bei einer bundesweiten Bevölkerungsbefragung würde man in der Regel zunächst eine Schichtung nach Bundesländern vornehmen. Innerhalb der Schichten sollte man eine möglichst hohe Zahl von Gemeinden (in der Regel z. B. 160, 210 oder 240)[45] proportional zu ihrer Einwohnerzahl ziehen (*Probability Proportional to Size*, PPS). In jeder gezogenen Gemeinde würde man eine konstante (kleine) Zahl (in der Regel weniger als 20) Personen aus der Einwohnermeldedatei ziehen. Praktisch steht eine solche bundesweite Ziehung vor dem Problem, dass die Gemeinden nicht kooperieren müssen und es zu entsprechenden Ausfällen auf Gemeindeebene kommt.

---

44 Versuche, bundesweite postalische Befragungen auf Haushaltsebene ohne Einwohnermeldedateien durchzuführen, sind selten. Ein entsprechendes Experiment war Bestandteil des Defect-Projekts (Schnell/Kreuter 2000).

45 Diese Zahlen gehen auf die Entwicklung der Stichproben für die Musterstichprobenpläne des Arbeitskreises Deutscher Markt- und Sozialforschungsunternehmen (ADM) zurück. Zur Begründung der Zahlen und einer Geschichte der Entwicklung diese Zahlen siehe Schnell 1997, 58–59.

Weiterhin variieren die Gebühren für solche Ziehungen in so erheblichem Umfang, dass man gelegentlich schon aus finanziellen Gründen auf einzelne Gemeinden verzichtet. Schließlich ist zu beachten, dass die Dauer der Ziehung in mehr als 160 Gemeinden mehr als ein halbes Jahr dauern kann: Dann sind im Mittel mehr als 5 % der Befragten bereits wieder verzogen. Postalische Befragungen mit Einwohnermeldedateien sind bundesweit also keineswegs unproblematisch.

9.3 Face-to-Face

Da in der Bundesrepublik kein vollständiges Zentralregister für die allgemeine Bevölkerung existiert, ist auch keine Zufallsstichprobe realisierbar, die solch eine Liste als *Sampling-Frame* benötigt. Demnach müssen andere Ansätze zur Konstruktion von Stichproben verwendet werden (Schnell 2012, 204–205).

9.3.1 Random Walks

Für bundesweite Erhebungen hat sich in Deutschland seit Ende der 70er Jahre ein Stichprobenplan als Standard etabliert, der auf die Musterstichprobenpläne des „Arbeitskreises deutscher Markt- und Sozialforschungsinstitute“ (ADM) zurückgeht (für die Historie der ADM-Stichproben siehe Löffler u. a. 2014). Im klassischen ADM-Design wurden zwischen 160 und 240 *Sampling-Points* aus einer Datei von ca. 80.000 Bundestagsstimmbezirken proportional zur Zahl der Stimmberechtigten gezogen. In den ausgewählten Stimmbezirken wurde dann ein *Random Walk* durchgeführt, bei dem eine Person ausgehend von einem zufällig gewählten Startpunkt einen Zufallsweg beschreitet. Die Grundregel für einen solchen Zufallsweg könnte z. B. lauten: Auf der linken Straßenseite gehen, rechts abbiegen wann immer es möglich ist. Im Detail werden die Regelwerke aufwendig (was ist eine Straße, was passiert in Sackgassen etc.), aber im Prinzip könnte ein Zufallsweg entstehen.[46] Die begehende Person listet auf ihrem Zufallsweg dann z. B. jeden dritten Haushalt (in der Regel: Klingeln). Die Liste dieser Haushalte (Name und Adresse) bildet dann die Auswahlgrundlage einer Haushaltsstichprobe.[47] In der Praxis wurden die Begehung und die Befragung häufig derselben Person übertragen (*Standard-Random*), was zu vielen Implementierungsproblemen

---

46 Details finden sich bei Schnell (2012, 206–207).

47 In den Haushalten wurde dann noch eine Person zumindest näherungsweise zufällig ausgewählt, in der Regel mit einer speziellen Zufallszahlentabelle („Schwedenschlüssel“) (siehe hierzu Schnell u. a. 2013, 276).

geführt hat. Durch die Verfügbarkeit digitaler Karten und elektronischer Gebäudedateien ist in modernen Gesellschaften ein *Random Walk* weitgehend obsolet geworden. *Random Walks* sollten für Befragungen nur noch dann verwendet werden, wenn keine digitalen Karten und Gebäudedateien verfügbar sind, z. B. in Entwicklungsländern, Katastrophen- oder Kriegsgebieten.

9.3.2 Einwohnermeldeamtsstichproben

Als Goldstandard gilt für Deutschland seit Mitte der 90er Jahre die Durchführung von Einwohnermeldestichproben.

Zu diesem Zweck ist die Kooperation der Einwohnermeldeämter unverzichtbar. Erschwert wird die Stichprobeziehung aus den Einwohnermeldeämtern dadurch, dass weder der Zugang zu den Daten bundeseinheitlich geregelt ist, noch die Gebührensätze über alle Gemeinden einheitlich sind.[48] Eher im Gegenteil schwanken die Gebührensätze in einem erstaunlichen Ausmaß. Die notwendige Kooperation mit mehreren Hundert Gemeinden, die sowohl frei über die Kooperation als auch die Gebührensätze entscheiden können, führt bei bundesweiten Stichproben neben hohen Kosten auch zu einer langen Dauer für die Stichprobenziehung (Schnell 2012, 194).

Doch obwohl die Daten der Einwohnermeldeämter in der Bundesrepublik als bestmögliche Auswahlgrundlage für Stichproben der allgemeinen Bevölkerung gelten, sind auch diese Daten mit Problemen behaftet.[49] So leiden auch die Daten der Einwohnermeldeämter unter Overcoverage und Undercoverage. Ein vom statistischen Bundesamt im Rahmen der Vorbereitung des Zensus 2011 durchgeführter Registertest (Stichtag 5. Dezember 2001) erbrachte bundesweit 1,7 % Personen, die zwar angetroffen, aber nicht in den Registern gefunden wurden (Undercoverage). Der gegenteilige Fall („Karteileichen“, Overcoverage) belief sich bundesweit auf 4,1 % der Einträge. Unter Berücksichtigung der zeitlichen Verzögerung zwischen Bevölkerungsbewegung und der Aktualisierung der Registereinträge („temporäre Karteileichen“) sinkt dieser Wert auf 2,9 %. Allerdings liegen deutliche Unterschiede hinsichtlich der Raten für Undercoverage (zwischen 1,0 % und 3,1 %) und Overcoverage (zwischen 2,6 % und 8,1 %) zwischen den Bundesländern vor. Ebenso deutliche Unterschiede zeigen sich je nach Größe für die einzelnen Gemeinden, wobei größere Gemeinden auch größere Fehlerraten aufweisen (Schnell 2012,

---

48 Einen detaillierteren Überblick über Einwohnermeldestichproben findet sich in von der Heyde 2014.

49 Für die vielen praktischen Probleme von Zufallsstichproben aus Einwohnermelderegistern siehe Albers 1997.

194; siehe auch Tabelle 1 in Statistische Ämter des Bundes und der Länder 2004, 814). Demnach liegen Coverage-Probleme insbesondere in Großstädten vor. Sollten sich die Coverageraten zwischen soziodemografischen Gruppen unterscheiden, werden diese Anteile verzerrt geschätzt. Da die Daten der Zensus-Testerhebung für wissenschaftliche Analysen nicht zur Verfügung stehen, kann dies nicht geprüft werden. Entsprechende Untersuchungen in den USA und dem Vereinigten Königreich zeigen aber, dass es sich bei unterrepräsentierten Personen eher um mobile und sozial randständige Bevölkerungsgruppen handelt (Schnell 2012, 195). Liegen hier systematisch höhere Viktimisierungsraten vor, so würde die tatsächlich vorliegende Kriminalitätsbelastung über eine Einwohnermeldestichprobe unterschätzt.

9.3.3 Gebäudestichproben

Seit 2006 existiert eine neue amtliche Datenbasis, in der alle Gebäude in Deutschland enthalten sind. Da durch diese Liste jedem Gebäude eine eindeutige Auswahlwahrscheinlichkeit zugeordnet werden kann, ist die Ziehung einer Stichprobe aus der allgemeinen Bevölkerung mit dieser Liste als Auswahlgrundlage möglich.[50] Da Einwohnerzahlen für Aggregate unterhalb der Ebene „Stadt“ in der Praxis aus verschiedenen Gründen schwierig zu erhalten sind, wird dieser Schritt in der vorgeschlagenen Stichprobenziehung umgangen und stattdessen zur Auswahl Wohngebäude verwendet. Die Auswahl von Gebäuden ist sowohl als einfache Zufallsstichprobe als auch als geschichtete und/oder geklumpte Stichprobe möglich. Beispielsweise ist es für eine Face-to-Face-Befragung sinnvoll, Städte als natürlich vorkommende Klumpen im Sampling-Prozess zu verwenden, um die Interviewer-Reisekosten im Vergleich zu einer einfachen Zufallsstichprobe geringzuhalten. Im Hinblick auf die unterschiedliche Anzahl der Wohnungen pro Gebäude in Klein- und Großstädten scheint eine Schichtung bezüglich der Größe der Städte sinnvoll zu sein. Innerhalb dieser Schichten wird die Auswahl einer für alle Städte gleichen Zahl von Gebäuden (*Secondary Sampling Unit*, SSU) aus den Städten (*Primary Sampling Unit*, PSU) per PPS-Sampling empfohlen (Schnell 2008, 7–8).[51]

Für Wohngebäude mit mehreren Wohnungen wird die Auswahl jeweils einer Wohnung empfohlen. Die Zahl der Wohngebäude (ohne Wohnheime) wird in

---

50 Das Design einer bundesweiten Bevölkerungsstichprobe auf Basis der Gebäudedatei geht auf einen Antrag des Erstautors bei der Deutschen Forschungsgemeinschaft (DFG) aus dem Jahr 2007 zurück. Details dieses „G-Plans“ finden sich bei Schnell 2008, 7.

51 Dies entspricht einer PPS-Stichprobe auf Gebäudeebene und erscheint sinnvoller als eine PPS-Stichprobe auf Personenebene, da sich die Gebäudestatistik langsamer ändert, als die Bevölkerung.

Deutschland für das Jahr 2011 auf 19.050.663 Gebäude mit 40.857.381 Wohnungen geschätzt (Statistische Ämter des Bundes und der Länder 2014, 5). Da 82,3 % der Wohngebäude in Deutschland aus höchstens zwei Wohnungen bestehen, ist eine Auswahl der Wohnung innerhalb eines Wohngebäudes mit nur einer Wohnung (65,1 %) nicht notwendig oder in Wohngebäuden mit zwei Wohnungen (17,2 %) durch Münzwurf zu realisieren (diese Zahlen basieren auf den Angaben des Statistischen Bundesamts (Statistische Ämter des Bundes und der Länder 2014, 8)). Für Gebäude mit mehr als zwei Wohnungen (17,7 %) müssen andere Auswahlmechanismen genutzt werden. Schnell (2008, 8) schlägt hierfür vor, die Klingeltafeln per Mobiltelefon zu fotografieren und per MMS an den Supervisor zu senden, der dann die zu kontaktierende Wohnung per einfacher Zufallsziehung ohne Zurücklegen festlegt und diese Information per SMS an den Interviewer übergibt.

Für das Jahr 2012 wird die Anzahl an Privathaushalten vom Statistischen Bundesamt auf 40.656.000 geschätzt. Davon sind 41 % Ein- und 35 % Zweipersonenhaushalte. Damit stellen Haushalte mit drei oder mehr Mitgliedern 24 % aller Haushalte dar. Innerhalb der Haushalte wird die Anzahl aller erwachsenen Haushaltsmitglieder durch die Interviewer erfragt und in ein CAPI-System eingegeben. In Ein-Personen-Haushalten oder Mehrpersonenhaushalten mit nur einer erwachsenen Person ist eine Auswahl nicht notwendig. Für Haushalte mit zwei erwachsenen Personen erfolgt die Auswahl über das CAPI-System mit einer Wahrscheinlichkeit von jeweils 50 % für die Kontaktperson oder die andere erwachsene Person im Haushalt. In Mehrpersonenhaushalten erfolgt die zufällige Auswahl aus einer Liste aller erwachsenen Haushaltsmitglieder durch das CAPI-System (Schnell 2008, 9).[52]

9.4 Web-Surveys

Das zentrale Problem aller Web-Surveys ist die Tatsache, dass es keine geeigneten Auswahlgrundlagen für allgemeine Bevölkerungsstichproben gibt. Für wenige hochspezialisierte Populationen sind vollständige, aktuelle und in aktiver Nutzung befindliche E-Mail-Listen verfügbar, wenngleich sehr seltene Ausnahmen. Damit verbleiben nur wenige Optionen für die Herstellung solcher Listen (Einzelheiten finden sich bei Schnell 2012).

Weitverbreitet ist das Ziehen einer Zufallsstichprobe aus einer administrativen Liste oder einer Telefonstichprobe, die dann online befragt wird, soweit

---

[52] Erfolgt die Auswahl nicht zufällig über eine bestimmbare Regel, sondern willkürlich über die Interviewer, so ist eine Berechnung der Inklusionswahrscheinlichkeiten nicht möglich. Somit handelt es sich bei einer solchen Stichprobe um keine Zufallsstichprobe (siehe Unterkapitel 5.1).

dies möglich ist. Manchmal wird versucht, den ausgewählten Personen einen Internetzugang zu ermöglichen, falls dies vor der Ziehung nicht der Fall war. In der Regel sind die Ausfälle von Offline-Rekrutierungen zur Online-Befragung erheblich[53] und immer systematisch: Sensible Populationen (z. B. illegale, alte und/oder kranke Personen) fallen spätestens bei der Online-Befragung aus.

Bei Studien zur Art der Nutzung sozialer Medien mag dies irrelevant sein, nicht aber für Gesundheitssurveys oder Viktimisierungsbefragungen. Solche systematischen Ausfälle führen immer zu einer Unterschätzung der Viktimisierung und lassen sich kaum durch Gewichtungen kompensieren.[54]

Der Nachweis, dass eine Internet-basierte Befragung einer allgemeinen Bevölkerung zu vergleichbaren Resultaten wie eine dem Stand der Surveymethodologie entsprechende Zufallsstichprobe führt, steht weltweit noch aus und ist auf lange Zeit angesichts der selektiven Internetnutzung (Schnell 2012) in der allgemeinen Bevölkerung nicht zu erwarten. So lautet eine klare Empfehlung der *American Association for Public Opinion Research* (AAPOR) im Hinblick auf Online-Befragungen: „Researchers should avoid nonprobability online panels when one of the research objectives is to accurately estimate population values“ (AAPOR 2010, 758).

Wir raten daher dringend von der Verwendung von Internetsurveys bei Viktimisierungsstudien ab.

10 Zum Begriff der Repräsentativität

Außerhalb der fachlichen Diskussion wird Repräsentativität als ein wichtiges Merkmal für die Beschreibung von Stichproben begriffen. Hiermit soll ausgedrückt werden, dass die Verteilung der in einer Stichprobe vorliegenden Merkmale deren Verteilung in der Grundgesamtheit entspricht. Doch im Gegensatz zur landläufigen Meinung stellt ‚Repräsentativität‘ keinen in der Stichprobentheorie verwendeten Begriff dar.[55]

---

53 Für Deutschland berichten Bandilla u. a. 2009 von 11 % der Befragten des Allbus 2006, die tatsächlich online an einer Befragung teilnahmen.

54 An dieser Stelle muss darauf hingewiesen werden, dass die Beweislast für die Gültigkeit einer Methode bei denen liegt, die eine neue Methode vorstellen, nicht umgekehrt. Ein solcher Nachweis kann auch nicht durch das einmalige Präsentieren eines günstigen Ergebnisses erbracht werden, da bei vielen Versuchen ein Ergebnis immer zufällig korrekt sein könnte.

55 Einzelheiten zur nahezu ausschließlich missbräuchlichen Verwendung des Begriffs ‚Repräsentativität‘ finden sich in einer Artikelserie bei Kruskall (1979a; 1979b; 1979c; 1980).

Die einzige Möglichkeit, die Übereinstimmung der Merkmalsverteilung zwischen Stichprobe und Grundgesamtheit innerhalb berechenbarer Fehlergrenzen sicherstellen zu können, liegt in der Verwendung von Zufallsstichproben. Nur auf diese Weise sind die beiden Begriffe ‚repräsentativ' und ‚Zufallsauswahl' synonym (Schnell u. a. 2013, 296). Die Bezeichnung einer nicht zufällig gezogenen Stichprobe wie z. B. einer Quotenstichprobe als repräsentativ ist demnach bedeutungslos. Auch der Nachweis, dass die Verteilung einiger Merkmale einer Quotenstichprobe oder eines Websurveys der Verteilung der Merkmale in der Grundgesamtheit entspricht, sagt nichts über die Verteilung der restlichen in der Stichprobe vorliegenden Merkmale aus. Beispielsweise sagt die Unverzerrtheit demografischer Merkmale – wie Alter, Geschlecht, Bildungsstand – nichts über die Unverzerrtheit anderer inhaltlich relevanter Merkmale wie Viktimisierungserfahrungen oder Kriminalitätsfurcht.[56]

11 Nonresponse

Ein zentrales Problem der empirischen Sozialforschung liegt im Ausfall für die Befragung vorgesehener Personen. Diese Ausfälle werden als *Nonresponse* bezeichnet.[57] Je nachdem, ob es sich um einen totalen Ausfall der zur Befragung vorgesehenen Person handelt oder die befragte Person nur einige Fragen nicht beantwortet, wird nach Unit- und Item-Nonresponse unterschieden.

Um die statistischen Konsequenzen solcher Ausfälle abschätzen zu können, sind die möglichen Ausfallmechanismen von besonderer Bedeutung. Je nach den Eigenschaften des vermuteten Ausfallmechanismus ist mit verzerrten Schätzungen zu rechnen. Demnach ergeben sich die weiteren Schritte der Datenanalyse daraus, welcher Ausfallmechanismus angenommen wird. In der neueren Literatur werden drei Prozesse unterschieden:

– MCAR: missing completely at random,

– MAR: missing at random,

– MNAR: missing not at random.

---

56 Theoretisch begründet und mit Simulationsbeispielen belegt findet sich dieses Ergebnis erstmals bei Schnell 1993. Ein neuerer empirischer Hinweis unter Verwendung medizinischer Registerdaten mit fast 20.000 Fällen findet sich bei Vercambre/Gilbert 2012.

57 Weder Populationssurveys im Allgemeinen noch Nonresponse im Besonderen scheinen im Rahmen kriminologischer Forschung bisher die notwendige Aufmerksamkeit erfahren zu haben: Die 5.662 Seiten starke „Encyclopedia of Criminology and Criminal Justice" von Bruinsma/Weisburd (2014) enthält neben den Beiträgen von Johnson (2014) zu „Sample Selection Models" und Aebi/Linde (2014) zu „National Victimization Surveys" keine weiteren Beiträge zu diesen Themen.

Im einfachsten Fall liegt MCAR vor. Die Befragten fehlen damit völlig zufällig, das Fehlen ist also durch keine Variable vorhersagbar. Sowohl die Nonrespondenten als auch die Respondenten stellen somit eine Zufallsstichprobe aus allen zur Befragung vorgesehenen Personen dar.[58] Im Vergleich zu einer Stichprobe ohne Nonresponse werden die Schätzungen durch die ausfallbedingte Reduktion der Stichprobengröße lediglich etwas unpräziser ausfallen.[59] MCAR würde also dann vorliegen, wenn Personen rein zufällig ausfallen und dies nicht beispielsweise von Bildungsstand, Geschlecht, Alter oder beruflicher Stellung abhängt.

Falls hingegen der Ausfall durch Merkmale wie Bildungsstand, Geschlecht, Alter oder berufliche Stellung erklärt werden kann, dann liegt MAR vor. Die Schätzungen können dann immer noch unverzerrt durchgeführt werden, wenngleich spezielle Analysemethoden notwendig sind (für Unit-Nonresponse siehe beispielsweise Särndal/Lundström 2005; Bethlehem u. a. 2011 oder Valliant u. a. 2013; für Item-Nonresponse beispielsweise Schnell 1986; Schafer 1997; Little/Rubin 2002 oder Enders 2010).

Sollte aber der Ausfall direkt mit dem Thema der Befragung derart in Zusammenhang stehen, dass das Fehlen einer Beobachtung nur durch die fehlende Information selbst erklärt werden kann, dann liegt MNAR vor. Bei Viktimisierungssurveys wäre der Ausfall der Person z. B. vom Viktimisierungsstatus selbst abhängig, wobei dieser nicht durch andere Variablen vorhergesagt werden kann. In diesem Fall ist eine Korrektur nur über komplexe Modelle möglich (Sample-Selection-Modelle), für die die explizite Modellierung des Ausfallmechanismus erforderlich ist. Sollte die korrekte Modellierung des

---

58 Und damit ebenfalls eine Zufallsstichprobe aus der Grundgesamtheit, sofern es sich bei der Stichprobe um eine Zufallsstichprobe gehandelt hat.

59 Liegt als Ausfallmechanismus nicht MCAR vor, so kann der Ausfall von Untersuchungseinheiten nicht einfach ignoriert werden (siehe auch Unterkapitel 11.4). Aber auch das „Nachziehen“ in einer vor der Durchführung der Untersuchung nicht exakt definierten Stichprobe, um trotzdem eine gewünschte Stichprobengröße und damit Präzision zu erreichen, ist in der Realität nicht geeignet, dieses Problem zu lösen. Ausfälle durch nicht erreichte oder verweigernde Personen können nicht einfach durch leicht erreichbare oder kooperationsbereite Personen ersetzt werden, wenn sich diese Gruppen systematisch von den teilnehmenden Personen unterscheiden. Hieraus resultieren verzerrte Schätzungen (siehe Unterkapitel 11.3). Somit würde das Nonresponse-Problem nicht gelöst, sondern durch das Nachziehen lediglich verdeckt, wodurch das Ausmaß des Nonresponse nicht mehr angegeben werden kann. Die Stichprobe würde damit faktisch zu einer Quotenstichprobe und wäre damit als willkürliche Stichprobe nicht mehr verallgemeinerbar (Schnell u. a. 2013, 307).

unbekannten Ausfallmechanismus jedoch nicht möglich sein, ist auch mit diesen Modellen keine korrekte Analyse erreichbar.[60]

Insgesamt kann festgehalten werden, dass das Nonresponse-Problem praktisch nicht ignoriert werden kann, da auch das Nichtbeachten der Ausfälle faktisch unterstellt, das der unproblematische Ausfallmechanismus MCAR vorliegt. Dies ist in der Praxis kaum gegeben: Im Regelfall dürften die meisten Ausfallmechanismen in der empirischen Sozialforschung MAR sein und bedürfen entsprechender Berücksichtigung im Design der Studie und der Analyse. Sollten sich hingegen Hinweise auf MNAR ergeben, dann ist eine aufwendige statistische Modellbildung erforderlich. Dabei muss aber beachtet werden, dass Annahmen der resultierenden Modelle prinzipiell nicht mit den zur Verfügung stehenden Daten getestet werden können. Für die Forschungspraxis würden wir bei Hinweisen auf MNAR zu zusätzlichen empirischen Studien mit anderen Methoden raten (z. B. Nutzung administrativer Daten, verdeckte Beobachtung usw.), nicht hingegen zur statistischen Modellierung.

Natürlich ist das Ausmaß eines Nonresponse-Problems vom Mechanismus und der Größe des Nonresponse abhängig. Die Diskussion der Größe des Nonresponse ist einfacher als die Diskussion seiner Mechanismen, daher beschränken sich sogar die meisten methodischen Arbeiten allein auf die Größe des Nonresponse. Die Feststellung der Größe des Nonresponse erfolgt meistens mit der Berechnung einer Ausschöpfungsquote.

11.1 Ausschöpfungsquote

Das quantitative Ausmaß des Nonresponse wird meistens über die Ausschöpfungsquote als Gegenteil der Nonresponse-Quote angegeben.[61] Für die Berechnung der Ausschöpfungsquote liefert die *American Association for Pub-*

---

60 Sample-Selection-Modelle werden nach ihrem Erfinder auch als Heckman-Modelle bezeichnet. Sie bestehen aus einer Modell- und einer Selektionsgleichung. Die Modellgleichung lautet $y_{1i}^* = x_{1i}\beta_1 + u_{1i}$, wobei y_{1i}^* das latente Gegenstück zu der beobachtbaren Variable y_{1i} darstellt. Die Selektionsgleichung ist über $y_{2i}^* = x_{2i}\beta_1 + u_{2i}$ gegeben. Wenn $y_{2i}^* > 0$ gilt, ist y_{1i} mit $y_{1i} = y_{1i}^*$ beobachtet, im Fall $y_{2i}^* \leq 0$ liegt keine Beobachtung für y_{1i} vor, also $y_{1i} = 0$ (Puhani 2000, 54). Das Hauptproblem dieser Modelle liegt in der Tatsache, dass sich ihre statistischen Annahmen mit den gegebenen Daten prinzipiell nicht prüfen lassen (Schnell 2012, 178). Simulationsstudien zu diesem Problem lassen die routinemäßige Anwendung dieser Modelle höchst fragwürdig erscheinen (Stolzenberg 1997; Vella 1998).

61 Die Angaben der Nonresponse-Quote sind nicht immer einfach zu bewerten. So werden in Quotenstichproben „Ausfälle“ durch andere Personen mit passenden Quotenmerkmalen ersetzt. Dieses Vorgehen verdeckt das Nonresponse-Problem lediglich, ohne es zu lösen.

lic Opinion Research insgesamt sechs Definitionen (AAPOR 2011, 44–45), deren eindeutigste als *minimum response rate* (RR1) bekannt ist. Sie ist über

$$RR1 = \frac{I}{(I+P)+(R+NC+O)+(UH+UO)} \tag{13}$$

gegeben (AAPOR 2011, 44). Die einzelnen Größen bezeichnen vollständige Interviews (*I*), partielle Interviews (*P*), Verweigerungen und Abbrüche (*R*), nicht Erreichte (*NC*), andere Ausfallgründe (*O*), unbekannt, ob es sich um einen Haushalt handelt oder nicht (*UH*) und andere unbekannte Ursachen (*UO*).

Um den Unterschied zwischen *O* einerseits sowie *UH* und *UO* andererseits zu verdeutlichen, soll hierauf weiter eingegangen werden. *O* liegt vor, wenn eine Zielperson zur Grundgesamtheit gehört, nicht verweigert, es aber trotzdem durch Krankheit oder Sprachprobleme nicht zu einem Interview kommt. Im Gegensatz ist für *UH* bzw. *UO* beispielsweise durch nicht aufgefundene oder nicht bearbeitete Adressen unbekannt, ob in dem Haus überhaupt eine Zielperson existiert (Schnell 2012, 163).

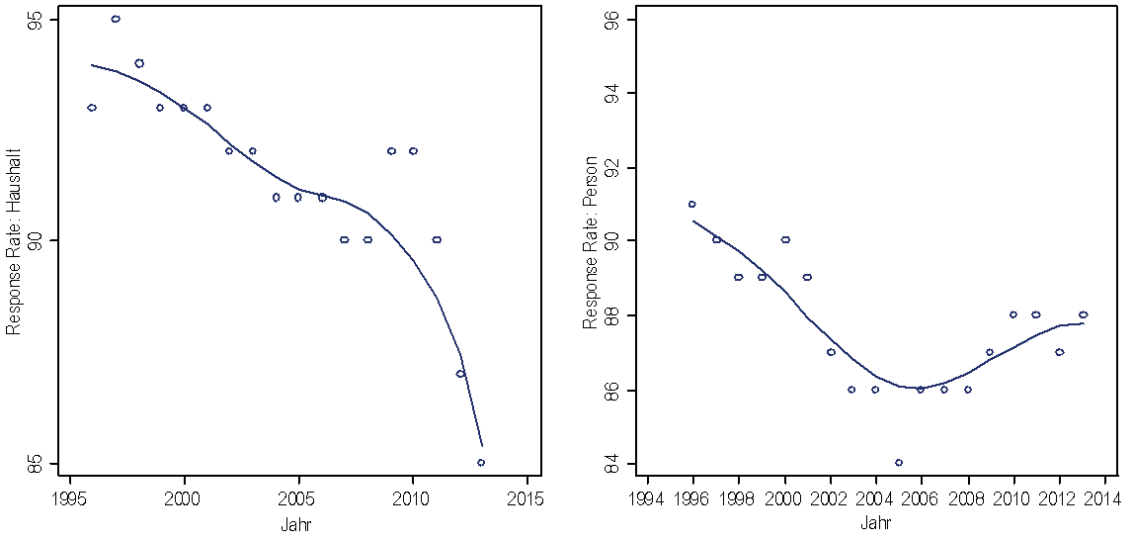
Trotz der starken Bemühungen sind die Ausschöpfungsquoten auch in den aufwendigsten Viktimisierungsstudien in den letzten Jahren zurückgegangen (*Abbildung 9*).[62] Der Effekt der verstärkten Bemühungen gegeben einen Kontakt ist in der rechten Abbildung deutlich zu sehen. Man beachte, dass trotz des Rückgangs beider Ausschöpfungsquoten die Höhe deutlich über den entsprechenden Zahlen für Deutschland liegt. Dies mag zum Teil sicherlich an den erheblich höheren Kosten pro Fall liegen, die für den amerikanischen *National Crime Victim Survey* (NCVS) akzeptiert werden.

---

62 Die der Abbildung zugrunde liegenden Daten wurden den Methodenberichten der jeweiligen Erhebung („National Crime Victimization Survey Technical Documentation“) des NCVS auf der Homepage des amerikanischen Justizministeriums (www.bjs.gov/content/pub/pdf/ncvstd13.pdf) entnommen.

Abbildung 9:

Ausschöpfungsrate (*Response Rate*) für Haushalte und Personen im NCVS 1996–2013. Eingezeichnete Linien: Lowess, Glättungsparameter f = 0,8



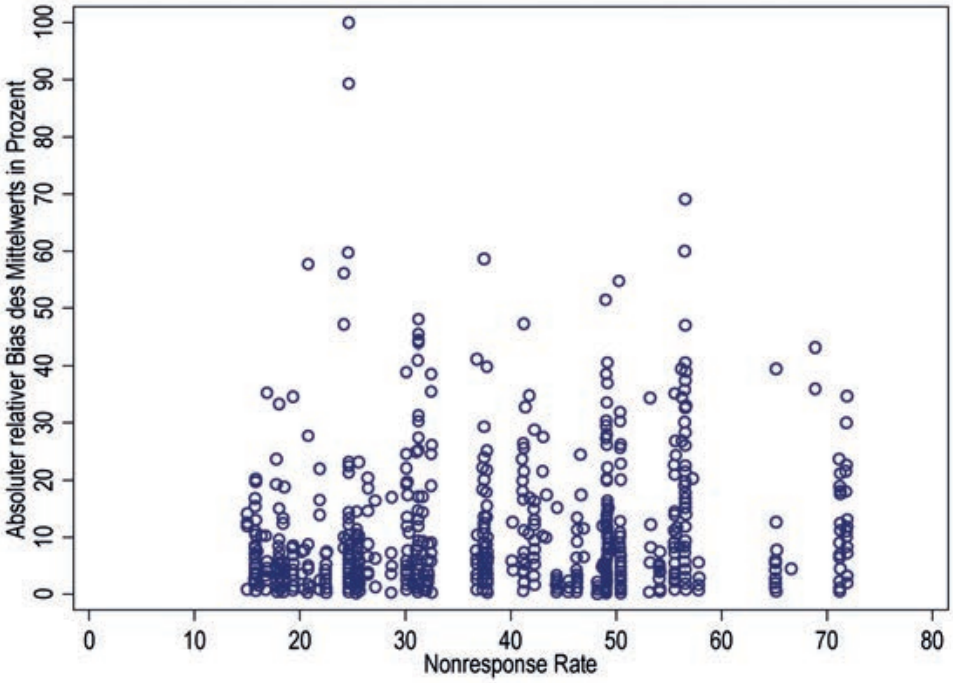
Die bloße Angabe der Ausschöpfungsquote ist für eine angemessene Abschätzung des Nonresponse-Problems allerdings nicht ausreichend. So konnten Groves und Peytcheva (2008) in der von ihnen durchgeführten Meta-Analyse anhand von 959 Nonresponse-Bias-Schätzungen aus 59 Studien zeigen, dass die Nonresponse-Rate nahezu keine Erklärungskraft für den Nonresponse-Bias besitzt (*Abbildung 10*).[63] Über eine einfache lineare Regression werden lediglich 4 % der Varianz des Nonresponse-Bias durch die Nonresponse-Rate erklärt (R^2 = 0,04; Groves und Peytcheva 2008, 174).[64] Für das Auftreten von Nonresponse-Bias müssen also weitere Größen von Bedeutung sein.

63 Zu diesem Ergebnis kommen auch Klausch u. a. 2015.

64 Die Abbildung entspricht der zweiten Abbildung bei Groves/Peytcheva (2008, 172). In Ermangelung des Datensatzes wurden die Daten aus dieser Abbildung für Abbildung 10 rekonstruiert. Die Grundlage der Abbildung bilden 959 Datenpunkte. Da durch Overplotting aber nur 527 Datenpunkte erkennbar sind, stimmt zwar die Darstellung optisch mit dem Original überein, Berechnungen mit der rekonstruierten Datenmatrix wären aber irreführend.

Abbildung 10:

Absoluter relativer Nonresponse-Bias für 959 Schätzungen in Abhängigkeit der Nonresponse-Rate in insgesamt 59 Surveys (Datenquelle: Groves/Peytcheva 2008, 172)



11.2 Nonresponse-Bias

Um zu verstehen, warum Nonresponse zu verzerrten Schätzungen führen kann, hilft die Betrachtung der Formel des Nonresponse-Bias. Dieser Bias ist in seiner einfachsten Form über

$$\bar{y}_{res} - \bar{y}_{all} = \frac{n_{non}}{n}(\bar{y}_{res} - \bar{y}_{non}) \qquad (14)$$

gegeben mit den entsprechenden Werten für alle Befragten (*all*), Respondenten (*res*) und Nonrespondenten (*non*) (z. B. Groves 1989, 134). An dieser Formel kann abgelesen werden, dass das primäre Problem – wie bereits gesehen – nicht im Ausmaß des Nonresponse, also dem Anteil von Nonrespondenten an allen zur Befragung vorgesehenen Personen $\frac{n_{non}}{n}$, sondern vielmehr in der Differenz zwischen Respondenten und Nonrespondenten $\bar{y}_{res} - \bar{y}_{non}$ liegt. Ist diese Differenz klein, unterscheiden sich Respondenten und Nonrespondenten also nicht (oder kaum), so kann auch bei einem großen Anteil an Nonrespondenten trotzdem von unverzerrten Schätzungen ausgegangen werden. Unterscheiden sich Respondenten und Nonrespondenten

allerdings systematisch, so sind die Schätzungen auch bei einem kleinen Anteil an Nonrespondenten verzerrt.[65] Dies ist besonders dann der Fall, wenn ein Zusammenhang zwischen Ausfallursache und dem Thema der Befragung besteht.

Im Hinblick auf die Ausfallursache ist an dieser Stelle darauf hinzuweisen, dass es sich bei Nonrespondenten um keine homogene Population handelt, auch wenn dies implizit in Formel 14 unterstellt wird. Für einen angemessenen Umgang mit Nonresponse ist somit eine weitere Unterteilung des Nonresponse in verschiedene Ausfallursachen notwendig.

11.3 Ausfallursachen

In der Literatur werden im Allgemeinen mindestens drei Kategorien von Ursachen für den Ausfall einer zur Befragung vorgesehenen Person genannt:

1. Verweigerung

2. Teilnahmeunfähigkeit

3. Nichterreichbarkeit

Interessanterweise ist es möglich, dass sich die Effekte des Nonresponse in den jeweiligen Gruppen unterscheiden, also verschiedene Gruppen von Nonrespondenten im Vergleich zu den Respondenten systematisch niedrigere

---

65 Eine Möglichkeit zur Schätzung des maximalen Bias beruht auf der Annahme, dass sich die Wahrscheinlichkeit für eine Teilnahme einer Person als *Responsepropensity* ρ schätzen lässt. Der maximal mögliche Bias ist dann durch

$$B_{max}(y,\rho) = \frac{\big(1 - R(\rho)\big)S(y)}{2\bar{\rho}} \geq \left|\frac{\mathrm{Cov}(y,\rho)}{\bar{\rho}}\right|$$

mit

$$R(\rho) = 1 - 2S(\rho)$$

gegeben, wobei $S(\rho)$ die Standardabweichung der Responsepropensities, $S(y)$ die Populationsvarianz der abhängigen Variablen, $\bar{\rho}$ den Mittelwert der Responsepropensities und $\mathrm{Cov}(y, \rho)$ die Kovarianz zwischen den Responsepropensities und der abhängigen Variablen darstellt. Diese Wahrscheinlichkeit ρ, dass eine für die Stichprobe ausgewählte Person auch antwortet, wird dabei durch eine Reihe von Hilfsvariablen x_j, beispielsweise über eine logistische Regression, geschätzt (Schouten u. a. 2009, 105). Der Bias ist dabei umso größer, je stärker die Korrelation der Responsepropensities ρ mit der untersuchten Variablen y ausfällt (Schouten u. a. 2009, 107). Die zentrale Schwäche dieses sogenannten R-Indikatoren-Ansatzes besteht in der Auswahl der Hilfsvariablen x_j. Wenn der Nonresponse-Mechanismus nicht mit den zur Schätzung der Responseprobabilities verwendeten Hilfsvariablen korreliert, bleibt der Bias unbemerkt (Schnell 2012, 174). Somit würde eine unverzerrte Schätzung angenommen, obwohl dies nicht der Fall ist.

oder höhere Mittelwerte aufweisen oder auch gar keine Differenz zu beobachten ist. Daher ist es prinzipiell möglich (wenn auch in der Praxis eher unwahrscheinlich), dass eine Erhöhung der gesamten Ausschöpfung zu einer Vergrößerung der Verzerrung führen kann, und zwar wenn Subgruppen mit kleiner Differenz zu den Respondenten stärker ausgeschöpft werden als Subgruppen mit großer Differenz. Eine Erweiterung von Formel (14) soll dies verdeutlichen. Diese erweiterte Formel ist über

$$\bar{y}_{res} - \bar{y}_{all} = \frac{n_{nc}}{n}(\bar{y}_{res} - \bar{y}_{nc}) + \frac{n_{rf}}{n}(\bar{y}_{res} - \bar{y}_{rf}) + \frac{n_{na}}{n}(\bar{y}_{res} - \bar{y}_{na}) + \frac{n_{ot}}{n}(\bar{y}_{res} - \bar{y}_{ot}) \quad (15)$$

gegeben, wobei *nc* für nicht erreichte Personen (*noncontacts*), *rf* für Verweigerer (*refusals*), *na* für teilnahmeunfähige Personen (*not able*) und *ot* für andere Gründe (*other*) steht (siehe Schnell 2012, 171). Im Folgenden soll auf die einzelnen Ausfallursachen näher eingegangen werden.

11.3.1 Verweigerung

In der Literatur werden viele Gründe als Ursachen für die Verweigerung der Teilnahme an einem Survey diskutiert, beispielsweise die Belastung durch Länge oder Häufigkeit der Befragung, politisches Desinteresse, altersbedingter Rückzug aus öffentlichen Angelegenheiten, Kriminalitätsfurcht oder unklare Konsequenzenbefürchtungen (Schnell 2012, 159). All diesen Hypothesen ist gemein, dass sie sich als Spezialfälle der Rational-Choice-Theorie interpretieren lassen (Schnell 1997, 157–216). Demnach ist eine Teilnahme dann zu erwarten, wenn der erwartete Nutzen die erwarteten Kosten übersteigt. Die Kosten und der Nutzen werden von den einzelnen Befragten individuell bewertet, sodass deren Einschätzung von den Besonderheiten und jeweiligen situativen Bedürfnissen und Erwartungen der Befragten abhängig ist.

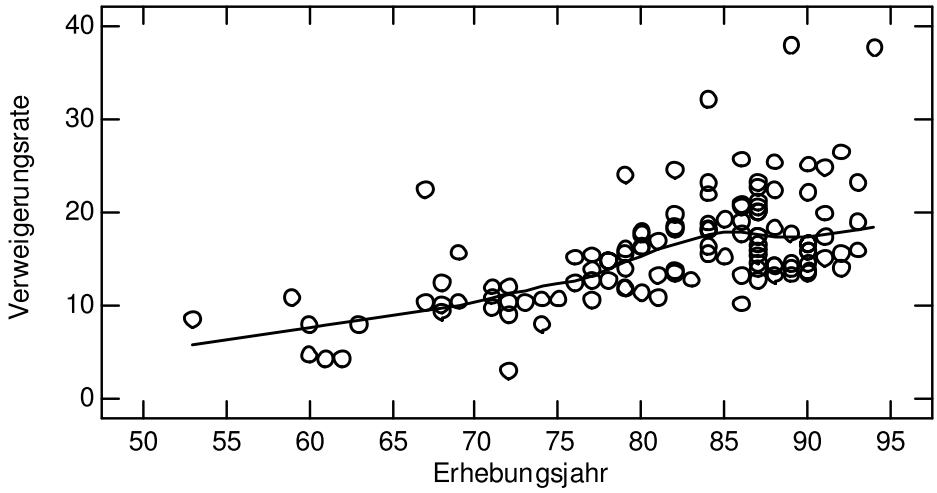
Durch die starke Routinisierung großer Teile des alltäglichen Lebens und Handelns reichen oftmals schon kleine Reize aus, um entsprechende Handlungsskripte bei den Befragten auszulösen (Schnell 2012, 159). Somit wird erklärbar, warum gelegentlich schon minimale Veränderungen in der Erhebungssituation (kleine Geschenke, Sprache oder Aussehen der Interviewer) zu größeren Veränderungen im Anteil der Verweigerungen führen können. Die Entscheidung zur Verweigerung scheint also stark von der Situation abzuhängen und ist somit nicht als über die Situation hinaus stabile Eigenschaft der Befragten zu sehen. Sowohl die hohen Konvertierungsraten von Verweigerern (bis zu maximal 30 %) als auch der Zeitpunkt der Verweigerung (gewöhnlich in den ersten Sekunden des Interviews, noch bevor das Thema der

Befragung erläutert wurde) können als Hinweis darauf interpretiert werden.[66] Dazu passend zeigen sich in der Regel nur schwache Korrelationen zwischen Verweigerungsverhalten und Hintergrundvariablen. Entsprechend existieren keinerlei empirische Hinweise auf einen „harten Kern“ von Verweigerern (Schnell 1997, 151, 186; Schnell 2012, 159–160).

Trotzdem muss bei jeder Studie geklärt werden, ob Verweigerungsgründe systematisch mit dem Thema der Untersuchung in Zusammenhang stehen. In einem solchen Fall sind verzerrte Schätzungen zu erwarten. So könnten beispielsweise ältere Personen Befragungen eher verweigern, weil sie die Kontaktaufnahme als Teil eines potenziellen Trickbetrugs wahrnehmen. Von weiblichen Personen könnte die Kontaktaufnahme in persönlichen Befragungen als Versuch wahrgenommen werden, Zutritt zur Wohnung in Absicht eines sexuellen Übergriffs zu erlangen. In diesen Fällen würde der Ausfallgrund mit den inhaltlichen Merkmalen einer Viktimisierungsstudie zusammenhängen, sodass beispielsweise das Kriminalitätsfurchtniveau älterer Personen vor Trickbetrug oder die Angst weiblicher Personen vor sexuellen Übergriffen unterschätzt werden würden.

Abbildung 11:

Entwicklung der Verweigerungsraten in akademischen Surveys in der BRD 1953–1994 (Schnell u. a. 2013, 301)



66 Die Höhe der Konvertierungsraten hängt maßgeblich von den Merkmalen und dem Verhalten der Interviewer ab. Ein Beispiel für ein vollständiges Verweigerungs-Reduktions-Training in deutscher Sprache findet sich bei Schnell (2012, 223–225).

Wie *Abbildung 11* zu entnehmen sind die Verweigerungsraten spätestens seit den 70er Jahren deutlich gestiegen. Man muss bedenken, dass dieser Trend trotz der gegenteiligen Bemühungen der Institute zu beobachten ist. Generell dürfte diese ansteigende Verweigerungstendenz unumkehrbar sein. Allerdings zeigen die großen Streuungen der Verweigerungsraten zwischen verschiedenen Studien gleichsam, dass ein großer Teil des Verweigerungsverhaltens offensichtlich von den Details der Feldarbeit abhängt. Es macht für eine endgültige Verweigerung einen deutlichen Unterschied, ob alle Regeln für ein erfolgreiches Interview (Ankündigung, Belohnung, Interviewerwechsel, Konvertierungsversuche, Wechsel des Erhebungsmodus etc., zu den Details Schnell 2012, 181–183, 223–225) beachtet wurden oder nicht. Entscheidend ist dabei, dass nicht eine einzelne Maßnahme zu einer deutlichen Verbesserung der Ausschöpfung führt, sondern nur die konsequente Anwendung aller Maßnahmen. Entsprechend kostenintensiv sind korrekt durchgeführte Erhebungen.

11.3.2 Erkrankung/Teilnahmeunfähigkeit

Es existieren verschiedene Gründe, warum eine Person nicht an einer Befragung teilnehmen kann, beispielsweise nicht ausreichende Sprachkenntnisse, Analphabetismus in einer postalischen Befragung, psychische Probleme, chronischer Alkohol- und Drogenmissbrauch oder schwere Erkrankungen (z. B. Demenz). Sollte der Grund der Nichtbefragbarkeit mit dem Thema des Surveys in Zusammenhang stehen, so ist mit verzerrten Schätzungen zu rechnen (Schnell u. a. 2013, 302).

Dies wäre z. B. dann der Fall, wenn eine zur Befragung vorgesehene Person aufgrund einer Viktimisierung entweder physisch oder psychisch nicht in der Lage ist, an der Befragung teilzunehmen. Ebenso würden die Viktimisierungsraten unterschätzt, wenn ein systematischer Zusammenhang zwischen einer vorgefallenen Viktimisierung und Deutschkenntnissen bestünde.

11.3.3 Nichterreichbarkeit

Für die meisten Studien stellen weder Verweigerungen noch Befragungsunfähigkeit prinzipiell das größte Problem dar, sondern schwer- und nicht erreichbare Personen. Damit sind Personen gemeint, die trotz mehrfacher Kontaktversuche an ihrem Wohnsitz nicht angetroffen werden. Neben Personen mit besonders vielen Sekundärkontakten (z. B. politisch Aktive, Vereinsmitglieder usw.) trifft dies auch auf längere Zeit Verreiste, Personen, deren tatsächlicher Aufenthaltsort nicht mit ihrem Wohnsitz übereinstimmt (z. B. Montagearbeiter), und Personen mit ungewöhnlichen Arbeitszeiten (z. B.

Krankenpflegepersonal) zu (Schnell u. a. 2013, 303). Diese Ausfälle erfolgen nicht zufällig, sondern hängen offensichtlich mit bestimmten Merkmalen der Personen zusammen. Übereinstimmend damit zeigt sich, dass die Erreichbarkeit der zur Befragung vorgesehenen Personen mit zahlreichen sozialwissenschaftlich relevanten Variablen korreliert. Aus diesem Grund können schwierig erreichbare Personen nicht einfach durch leicht erreichbare Personen ersetzt werden (Schnell 1998). Sind Personen schwieriger oder nicht erreichbar, weil sie vielen Aktivitäten außerhalb der eigenen Wohnung nachgehen, und weisen diese Personen eine erhöhte Wahrscheinlichkeit auf, Opfer eines Verbrechens zu werden, so wird die entsprechende Viktimisierungsrate unterschätzt. Dieser Zusammenhang wird zum Beispiel von Hindelang u. a. (1978, 250) oder Cohen/Felson (1979, 589) beschrieben.

Daher müssen auch schwierig Erreichbare in die Stichprobe aufgenommen werden. Zumeist wird über mehrere Kontaktversuche (*Callbacks*) zu verschiedenen Tageszeiten versucht, die zur Befragung vorgesehene Person zu erreichen. Erfolgsversprechende Zeiten für Kontaktversuche liegen in den frühen Abendstunden oder am Wochenende, wobei bei persönlichen Befragungen auch ein Wechsel der Interviewer wünschenswert ist, da sich Interviewer in ihren Kontaktstrategien unterscheiden. Interessanterweise lässt sich auch die Nichterreichbarkeit durch Incentives beeinflussen. Insgesamt zeigt sich, dass sich die Zahl der Nichterreichten durch eine flexible Kontaktstrategie und eine hohe Zahl von Callbacks deutlich reduzieren lässt. Allerdings steigen mit zunehmender Callback-Zahl auch die Kosten pro Interview (Schnell u. a. 2013, 303).

11.4 Vermeidung und Kontrolle von Nonresponse statt Korrektur

Wie bereits ausgeführt ist die Responserate allein nicht geeignet, um Aussagen über mögliche auf Nonresponse zurückgehende Verzerrungen zu treffen. Man benötigt für jede neue Studie erneut eine Analyse der Ursachen des Nonresponse.[67] In der Regel bedeutet dies, für jede Gruppe der Ausfallmechanismen (Erkrankung, Verweigerung, Nichterreichbarkeit) die möglichen Effekte auf die jeweilige Studie zu analysieren. Dies muss bereits vor der ersten

[67] Hierfür sind Informationen über die Erhebung, sogenannte Paradaten (z. B. Datum, Uhrzeit, Nummer und Ergebnis des Kontaktversuchs) notwendig. Diese Daten sind für die Erhebungsinstitute prinzipiell leicht verfügbar und nahezu kostenneutral zu erhalten, trotzdem sind nicht alle Erhebungsinstitute bereit, diese Paradaten auch zur Verfügung zu stellen. Demnach sollte im Vorfeld durch den Auftraggeber vertraglich festgehalten werden, welche (Para-)Daten als Teil des Datensatzes vom Erhebungsinstitut zu liefern sind (siehe hierzu Anhang F in Schnell u. a. 2013).

Feldphase erfolgen. Im Anschluss an diese Analyse muss dann das Design der Studie (z. B. Erhebungsmodus, Klumpung und Schichtung, Interviewerkontrolle, Interviewerallokation, Incentives, Tracking-Maßnahmen, Interviewerschulung, Verweigerungstraining etc.) angepasst werden.

An dieser Stelle sei explizit darauf hingewiesen, dass die einzig erfolgsversprechende Strategie im Umgang mit dem Nonresponse-Problem darin besteht, bereits in der Feldphase alle möglichen Schritte zu unternehmen, um Unit-Nonresponse so weit wie möglich zu verhindern. Die beste Lösung des Nonresponse-Problems besteht darin, kein Nonresponse-Problem zu haben.[68] Methodologen sind sich einig, dass Nonresponse ein bereits vor einer Erhebung zu berücksichtigendes und minimierendes Problem ist. In der Praxis zeigen sich hingegen naive Anwender durch ein hohes Ausmaß an Nonresponse überrascht und betrachten es fälschlich als unabwendbares Naturereignis. Weder Verweise auf zahlreiche andere Studien mit Nonresponse-Problemen noch heroische Annahmen über die vermeintliche Neutralität der Ausfälle sind akzeptable Handlungsstrategien. Dies gilt auch für alle Versuche, Nonresponse nachträglich allein durch spezielle Gewichtungsverfahren zu „korrigieren“ (siehe Unterkapitel 11.6).

11.5 Nonresponse in neueren deutschen Viktimisierungssurveys

Da für die Bundesrepublik im Gegensatz zu den Vereinigten Staaten und zum Vereinigten Königreich kein amtlicher Viktimisierungssurvey wie der NCVS bzw. der *Britisch Crime Survey* (BCS) existiert, ist die kriminologische Forschung in Deutschland auf eigene Erhebungen angewiesen. Interessanterweise existiert aber bislang keine Übersicht über Nonresponse in neueren Erhebungen, in denen kriminologisch relevante Fragen Teil des Frageprogramms waren. Diese Lücke wurde durch die von den Autoren betreute Abschlussarbeit von Klingwort (2014) geschlossen. Gesucht wurden Surveys anhand folgender Kriterien:

– Feldzeit ab dem 01. 01. 2001 und

– Stichprobengröße ≥ 1.500.

---

[68] Anderson u. a. (1983) schreiben in Hinsicht auf fehlende Werte allgemein dieses Motto ohne Quellenangabe bereits Snedecor zu.

Weiterhin sollten Fragen nach

– Viktimisierung in Form von körperlicher Gewalt und/oder

– Viktimisierung in Form von Wohnungseinbruch und/oder

– zur Kriminalitätsfurcht

erhoben worden sein. In die resultierende Auflistung wurden sowohl spezielle Viktimisierungssurveys als auch Mehrthemenbefragungen mit einem kriminologischen Teil aufgenommen.

Besonderes Augenmerk lag dabei auf der Definition der Grundgesamtheit als „bundesweite allgemeine Bevölkerung". Trotzdem wurden auch die Erhebungen in die Analyse einbezogen, die zwar den ersten Kriterienkatalog erfüllten, sich aber nur auf einzelne Regionen oder spezielle Subpopulationen bezogen.[69]

Von den 34 identifizierten Projekten mit insgesamt 105 Erhebungen mit kriminologischem Bezug seit 2001, die alle Kriterien erfüllten, beziehen sich lediglich 13 Projekte auf die allgemeine Bevölkerung (Klingwort 2014, 10–32).[70]

---

69 Hierbei handelt es sich um die Projekte „Lebenssituation, Sicherheit und Gesundheit von Frauen in Deutschland", „Kriminalität und Gewalt im Leben alter Menschen", „Jugendliche in Deutschland als Opfer und Täter von Gewalt", „Lebenssituation und Belastung von Frauen mit Behinderungen und Beeinträchtigungen in Deutschland", „Gender-based Violence, Stalking and Fear of Crime", „Repräsentativbefragung zu Viktimisierungserfahrungen in Deutschland", „Violence against Women: an EU-wide Survey", „Muslime in Deutschland" (Erwachsene, Studenten, Schüler), „Jugendgewalt und Jugenddelinquenz in Hannover", „Second International Self-Reported Study of Delinquency", „Sicherheit und Kriminalität in Stade" (Schüler, Erwachsene), „European Union Minorities and Discrimination Survey", „Jugendliche als Opfer und Täter von Gewalt im Bundesland Sachsen-Anhalt", „Kriminalitäts- und Terrorismusfurcht in Hessen", „Jugendliche als Opfer und Täter von Gewalt in Wolfsburg", „Kinder- und Jugenddelinquenz im Bundesland Saarland", „Jugendliche als Opfer und Täter von Gewalt in Berlin", „Jugendliche als Opfer und Täter von Gewalt im Landkreis Emsland", „Gewalt im Strafvollzug", „Youth Deviance and Youth Violence" und „Jugendkriminalität in der modernen Stadt" (2001–2009, 2011, 2013) (Klingwort 2014, 33–55).

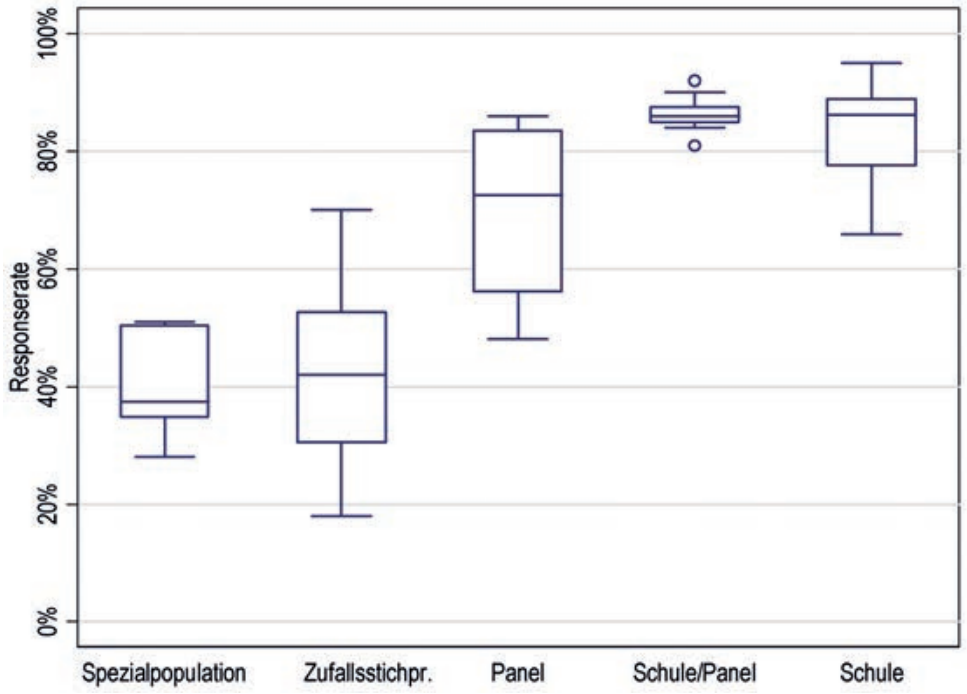
70 Hierbei handelt es sich um folgende Projekte: „Die Ängste der Deutschen" (jährlich 2001–2014), „Kriminalitätsfurcht, Strafbedürfnisse und wahrgenommene Kriminalitätsentwicklung" (2004, 22006, 2010), „European Survey of Crime and Safety 2005 (EU ICS)", „Studie zur Gesundheit Erwachsener in Deutschland", „Kriminalität und Sicherheitsempfinden", „International Crime Victims Survey-Pilotstudie 2010" (CATI und Online-Panel), „Deutscher Viktimisierungssurvey 2012 (Barometer Sicherheit in Deutschland)", „Sicherheitsreports" (2011, 2012, 2013), „Das Sozio-oekonomische Panel (SOEP, jährlich 2001–2012)", „Eurobarometer" (Standard Eurobarometer zwei Mal jährlich 2001–2013, Nr. 56–59, 61–79), „Special Eurobarometer 2010", „European Social Survey" (2002, 2004, 2006, 2008, 2010, 2012), „Allgemeine Bevölkerungsumfrage der Sozialwissenschaften 2008" und „Allgemeine Bürgerbefragungen der Polizei in Nordrhein-Westfalen" (Klingwort 2014, 10–32). Die letztgenannte Studie würde die Aufnahmekriterien eigentlich nicht erfüllen, da sie auf NRW beschränkt ist. Da NRW aber fast 22 % der Bevölkerung der Bundesrepublik umfasst, wird diese Studie trotzdem aufgeführt.

In diesen Projekten wurden insgesamt 71 Erhebungen durchgeführt. Davon sind 17 Erhebungen Quotenstichproben. Da sich für Quotenstichproben keine Nonresponse-Quoten angeben lassen (obwohl Quotenstichproben auch ein Nonresponse-Problem besitzen), sind diese Studien für Nonresponse-Analysen irrelevant und wurden aus der Betrachtung ausgeschlossen. Die folgenden Analysen basieren auf 18 Erhebungen aus dieser Gruppe.[71]

Die verbleibenden 21 Projekte bestehen aus 34 Erhebungen, wobei für drei dieser Erhebungen keine Responseraten angegeben werden können.[72] Somit basiert *Abbildung 12* auf 18 + 31 = 49 Erhebungen.

Abbildung 12:

Responseraten nach Auswahlverfahren und Art der Zielpopulation (Datengrundlage: Klingwort 2014)



71 Für die Eurobarometererhebungen (24 Erhebungen) sind in den Dokumentationen keine Informationen über Nonresponse enthalten. Für das SOEP (12 Wellen) wurden keine Responseraten recherchiert, da dies hier für ein seit vielen Jahren laufendes Panel nicht sinnvoll erscheint. Informationen zu den Ausschöpfungsquoten der einzelnen SOEP-Wellen können Kroh 2014 entnommen werden.

72 Dies sind die Erhebungen „Gender-based Violence, Stalking and Fear of Crime“ (nicht dokumentiert), „Repräsentativbefragung zu Viktimisierungserfahrungen in Deutschland“ (Quotenstichprobe) und „Jugendkriminalität in der modernen Stadt 2013“ (Information liegt nicht vor, aktuellster Methodenbericht von Bentrup und Verneuer 2014 bezieht sich auf 2011).

Die Responseraten dieser Erhebungen sind nach Erhebungstyp zusammengefasst. Hier zeigt sich für die Random-Erhebungen eine Responserate von durchschnittlich ca. 41 %, für die Erhebungen von Spezialpopulationen von ungefähr ca. 40 %. *Abbildung 12* enthält eine Reihe interessanter Details.

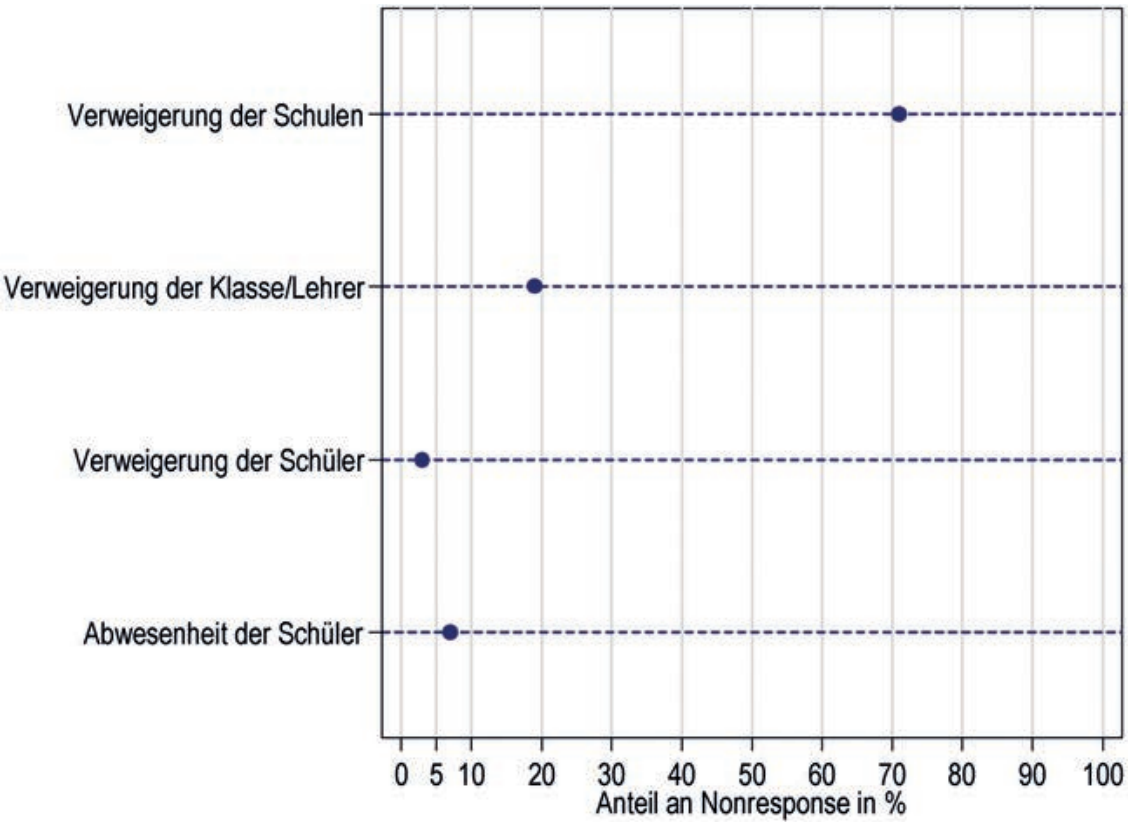
Erstens kann eine große Streuung innerhalb einiger der Gruppen beobachtet werden. So liegen beispielsweise die Responseraten für Erhebungen auf der Basis von Zufallsstichproben zwischen 20 % und 70 %. Schwankungen dieser Größe innerhalb von wenigen Jahren sind nicht zufällig. Die oben beschriebenen kumulativen Effekte scheinbar trivialer Details wie Ankündigung, Feldzeit, Incentives, Konvertierungsversuche etc. resultieren in eindrucksvollen Differenzen von mehr als 50 % Unterschied in der Ausschöpfung einer Zufallsstichprobe aus der bundesweiten „allgemeinen" Bevölkerung. Dieser große Unterschied in der Ausschöpfungsquote ist das wichtigste Ergebnis der Arbeit von Klingwort (2014).

Nicht so bedeutsam für die Methodenforschung im Allgemeinen ist das erwartbare zweite Ergebnis der Studie: die Beobachtung hoher Responseraten für Panelstudien und Schulbefragungen sowie besonders für Panelstudien mit Schulen. Allgemein sinkt die Kooperationsbereitschaft der Befragten in Panelstudien – falls die erste Befragung nicht traumatisch war – mit der Anzahl der teilgenommenen Wellen nur langsam. Der Ausfall aus einem Panel (*Dropout*, *Attrition*) lässt sich durch Belohnungen für die Teilnehmer (Untersuchungsberichte sind keine geeigneten Belohnungen) und hohen Aufwand für das Verfolgen kooperationsbereiter, aber verzogener Personen („Tracking") weiter reduzieren.

Schülerbefragungen sind aus verschiedenen Gründen als problematisch anzusehen. Zwar ist bei einer Befragung der Schülerschaft einer Schule innerhalb der Klassen eine hohe Kooperationsbereitschaft erwartbar, doch setzt dies voraus, dass die jeweilige Schule kooperiert. Dies ist keineswegs selbstverständlich. Der Großteil des Nonresponse in der Studie „Second International Self-Reported Study of Delinquency" (2006) ging beispielsweise auf die komplette Verweigerung der Teilnahme einiger Schulen zurück (Enzmann 2010, 51, siehe *Abbildung 13*). Entsprechend sollte bei der Beurteilung von Schülerbefragungen beachtet werden, ob ein solcher Komplettausfall einer Schule (korrekt) als Nonresponse codiert oder als vermeintlich unsystematischer Ausfall betrachtet wurde.

Abbildung 13:

Nonresponse nach Ursache in der Schülerbefragung „Second International Self-Reported Study of Delinquency“ von 2006 (Enzmann 2010, 51)



Weiterhin ist in Deutschland eine Einwilligung der Eltern zur Befragung Minderjähriger notwendig (Schnell 2012, 166). Auch diese Einwilligung kann durchaus systematisch mit kriminologischen Variablen zusammenhängen.

Das schwerwiegendere Problem bei Schülerbefragungen besteht aber darin, dass Absentismus der Schulpflichtigen an die Delinquenz eben dieser gekoppelt zu sein scheint (Vaughn u. a. 2013, 773). Wenn also gerade delinquente Jugendliche nicht erscheinen oder in einer Panelstudie die Schule eher abbrechen als andere Schülerinnen und Schüler, so stellen die verbleibenden Panelteilnehmer keine zufällige Auswahl aus der gesamten Schülerschaft dar.[73]

[73] Ein deutsches Beispiel für diesen Effekt findet sich in der Duisburger kriminologischen Schuluntersuchung bei Pöge 2007.

Daher ist auch die Dauer der Feldzeit innerhalb der Schulen für Schülerbefragungen von Bedeutung. Wird die Befragung nur an einem Stichtag durchgeführt und nicht über einen längeren Zeitraum ist mit deutlich niedrigeren Responseraten und eher mit systematischen Ausfällen zu rechnen. Bei Studien mit solchen durch Absentismus systematisch erfolgenden Ausfällen ist die geschätzte Delinquenz demnach lediglich als Untergrenze der tatsächlichen Delinquenz anzusehen. Da auch bei Jugendlichen aufgrund des sogenannten *Victim-Offender Overlap* Täter, insbesondere bei Gewaltdelikten überproportional häufig Opfer dieser Delikte werden (Shaffer/Ruback 2002), wird demnach durch Absentismus nicht nur die Anzahl der Täter, sondern auch die die Anzahl der Opfer unterschätzt.

Abschließend möchten wir nochmals betonen, dass die Ausschöpfungsquote allein keinen Hinweis auf mögliche Nonresponse-Effekte liefert. Allerdings sollten mittlere Ausschöpfungen von 40 % ein deutlicher Hinweis darauf sein, das Nonresponse bereits bei der Planung einer Erhebung berücksichtigt werden muss – und nicht erst bei der Analyse.

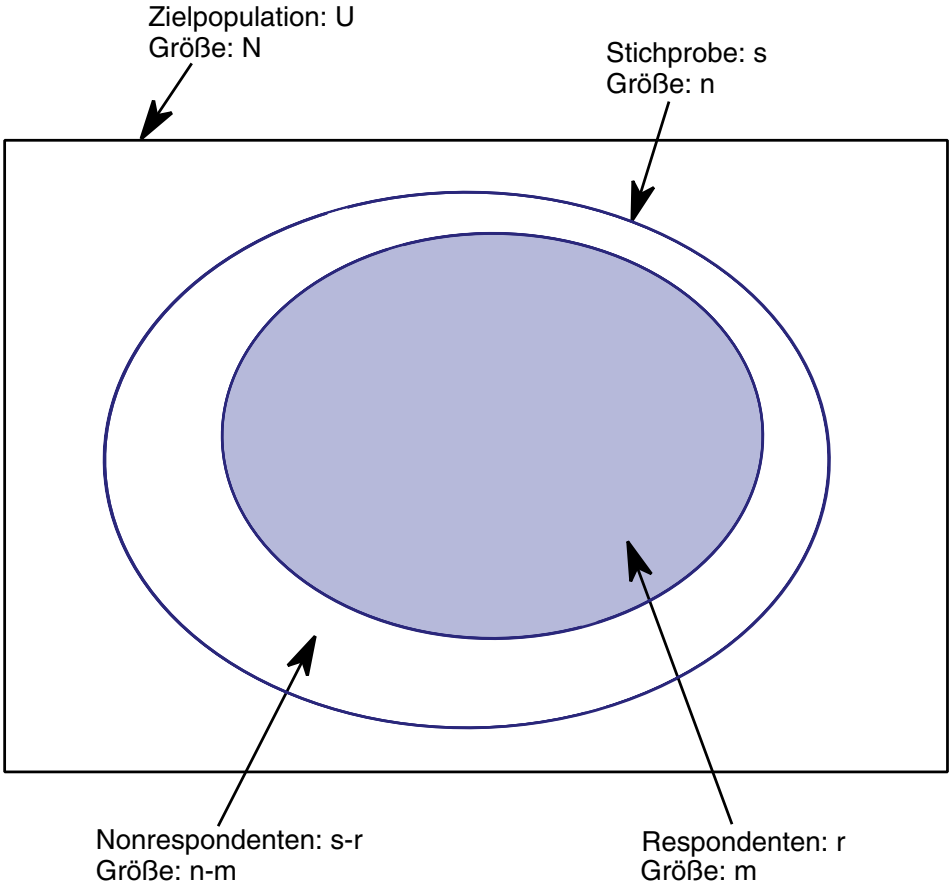
11.6 Korrekturverfahren für Nonresponse-Bias

Trotz aller Anstrengungen während der Feldphase ist Nonresponse unvermeidlich. Üblicherweise wird versucht, zumindest offensichtliche Verzerrungen der Stichprobe nachträglich durch Gewichtungsverfahren zu kompensieren.[74] Dies bedeutet, dass jedem Fall im Datensatz ein bei Analysen zu berücksichtigendes Gewicht zugeordnet wird. *Abbildung 14* illustriert die im Folgenden verwendete Notation.

[74] In der Statistik wird zwischen Design- und Korrekturgewichten unterschieden. Designgewichte korrigieren für unterschiedliche Auswahlwahrscheinlichkeiten durch das Design der Stichprobe, z. B. werden bei disproportionalen Stichproben Personen aus Bundesländern, die überproportional gezogen wurden, entsprechend geringer gewichtet. Designgewichte sind in der Statistik unstrittig. Ob und wie Korrekturgewichte verwendet werden sollten, ist nicht abschließend geklärt. Bei Längsschnittstudien kommen noch Längsschnittgewichte hinzu, die häufig wiederum Design- und Korrekturgewichte enthalten. Das Gesamtgewicht eines Falls ist immer das Produkt der Einzelgewichte.

Abbildung 14:

Gezogene und realisierte Stichprobe als Teilmengen der Zielpopulation (als identisch mit der Auswahlgesamtheit angenommen; in Anlehnung an Särndal/Lundström 2005, 44)



Für den Fall einer Zufallsstichprobe ohne Zurücklegen der Größe n mit identischen Inklusionswahrscheinlichkeiten $\pi_k = n/N$ und Gewichten $d_k = 1/\pi_k$ für alle Elemente der Stichprobe und ohne Nonresponse stellt das über die Werte y_k berechnete Stichprobentotal $\hat{Y}$ über den Horvitz-Thompson-Schätzer

$$\hat{Y} = \sum_r d_k y_k \tag{16}$$

eine unverzerrte Schätzung für das Total der Population Y mit

$$Y = \sum_U y_k \tag{17}$$

dar.[75] Fallen allerdings Personen durch Nonresponse aus, wird Y nicht mehr unverzerrt über $\hat{Y}$ geschätzt (Särndal und Lundström 2005, 57). Die Schätzung muss dann durch Gewichtung korrigiert werden und erfolgt für alle befragbaren Personen ($k \in r$) über

$$\hat{Y}_W = \sum_r w_k y_k, \tag{18}$$

wobei w_k die Korrekturgewichte der Respondenten sind. Es existieren mehrere Methoden, um diese Korrekturgewichte zu konstruieren, alle diese Methoden haben jedoch gemeinsam, dass der gewichtete Schätzer bessere Eigenschaften aufweisen soll, als die ungewichtete Schätzung.

Hierzu werden zusätzliche Informationen in Form von Hilfsvariablen (*auxiliary variables*) benötigt. Dies können zusätzliche Variablen sein, die für alle zur Befragung vorgesehenen Personen (also Respondenten und Nonrespondenten) oder die gesamte Population vorliegen, aber auch Informationen über die Verteilung der Hilfsvariablen in der Population. So können beispielsweise die zur Stichprobenziehung verwendeten Auswahlgrundlagen zusätzliche Variablen enthalten (z. B. bei Einwohnermeldedateien das Alter). Es können aber auch Daten aus anderen administrativen Registern, Aggregatstatistiken (z. B. Zahl der Krankenhausentlassungen pro Jahr) oder Daten über den Prozess der Datenerhebung selbst (Paradaten) zur Konstruktion der Korrekturgewichte verwendet werden.[76]

Notwendige Voraussetzung ist allerdings, dass die zur Gewichtung herangezogenen Hilfsvariablen mit dem Nonresponse-Mechanismus in Zusammenhang stehen. Sind sie vollkommen unabhängig von der Ausfallursache, wird die Gewichtung keine bessere Schätzung erbringen als die ungewichtete Schätzung.

Aktueller Standard in der Statistik für die Korrektur von Nonresponse ist der sogenannte Calibration-Ansatz (Deville/Särndal 1992), der zahlreiche ältere Verfahren als Spezialfall enthält. Der erste Schritt einer solchen Gewichtung besteht in der Auswahl geeigneter Hilfsvariablen x_k. Anschließend werden im zweiten Schritt Gewichte w_k gesucht, die die sogenannte Kalibrierungsgleichung

$$\sum_r w_k \mathrm{x}_k = \mathrm{X} \tag{19}$$

---

[75] Diese Inklusionswahrscheinlichkeiten können sich für komplexere Stichprobendesigns zwischen den Elementen der Stichprobe (bspw. in disproportional geschichteten Stichproben) unterscheiden. Ist π_i konstant, spricht man von einer EPSEM-Stichprobe (*Equal Probability of Selection Method*, Kish 1965, 21).

[76] Einzelheiten finden sich vor allem bei Särndal/Lundström 2005 und Valliant u. a. 2013, zu Paradaten allgemein siehe die Beiträge in dem Sammelband von Kreuter 2013.

erfüllen. Gewichte, die dieser Gleichung genügen, werden im Hinblick auf X als kalibriert bezeichnet (Särndal/Lundström 2005, 58).

Wie bereits erwähnt müssen im Falle von Stichproben mit ungleichen Inklusionswahrscheinlichkeiten oder der Schätzung des Population Totals über den Horvitz-Thompson-Schätzer (Formel (16)) zusätzlich die durch das Design erforderlichen Gewichte d_k berücksichtigt werden, die als Kehrwert der Inklusionswahrscheinlichkeit über $d_k = 1/\pi_k$ definiert sind. Designgewichte sind notwendig, aber nicht in der Lage, Nonresponse zu korrigieren. Die Designgewichte werden daher mit einem zusätzlichen Faktor υ_k korrigiert. Diese neuen korrigierten Gewichte sind dann als

$$w_k = d_k \upsilon_k \tag{20}$$

definiert.

Wird ein linearer Zusammenhang zwischen den Hilfsvariablen x_k und dem Korrekturfaktor v_k angenommen, ergibt sich der Korrekturfaktor über

$$\upsilon_k = 1 + \lambda' \mathrm{x}_k, \tag{21}$$

wobei in einem nächsten Schritt der Vektor λ aus dieser Gleichung bestimmt werden muss. Wird $w_k = d_k\,(1+\lambda' x_k)$ in Formel (19) eingesetzt und nach λ aufgelöst, so ergibt sich

$$\lambda' = \left(\mathrm{X} - \sum_r d_k \mathrm{x}_k\right)' \left(\sum_r d_k \mathrm{x}_k \mathrm{x}'_k\right)^{-1} \tag{22}$$

als Lösung. Die Korrekturgewichte w_k ergeben sich dann über

$$w_k = d_k + d_k \lambda' \mathrm{x}_k \tag{23}$$

und führen damit zur kalibrierten Schätzung

$$\hat{Y}_W = \sum_r w_k y_k \tag{24}$$

als Produkt der konkreten Messung y_k und des Korrekturgewichts w_k (Särndal/Lundström 2005, 57–59).[77]

[77] Für Kalibrierungen stehen für Statistikprogramme wie STATA oder R kostenlose zusätzliche Makros zur Berechnung zur Verfügung, daneben gibt es auch teilkommerzielle Lösungen wie z. B. BASCULA als Teil von BLAISE, dem CATI-System des niederländischen CBS.

Es muss nochmals betont werden, dass die Übereinstimmung einiger Verteilungen der Stichprobe mit einigen Verteilungen in der Grundgesamtheit nicht beweist, dass die Stichprobe eine „repräsentative“ Stichprobe, frei von Verzerrungen oder eine Zufallsstichprobe ist. Solche Übereinstimmungen zeigen nur, dass der Selektionsmechanismus nicht mit den überprüften Variablen zusammenhängt (Schnell 1993). Trotz übereinstimmender Randverteilungen (gleichgültig ob vor oder nach einer Gewichtung oder Kalibrierung) zwischen Stichprobe und Grundgesamtheit sind irreführende Verallgemeinerung oder verzerrte Schätzungen keineswegs ausgeschlossen oder weniger wahrscheinlich. Nehmen wir z. B. an, ältere Menschen würden aus gesundheitlichen Gründen weniger an Befragungen teilnehmen. Trotzdem wird es einige ältere Menschen in der Stichprobe geben, die vermutlich etwas gesünder als die älteren Nichtteilnehmer sind. Werden diese gesunden Älteren nun höher gewichtet, dann wird der Anteil der Gesunden in der Grundgesamtheit überschätzt, obwohl der Anteil der Älteren, eventuell auch Geschlecht und Bildung mit der Verteilung in der Grundgesamtheit übereinstimmen. Gewichten verringert Verzerrungen nur dann, wenn die Gewichtungsvariablen mit dem Selektionsmechanismus stark zusammenhängen. Das lässt sich empirisch anhand einer gegebenen Stichprobe allein nicht überprüfen, sondern nur mittels einer Stichprobe ohne Nonresponse oder der Grundgesamtheit. Stehen diese Daten nicht zur Verfügung, müssen heroische Annahmen getroffen werden. Das ist in keiner Weise problematisch, muss aber klar in einer Studiendokumentation thematisiert werden.

Wir haben hierzu ein fiktives Beispiel mit den Daten des *British Crime Survey* (BCS) 2010–2011 berechnet. Für dieses Beispiel wurde die Stichprobe des BCS (etwa $n = 47.000$) als Population verwendet, aus der eine Zufallsstichprobe der Größe $n = 3.000$ gezogen wurde. Für diese Stichprobe wurde dann der Ausfall von ca. 25 % der Befragten in Abhängigkeit von der Anzahl der erwachsenen Haushaltsmitglieder, des durch die Interviewer wahrgenommenen Ausmaßes an Incivilities sowie der Lage des Haushalts (Innenstadt: ja/nein) modelliert.[78] Als Hilfsinformationen für die Kalibrierung konnten zwei Quellen verwendet werden: einerseits die Informationen aus der Population (BCS-Gesamtstichprobe). Hier werden „Populationstotale“ für fünf Altersklassen und Geschlecht verwendet, die beispielsweise der amtlichen Statistik entnommen werden könnten. Weiterhin können die für alle Personen (also Respondenten und Nonrespondenten) vorliegenden Variablen „Zahl der er-

[78] Die Gewichte, mit denen diese Variablen in die Ausfallmodellierung eingingen, wurden über eine Regression des Kriminalitätsfurcht-Standardindikators auf diese Variablen ermittelt. Zu den resultierenden vorhergesagten Werten wurde eine Zufallskomponente addiert ($N(0; 0.25)$) und die 25 % größten Werte dann als fehlend markiert. Ohne diese Zufallskomponente wäre durch die Kalibrierung eine unverzerrte Schätzung erreicht worden.

wachsenen Haushaltsmitglieder“, „durch Interviewer wahrgenommene Incivilities“ sowie „Innenstadtlage des Haushalts“ verwendet werden. Diese Art von Informationen kann entweder im Sampling-Frame vorliegen oder durch die Interviewer während der Erhebung gesammelt werden (Paradaten). Mit diesen Informationen wurden dann die kalibrierten Korrekturgewichte in einem zweistufigen Verfahren konstruiert (Typ B, siehe Särndal/Lundström 2005, 81–83). Für dieses Beispiel ergibt die ungewichtete Berechnung des Anteils „ängstlicher Personen“ einen Wert von 24,8 %.[79] Wird die Schätzung mit den durch das Calibrate-Verfahren konstruierten Korrekturgewichten durchgeführt, ergibt sich ein Wert von 26,8 %, was einer Steigerung des Anteils ängstlicher Personen von ungewichteter zu gewichteter Schätzung von 8,1 % entspricht. Der Populationsparameter μ beträgt 26,2 %. Damit ist die kalibrierte Schätzung $\bar{y}_w$ mit einer Abweichung von 0,6 % deutlich genauer als die unkalibrierte Schätzung $\bar{y}_u$ mit einer Abweichung von −1,4 %.

Häufig wird übersehen, dass sich durch Gewichtung der Gesamtfehler gemessen als MSE (siehe Gleichung 2) erhöhen kann. Dies ist dann der Fall, wenn der gewichtungsbedingte Präzisionsgewinn durch Biasreduktion durch eine ebenfalls gewichtungsbedingte Varianzinflation des Schätzers übertroffen wird (Kish 1965, 424–433; Elliot/Little 2000, 192). Daher können ungewichtete, verzerrte Schätzer bessere Ergebnisse liefern als gewichtete, unverzerrte Schätzer.[80] Damit ist insbesondere dann zu rechnen, wenn die Gewichte selbst eine große Streuung aufweisen oder es sich um kleine Stichproben handelt (Elliot/Little 2000, 192).[81]

Für ungewichtete Daten kann der MSE durch

$$MSE(\bar{y}_u) = S^2\left(1 + \frac{B^2}{S^2}\right) = S^2\left(1 + \left(\frac{\bar{y}_u - \bar{y}_w}{\sigma_{y_u}}\right)^2\right) \qquad (25)$$

---

79 Als Variable wurde der Standardindikator zur Kriminalitätsfurchtmessung verwendet und vor der Analyse dichotomisiert.

80 Dieser als *Bias-Variance-Tradeoff* bekannte Effekt ist dadurch erklärbar, dass sich die Varianz eines Schätzers, zum Beispiel des Mittelwerts, durch die Gewichtung von S^2/n auf $S^2\left(1 + s_k^2/\bar{k}^2\right)/n$ erhöht (Kish 1992). Hierbei stellt s_k^2 die Varianz der Gewichte und $\bar{k}$ den Mittelwert der Gewichte dar.

81 Aus diesem Grund werden in der Praxis Gewichte häufig numerisch begrenzt („getrimmt“), sodass kein Fall z. B. ein Gewicht über 3 oder 10 erhält. Der Vollständigkeit halber soll erwähnt werden, dass vor allem bei Kalibrierungen auch negative Gewichte entstehen können. Da dies in der Regel schwer zu vermitteln ist, werden Gewichte daher häufig auch nach unten begrenzt.

berechnet werden. Für die gewichteten Daten kann die Varianz über

$$Var(\bar{y}_w) = S^2(1 + L) = S^2\left(1 + \frac{s_k^2}{\bar{k}^2}\right) \qquad (26)$$

mit dem quadrierten Variationskoeffizienten (L) der Gewichte (k) berechnet werden (Kish 1992, 191).

Für das vorherige fiktive Beispiel mit den Daten des BCS ergibt sich ein Wert für $Var(\bar{y}_w)$ von $1 + L = 1.11$ und entsprechend ein Faktor für den $MSE(\bar{y}_u)$ von $1 + (B/S)^2 = 5{,}89$. Da bessere Schätzungen kleinere MSE-Werte aufweisen, sollten die Daten hier gewichtet werden.[82]

12 Empfehlungen

Die Gesamtkosten, die Probleme und die Dauer von Befragungen werden von Laien zumeist unterschätzt (Schnell 2012, 24–25). Das gilt insbesondere für die notwendigen Stichprobengrößen für Viktimisierungsstudien und die Maßnahmen zur Reduktion von Nonresponse.

Die tatsächlichen Konfidenzintervalle sind zumeist sehr viel größer, als es naive Analysen erwarten lassen. Aus diesem Grund sind vorbildliche Studien wie der NCVS und der BCS von beeindruckender Größe. Der British Crime Survey (nun *Crime Survey for England and Wales*, CSEW) umfasste 2012/2013 insgesamt 37.759 Personen; der NCVS lag 2013 bei 90.630 Haushalten mit 160.040 Personen. Entsprechend den hohen methodischen Anforderungen an Viktimisierungsstudien sind solche Surveys kostspielig: Die jährlichen Kosten des NCVS wurden 2012 auf 27 Millionen Dollar geschätzt.[83] Diese Kosten ergeben sich vor allem durch die zusätzlichen Maßnahmen zur Verhinderung von Nonresponse, vor allem durch aufwendige Mehrfachkontakte in verschiedenen Erhebungsmodi, Verweigerungskonvertierungen und den Einsatz sicherer Incentives.

---

82 Natürlich ist auch hier zusätzlich der in Kapitel 8 auf komplexe Stichprobendesigns zurückgehende Präzisionsverlust relevant.

83 Bereinigt man dies um die Einwohnerzahl (Faktor 0,25) und das Bruttosozialprodukt pro Kopf (Faktor 0,85) sowie den Währungskurs (Faktor 0,80), dann entspräche dies 4,6 Millionen Euro; in dieser Größenordnung dürfte auch der BCS liegen. Deutschland leistet sich mit dem Mikrozensus eine Erhebung von 830.000 Personen in ca. 370.000 Haushalten mit geschätzten Kosten von 21,6 Millionen Euro (Bundestagsdrucksache 17/10041 vom 19. 06. 2012). Eine Erhebung der Größe des NCVS oder des BCS/CSEW ist in Deutschland unter den gegebenen politischen Bedingungen kaum durchsetzbar.

Es gibt keine Möglichkeit, mit der solche Kosten vermieden werden können, falls man belastbare Aussagen für politisch interessante Subgruppen (wie Bundesländer oder Personen mit Migrationshintergrund) treffen möchte. Um das nochmals deutlich zum Ausdruck zu bringen: Es gibt weder statistische Zaubertricks noch „moderne Erhebungsmethoden“, die mit kleineren Kosten vergleichbar präzise Resultate wie sehr große, langwierige und aufwendige Surveys produzieren können. Dies wird sich keineswegs im Laufe der Zeit bessern – eher im Gegenteil.

Wir raten daher im Zweifelsfall eher von kleineren Erhebungen ab. Belastbare Aussagen lassen sich mit kommunalen Studien (gar durch schriftliche Befragungen oder Lehrforschungsprojekte) mit geringem Aufwand nicht erzielen. Die Zukunft wissenschaftlicher Viktimisierungsstudien sind wenige, sehr große und methodisch sehr aufwendige Erhebungen, die ein einzelnes Institut nicht leisten kann. Die für solche Erhebungen notwendigen finanziellen Mittel werden die Zahl solcher Studien erheblich reduzieren; vermutlich wird sich Deutschland eine solche Studie nur in größeren Intervallen leisten wollen.

Trotz dieser Überlegungen gibt es Fragestellungen, für die quantitative Vorstudien sinnvoll sein können. Hierzu gehören vor allem Erhebungen in Institutionen und die Erhebung von Spezialpopulationen.[84] Empfehlungen für die Durchführung solcher Erhebungen haben wir im Folgenden zusammengefasst.

– Erhebung in einer Institution: Verwendung einer Liste der Insassen, der dort arbeitenden oder wohnenden Personen. Uneingeschränkte Zufallsauswahl, bei sehr kleinen Institutionen Vollerhebung. Bei besonderem Interesse an kleinen Teilgruppen geschichtete Auswahl. Schriftliche Befragung mit mehrfacher Mahnung.

– Spezialpopulationen: Liegt eine Liste der Mitglieder der Population vor, sollte eine uneingeschränkte Zufallsstichprobe gezogen werden, wobei der Erhebungsmodus den Kosten entsprechend gewählt werden kann. Liegt keine Liste vor, müssen spezielle Verfahren eingesetzt werden; die Durchführung solcher Verfahren sollte delegiert werden. Von der Verwendung kumulierter Screening-Interviews in Telefonbussen oder Schneeballverfahren sollte abgesehen werden.

---

84 Die Ziehung von Zufallsstichproben seltener Populationen (konventionell weniger als 5 % oder 1 % der „allgemeinen“ Bevölkerung) ist ein eigenes Gebiet innerhalb der Stichprobenverfahren. Eine Übersicht findet sich bei Schnell u. a. 2013.

– Kommunale Erhebung: Einwohnermelderegister als Auswahlgrundlage, Stichprobe größer als 2.000, uneingeschränkte Zufallsauswahl. Schriftliche Befragung mit mehrfacher Mahnung. Nonresponse-Stichprobe mit Face-to-Face-Erhebung zur Kontrolle des Nonresponse-Bias. Aufwand mindestens zwei Personenjahre.

– Bundesweite Erhebung: Stichprobenziehung und Datenerhebung an ein großes sozialwissenschaftliches Institut delegieren. Keine Websurveys, schriftliche Befragungen kaum empfehlenswert. CATI als Erstmodus, Dual-Frame-Stichprobe größer als 2.000, Nonresponse-Stichprobe (Face-to-Face) dringend empfohlen. Reine Erhebungskosten oberhalb von 150.000 Euro.

13 Literatur

AAPOR (2011): Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys. Lenexa, KS: American Association for Public Opinion Research.

AAPOR (2010): AAPOR Report on Online Panels. Prepared for the AAPOR Executive Council by a Task Force operating under the auspices of the AAPOR Standards Committee, with members including: Reg Baker, Stephen Blumberg, J. Michael Brick, Mick P. Couper, Melanie Courtright, Mike Dennis, Don Dillman, Martin R. Frankel, Philip Garland, Robert M. Groves, Courtney Kennedy, Jon Krosnick, Sunghee Lee, Paul J. Lavrakas, Michael Link, Linda Piekarski,Kumar Rao, Douglas Rivers, Randall K. Thomas, Dan Zahs. In: Public Opinion Quarterly, 74, S. 711–781.

Aebi, Marcelo F.; Linde, Antonia (2014): National Victimization Surveys. In: Bruinsma, Gerben; Weisburd, David (Hg.): Encyclopedia of Criminology and Criminal Justice. New York: Springer, S. 3228–3242.

Ahlborn, Wilfried; Böker, Fred und Lehnick, Dirk (1993): Stichprobengrößen bei Opferbefragungen in der Dunkelfeldforschung, Wiesbaden: Bundeskriminalamt.

Albers, Ines (1997): Einwohnermelderegister-Stichproben in der Praxis. In: Gabler, Siegfried; Hoffmeyer-Zlotnik, Jürgen H. P. (Hg.): Stichproben in der Umfragepraxis. Opladen: Westdeutscher Verlag, S. 117–126.

Anderson, Andy B.; Basilevsky, Alexander und Hum, Derek P. I. (1983): Missing Data. A Review of the Literature. In: Rossi, Peter H.; Wright, James D. und Anderson, Andy B. (Hg.): Handbook of Survey Research. New York: Academic Press, S. 415–494.

Bandilla, Wolfgang; Kaczmirek, Lars; Blohm, Michael; Neubarth, Wolfgang; Jackob, Nikolaus; Schoen, Harald und Zerback, Thomas (2009): Coverage- and Nonresponse-Effekte bei Online-Bevölkerungsumfragen. In: Jackob, Nikolaus; Schoen, Harald und Zerback, Thomas (Hg.): Sozialforschung im Internet. Methodologie und Praxis der Online-Befragung. Wiesbaden: VS-Verlag, S. 129–144.

Bentrup, Christina; Verneuer, Lena (2014): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2011. In: Schriftenreihe Jugendkriminalität in der modernen Stadt – Methoden, Nr. 20/2014.

Bethlehem, Jelke; Cobben, Fannie und Schouten, Barry (2011): Handbook of Nonresponse in Household Surveys. Hoboken: Wiley.

Biemer, Paul P. (2010): Total Survey Error. Design, Implementation, and Evaluation. In: Public Opinion Quarterly, 74, S. 817–848.

Biemer, Paul P.; Lyberg, Lars E. (2003): Introduction to Survey Quality, Hoboken: Wiley.

Bortz, Jürgen (2005): Statistik für Human- und Sozialwissenschaftler, 6. Aufl. Heidelberg: Springer.
Bruinsma, Gerben; Weisburd, David (Hg.) (2014): Encyclopedia of Criminology and Criminal Justice. New York: Springer.
Bundesarbeitsgemeinschaft Wohnungslosenhilfe (2013): Zahl der Wohnungslosen in Deutschland weiter gestiegen. Pressemitteilung. URL: http://www.bagw.de/media/doc/PRM_2013_08_01_Zahl_der_Wohnungslosen.pdf. – Download vom 09. 04. 2015.
CLANDESTINO Project (2009): Undocumented Migration: Counting the Uncountable. Data and Trends Across Europe. Final Report. Athens.
Cohen, Lawrence E.; Felson, Marcus (1979): Social Change and Crime Rate Trends. A Routine Activity Approach. In: American Sociological Review, 44, S. 588–608.
Converse, Philip E.; Traugott, Michael W. (1986): Assessing the Accuracy of Polls and Surveys. In: Science, 234, S. 1094–1098.
Deville, Jean-Claude; Särndal, Carl-Erik (1992): Calibration Estimators in Survey Sampling. In: Journal of the American Statistical Association, 87, S. 376–382.
Doblhammer, Gabriele; Schulz, Anne; Steinberg, Juliane und Ziegler, Uta (2012): Demografie der Demenz. Bern: Verlag Hans Huber.
Elliot, Michael R.; Little, Roderick J. A. (2000): Model-based Alternatives to Trimming Survey Weights. In: Journal of Official Statistics, 16, S. 191–209.
Enders, Craig K. (2010): Applied Missing Data Analysis. New York: Guilford Press.
Enzmann, Dirk (2010): Germany. In: Junger-Tas, Josine; Marshall, Ineke H.; Enzmann, Dirk; Killias, Martin; Steketee, Majone und Gruszczynska, Beate (Hg.): Juvenile Delinquency in Europe and Beyond. Results of the Second International Self-reported Delinquency Study. Dordrecht: Springer, S. 47–64.
Fahrmeir, Ludwig; Künstler, Rita; Pigeot, Iris und Tutz, Gerhard (2007): Statistik, 6. Aufl. Berlin/Heidelberg: Springer.
Geiger, Marion; Styhler, Doris (2012): ZENSUS 2011: Erhebungsteil Sonderbereiche. In: Bayern in Zahlen, 5, S. 280–285.
Glemser, Axel; Meier, Gerd und Heckel, Christiane (2014): Dual-Frame: Stichprobendesign für CATI-Befragungen im mobilen Zeitalter. In: ADM Arbeitskreis Deutscher Markt und Sozialforschungsinstitute e. V. (Hg.): Stichproben-Verfahren in der Umfrageforschung. Eine Darstellung für die Praxis, 2. Aufl. Wiesbaden: Springer VS, S. 167–190.
Groves, Robert M. (1989): Survey Errors and Survey Costs. New York: Wiley.

Groves, Robert M.; Cork, Daniel L. (Hg.) (2008): Surveying Victims. Options for Conducting the National Crime Victimization Survey, Washington: The National Academies Press.
Groves, Robert M.; Magilavy, Lou J. (1986): Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys. In: Public Opinion Quarterly, 50, S. 251–266.
Groves, Robert M.; Peytcheva, Emilia (2008): The Impact of Nonresponse Rates on Nonresponse Bias. A Meta-analysis. In: Public Opinion Quarterly, 72, S. 167–189.
Häder, Sabine; Gabler, Siegfried und Heckel, Christiane (2009): Stichprobenziehung für die CELLA-Studie. In: Häder, Michael; Häder, Sabine (Hg.): Telefonbefragungen über das Mobilfunknetz. Konzept, Design und Umsetzung einer Strategie zur Datenerhebung. Wiesbaden: VS-Verlag, S. 21–49.
Haug, Sonja (2008): Sprachliche Integration von Migranten in Deutschland. Bundesamt für Migration und Flüchtlinge, Workingpaper 14.
Heckel, Christiane; Glemser, Alex und Meier, Gerd (2014): Das ADM-Telefonstichproben-System. In: ADM Arbeitskreis Deutscher Markt und Sozialforschungsinstitute e. V. (Hg.): Stichproben-Verfahren in der Umfrageforschung. Eine Darstellung für die Praxis, 2. Aufl. Wiesbaden: Springer VS, S. 137–166.
Heckel, Christiane; Wiese, Kathrin (2012): Sampling Frames for Telephone Surveys in Europe. In: Häder, Sabine; Häder, Michael und Kühne, Mike (Hg.): Telephone Surveys in Europe. Research and Practice. Berlin/Heidelberg: Springer, S. 103–119.
von der Heyde, Christian (2014): Einwohnermeldeamts-Stichproben (EWA-Stichproben). In: ADM Arbeitskreis Deutscher Markt und Sozialforschungsinstitute e. V. (Hg.): Stichproben-Verfahren in der Umfrageforschung. Eine Darstellung für die Praxis, 2. Aufl. Wiesbaden: Springer VS, S. 191–195.
Hindelang, Michael J.; Gottfredson, Michael R. und Garofalo, James (1978): Victims of Personal Crime. An Empirical Foundation for a Theory of Personal Victimization. Cambridge: Ballinger.
Hunsicker, Stefan; Schroth, Yvonne (2014): Dual-Frame-Ansatz in politischen Umfragen. Arbeitspapiere der Forschungsgruppe Wahlen e. V., Mannheim, Nr. 2 – April 2014.
Johnson, Brian (2014): Sample Selection Models. In: Bruinsma, Gerben; Weisburd, David (Hg.): Encyclopedia of Criminology and Criminal Justice. New York: Springer, S. 4561–4580.
Kalton, Graham (1983): Introduction to Survey Sampling, Newbury Park: Sage.

Kish, Leslie (1965): Survey Sampling, New York: Wiley.Kish, Leslie (1992): Weighting for Unequal P_i. In: Journal of Official Statistics, 8, S. 183–200.

Klausch, Thomas; Hox, Joop und Schouten, Barry (2015): Selection Error in Single- and Mixed Mode Surveys of the Dutch General Population. In: Journal of the Royal Statistical Society, Series A (Statistics in Society), Early View (Online Version of Record published before inclusion in an issue). URL: http://onlinelibrary.wiley.com/doi/10.1111/rssa.12102/full – Download vom 09. 04. 2015.

Klingwort, Jonas (2014): Nonresponse in aktuellen deutschen Viktimisierungssurveys, Bachelorarbeit. Universität Duisburg-Essen: Institut für Soziologie.

Kreuter, Frauke (Hg.) (2013): Improving Surveys with Paradata – Analytic Uses of Process Information. Hoboken: Wiley.

Kroh, Martin (2014): Documentation of Sample Sizes and Panel Attrition in the German Socio-Economic Panel (SOEP) (1984 until 2012). SOEP Survey Papers 177: Series D. Berlin: DIW/SOEP.

Kruskal, William; Mosteller, Frederick (1979a): Representative Sampling, I: Non-scientific Literature. In: International Statistical Review, 47, S. 13–24.

Kruskal, William; Mosteller, Frederick (1979b): Representative Sampling, II: Scientific Literature, Excluding Statistics. In: International Statistical Review, 47, S. 111–127.

Kruskal, William; Mosteller, Frederick (1979c): Representative Sampling, III: the Current Statistical Literature. In: International Statistical Review, 47, S. 245–265.

Kruskal, William; Mosteller, Frederick (1980): Representative Sampling, IV: the History of the Concept in Statistics 1895–1939. In: International Statistical Review, 48, S. 169–195.

Lepkowski, James M. (1988): Telephone Sampling Methods in the United States. In: Groves, Robert M.; Biemer, Paul P.; Lyberg, Lars E.; Massey, James T.; Nicholls, William L. und Waksberg, Joseph (Hg.): Telephone Survey Methodology. New York: Wiley, S. 73–98.

Link, Michael W.; Fahimi, Mansour (2008): Telephone Survey Sampling. In: Levy, Paul S.; Lemeshow, Stanley (Hg.): Sampling of Populations. Methods and Applications, 4. Aufl. Hoboken: Wiley, S. 455–487.

Little, Roderick J. A.; Rubin, Donald B. (2002): Statistical Analysis with Missing Data, 2. Aufl. Hoboken: Wiley.

Löffler, Ute; Behrens, Kurt und von der Heyde, Christian: (2014): Die Historie der ADM-Stichproben. In: ADM Arbeitskreis Deutscher Markt und Sozialforschungsinstitute e. V. (Hg.): Stichproben-Verfahren in der Umfrageforschung. Eine Darstellung für die Praxis, 2. Aufl. Wiesbaden: Springer VS, S. 67–83.

Lohr, Sharon L. (2010): Sampling: Design and Analysis, 2. Aufl. Boston: Brooks/Cole.
Lynch, James P. (2006): Problems and Promise of Victimization Surveys for Cross-national Research. In: Crime and Justice, 34, S. 229–287.
Lynn, Peter (1997): Sampling Frame Effects on the British Crime Survey. In: Journal of the Royal Statistical Society, Series A (Statistics in Society), 160, S. 253–269.
Mayer, Raimund (2013): Zensus 2011: Erhebung an Anschriften mit Sonderbereichen. In: Statistische Monatshefte Niedersachsen, 12, S. 672–679.
Mitofsky, Warren (1970): Sampling of Telephone Households, unveröffentlichtes CBS-News-Memorandum.
National Research Council (2012): Small Populations, Large Effects: Improving the Measurement of the Group Quarters Population in the American Community Survey, Washington DC: The National Academies Press.
National Research Council (2014): Estimating the Incidence of Rape and Sexual Assault, Washington DC: The National Academies Press.
Noack, Marcel (2015): Methodische Probleme bei der Messung von Kriminalitätsfurcht und Viktimisierungserfahrungen, Wiesbaden: Springer VS.
Pöge, Andreas (2007): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002 bis 2005 (Vier-Wellen-Panel). In: Schriftenreihe Jugendkriminalität in der modernen Stadt – Methoden, Nr. 13/2007.
Puhani, Patrick A. (2000): The Heckman Correction for Sample Selection and Its Critique. In: Journal of Economic Surveys, 14, S. 53–68.
Särndal, Carl-Erik; Lundström, Sixten (2005): Estimation in Surveys with Nonresponse. Chichester: Wiley.
Schafer, Joseph L. (1997): Analysis of Incomplete Multivariate Data. Boca Raton: Chapman & Hall/CRC.
Schnell, Rainer (1986): Missing-Data-Probleme in der empirischen Sozialforschung. Dissertation. Ruhr-Universität Bochum.
Schnell, Rainer (1991): Wer ist das Volk? Zur faktischen Grundgesamtheit bei „allgemeinen Bevölkerungsumfragen“: Undercoverage, Schwererreichbare und Nichtbefragbare. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 43, S. 106–137.
Schnell, Rainer (1993): Die Homogenität sozialer Kategorien als Voraussetzung für „Repräsentativität“ und Gewichtungsverfahren. In: Zeitschrift für Soziologie, 22, S. 16–32.
Schnell, Rainer (1997): Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen, Opladen: Leske+Budrich.

Schnell, Rainer (1998): Besuchs- und Berichtsverhalten der Interviewer. In: Statistisches Bundesamt (Hg.): Interviewereinsatz und -qualifikation, Nummer 11 in Spektrum Bundesstatistik. Stuttgart: Metzler-Poeschel, S. 156–170.

Schnell, Rainer (2008): Avoiding Problems of Traditional Sampling Strategies for Household Surveys in Germany: Some New Suggestions, DIW Data-Documentation 33, Berlin.

Schnell, Rainer (2012): Survey-Interviews. Methoden standardisierter Befragungen, Wiesbaden: VS-Verlag.

Schnell, Rainer; Hill, Paul. B. und Esser, Elke (2013): Methoden der empirischen Sozialforschung, 10. Aufl. München: Oldenbourg.

Schnell, Rainer; Hoffmeyer-Zlotnik, Jürgen H. P. (2002): Methodik für eine regelmäßige Opferbefragung, Gutachten im Auftrag des BMI/BMJ, Universität Konstanz.

Schnell, Rainer; Kreuter, Frauke (2000): Untersuchungen zur Ursache unterschiedlicher Ergebnisse sehr ähnlicher Viktimisierungssurveys. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 52, S. 96–117.

Schnell, Rainer; Kreuter, Frauke (2005): Separating Interviewer and Sampling-point Effects. In: Journal of Official Statistics, 21, S. 389–410.

Schnell, Rainer; Noack, Marcel (2014): The Accuracy of Pre-Election Polling of German General Elections. In: MDA – Methods, Data, Analyses, 8, S. 5–24.

Schonlau, Matthias; Weidmer, Beverly und Kapteyn, Arie (2014): Recruiting an Internet Panel Using Respondent-driven Sampling. In: Journal of Official Statistics, 30, S. 291–310.

Schouten, Barry; Cobben, Fannie und Bethlehem, Jelke (2009): Indicators for the Representativeness of Survey Response. In: Survey Methodology, 35, S. 101–113.

Schwarz, Norbert (2007): Retrospective and Concurrent Self-reports. The Rationale for Real-time Data Capture. In: Stone, Arthur A.; Shiffman, Saul; Atienza, Audie A. und Nebeling, Linda (Hg.): The Science of Real-time Data Capture. Self-reports in Health Research. New York: Oxford University Press, S. 11–26.

Shaffer, Jennifer N.; Ruback, R. Barry (2002): Violent Victimization as a Risk Factor for Violent Offending Among Juveniles, Washington DC: Office of Juvenile Justice and Delinquency Prevention.

Smith, Adrian (2006): Crime Statistics. An Independent Review. Carried out for the Secretary of State for the Home Department, London: Home Office.

Statistische Ämter des Bundes und der Länder (2004): Ergebnisse des Zensustests. In: Wirtschaft und Statistik, 8, S. 813–833.

Statistische Ämter des Bundes und der Länder (Hg.) (2014): Gebäude- und Wohnungsbestand in Deutschland. Erste Ergebnisse der Gebäude- und Wohnungszählung 2011, Hannover: Landesamt für Statistik Niedersachsen.

Stolzenberg, Ross M.; Relles, Daniel A. (1997): Tools for Intuition about Sample Selection Bias and Its Correction. In: American Sociological Review, 62, S. 494–507.

Tucker, Clyde (1983): Interviewer Effects in Telephone Surveys. In: Public Opinion Quarterly, 47, S. 84–95.

Valliant, Richard; Dever, Jill A. und Kreuter, Frauke (2013): Practical Tools for Designing and Weighting Survey Samples. New York: Springer.

Vaughn, Michael G.; Maynard, Brandy; Salas-Wright, Christopher; Perron, Brian E. und Abdon, Arnelyn (2013): Prevalence and Correlates of Truancy in the US. Results from a National Sample. In: Journal of Adolescence, 36, S. 767–776.

Vella, Francis (1998): Estimating Models with Sample Selection Bias. A Survey. In: Journal of Human Resources, 33, S. 127–169.

Vercambre, Marie-Noël; Gilbert, Fabien (2012): Respondents in an Epidemiologic Survey Had Fewer Psychotropic Prescriptions than Nonrespondents. An Insight into Health-related Selection Bias Using Routine Health Insurance Data. In: Journal of Clinical Epidemiology, 65, S. 1181–1189.

Vogel, Dita; Aßner, Manuel (2011): Umfang, Entwicklung und Struktur der irregulären Bevölkerung in Deutschland. Expertise im Auftrag der deutschen nationalen Kontaktstelle für das Europäische Migrationsnetzwerk (EMN) beim Bundesamt für Migration und Flüchtlinge. URL: http://www.bamf.de/SharedDocs/Anlagen/DE/Publikationen/EMN/Expertisen/emn-wp-41-expertise-de.pdf – Download vom 09. 04. 2015.

Waksberg, Joseph (1978): Sampling Methods for Random Digit Dialing. In: Journal of the American Statistical Association, 19, S. 103–113.

Weisberg, Herbert F. (2005): The Total Survey Error Approach. A Guide to the New Science of Survey Research, Chicago: The University of Chicago Press.

Weissenberger-Leduc, Monique; Weiberg, Anja (2011): Gewalt und Demenz. Ursachen und Lösungsansätze für ein Tabuthema in der Pflege. Wien: Springer.

Weyerer, Siegfried (2005): Altersdemenz. In: Gesundheitsberichterstattung des Bundes, 28, S. 1–33.

Wolter, Kirk M. (2007): Introduction to Variance Estimation, 2. Aufl. New York: Springer.

Ziegler, Uta; Doblhammer, Gabriele (2009): Prävalenz und Inzidenz von Demenz in Deutschland. In: Das Gesundheitswesen, 71, S. 281–290.