# Planung von Stichprobenerhebungen

Karl-August Schäffer

Statistische Erhebungen haben die Aufgabe, Informationen über eine abstrakt definierte, aber real existierende Zielgesamtheit (*target population*) von Einheiten (z.B. Personen, Institutionen, physikalischen Objekten) durch Erfassung von Merkmalen zu erlangen und daraus quantitative Aussagen abzuleiten. Derartige Aussagen stimmen generell nicht exakt mit der Realität überein, sondern unterscheiden sich von ihr um Erhebungsfehler.

Zwei Gruppen von Erhebungsfehlern sind zu unterscheiden: Stichprobenfehler (*sampling errors*) und systematische Fehler (*nonsampling errors*).

Stichprobenfehler ergeben sich zwangsläufig aus der Entscheidung, mit Rücksicht auf die verfügbaren finanziellen, zeitlichen und personellen Ressourcen anstelle der Gesamtheit nur eine Teilmenge davon zu untersuchen und aus den Eigenschaften der Teilgesamtheit, der Stichprobe (*sample*), einen Schluß auf die Eigenschaften der Gesamtheit zu ziehen. Strukturelle Unterschiede zwischen Gesamtheit und Stichprobe sind unvermeidbar; sie führen zu den Stichprobenfehlern.

Eine Sonderstellung nehmen Zufallsstichproben (*random samples*) ein. Sie sind definiert als Stichproben, die nach einer Zufallsprozedur so gezogen werden, daß jeder möglichen Stichprobe eine Wahrscheinlichkeit für ihre Realisierung zugeordnet wird. Zufallsstichproben bieten den Vorteil, daß die Größenordnung der Stichprobenfehler aufgrund der Stichprobentheorie bei der Planung gesteuert und nach der Erhebung allein aus den mit der Stichprobe gewonnenen Daten ermittelt werden kann.

Systematische Fehler umfassen alle übrigen Fehlerquellen, die zu den Erhebungsfehlern beitragen. Im Gegensatz zu den Stichprobenfehlern von Zufallsstichproben existiert für die Abschätzung der systematischen Fehler keine allgemein anwendbare Theorie.

Drei Arten von systematischen Fehlern sind zu unterscheiden (vgl. Lessler & Kalsbeek, 1992): Fehler im Erhebungsrahmen (*frame errors*), Fehler infolge von Antwortausfällen (*nonresponse errors*) und Meßfehler (*measurement errors*).

Fehler im Erhebungsrahmen beruhen darauf, daß die abstrakte Definition der Zielgesamtheit nicht geeignet ist, die zu ihr gehörenden Einheiten für die Erhebung abzugrenzen, zu identifizieren und zu erreichen. Für diese Zwecke sowie für das Ziehen von Zufallsstichproben wird ein Hilfsmittel (in Form einer Liste, Datei, Karte), der sogenannte Erhebungsrahmen (*frame*), benötigt. Die durch den Erhebungsrahmen explizit definierte Gesamtheit heißt Erhebungsgesamtheit (*frame population*). Falls diese Gesamtheit nicht deckungsgleich mit der Zielgesamtheit ist (z.B. weil sie nicht alle Einheiten der Zielgesamtheit enthält), sind systematische Fehler in den Ergebnissen zu erwarten.

https://doi.org/10.25521/HQM02

Fehler infolge von Antwortausfällen sind darauf zurückzuführen, daß für Einheiten, die zur Stichprobe gehören, trotz aller Bemühungen überhaupt keine oder keine vollständigen Informationen zu erlangen sind. Erfahrungsgemäß unterscheidet sich die Menge der nicht antwortenden Einheiten in ihrer Struktur deutlich von der Menge der antwortenden Einheiten, so daß sowohl der Ersatz fehlender Angaben durch vorliegende Werte als auch die Beschränkung der Studie auf die verfügbaren Daten zu Fehlern infolge der Antwortausfälle führen.

Meßfehler treten auf, falls einer Einheit unzutreffende Werte zugeordnet werden. Derartige Fehler entstehen z.B. dann, wenn eine Fragestellung mißverständlich oder ein Instrument nicht geeicht ist, unzutreffende Auskünfte bewußt oder infolge von Erinnerungsfehlern gegeben werden, Interviewer vorgegebene Fragen nicht korrekt stellen oder richtige Antworten falsch notieren oder richtige Werte infolge von Versehen unrichtig kodiert werden.

Die noch immer häufig anzutreffende Vorstellung, eine wie auch immer gezogene Stichprobe sei ein verkleinertes Bild der Gesamtheit, das ihr in struktureller Hinsicht genau entspreche und deshalb kurzerhand als „repräsentativ" für die Gesamtheit interpretiert werden könne, ist grundsätzlich nicht haltbar: Die Struktur von Stichproben weicht zwangsläufig – selbst im Idealfall – von der Struktur der Gesamtheit ab, und zwar umso mehr, je kleiner die Zahl der Einheiten in der Stichprobe und je kleiner die sachlich interessierenden Teilgesamtheiten sind. Die oben skizzierte Systematik zeigt zudem, daß die sachgerechte Interpretation der Ergebnisse einer Stichprobenerhebung nicht allein die Stichprobenfehler, sondern auch die verschiedenartigen systematischen Fehler der Ergebnisse berücksichtigen muß. Diese Forderung wird offensichtlich immer wieder mißachtet, wie auch das folgende Zitat von Lessler und Kalsbeek (1992, S. 7) zeigt:

> *In many cases researchers are simply reluctant to face the problems that may be present in the survey. An „ignorance is bliss" attitude seems to prevail and gratuitous assumptions are made about the quality of the data (the 7000 nonrespondents are adequately represented by the 3000 respondents).*

Die Tatsache, daß Ergebnisse von Erhebungen nicht fehlerfrei sein können, sollte jedoch nicht erst bei ihrer Interpretation, sondern bereits im Planungsstadium beachtet werden: Bei der Planung einer Stichprobenerhebung sind außer den relevanten realwissenschaftlichen, technischen und organisatorischen Aspekten alle möglichen Quellen von Erhebungsfehlern vorausschauend zu untersuchen und – gestützt auf empirische Erfahrungen sowie theoretische Kenntnisse – zu berücksichtigen.

In den folgenden Abschnitten 1 und 2 werden zunächst die einzelnen Fehlerquellen betrachtet. In Abschnitt 3 wird dann abschließend ein Konzept für die Planung von Stichprobenerhebungen vorgestellt.

## 1 Stichprobenfehler

### *1.1 Auswahl der Stichprobe*

Nur Stichproben, die nach dem Zufallsprinzip gezogen sind, bieten den Vorteil, daß die Größenordnung der Stichprobenfehler objektiv aufgrund der Daten geschätzt

werden kann, die mit der Stichprobe gewonnen worden sind. Andere Auswahlverfahren erlauben keine derartigen, theoretisch abgesicherten Aussagen über die Qualität der Stichprobenergebnisse und sind deshalb auf rein subjektive Urteile über die Genauigkeit der Ergebnisse angewiesen. Für Stichprobenerhebungen, die nicht auf eine vorläufige Exploration, sondern auf wissenschaftlich abgesicherte Ergebnisse ausgerichtet sind, sollte deshalb als Auswahlverfahren nur die Zufallsauswahl verwendet werden.

Das gelegentlich vorgetragene Argument, für das geplante wissenschaftliche Projekt sei eine Zufallsauswahl praktisch nicht realisierbar, beruht meist darauf, daß die vielfältigen Gestaltungsmöglichkeiten der Zufallsauswahl verkannt werden. Die folgenden Unterabschnitte skizzieren die Flexibilität des Auswahlverfahrens; detaillierte Darstellungen geben die in Abschnitt 4 genannten Werke.

### Festlegung der Auswahleinheiten

Die Erhebungseinheit (*reporting unit*) ist definiert als die Einheit, die bei der Erhebung herangezogen wird; sie ist in der Regel eindeutig durch das sachliche Ziel der Erhebung determiniert. Das gilt jedoch nicht für die Auswahleinheiten (*sampling units*), d.h. die Einheiten, die einem Auswahlvorgang zugrunde liegen und von denen jede ausgewählt werden könnte. Diese Hilfseinheiten können nach arbeitstechnischen Gesichtspunkten festgelegt werden, weil sie nur den Zugang zu den Erhebungseinheiten schaffen sollen, die in die Erhebung einzubeziehen sind (z.B. sind im Mikrozensus Personen die Zähleinheiten, Haushalte die Erhebungseinheiten, dagegen werden als Auswahleinheiten regional abgegrenzte Klumpen von Häusern bzw. Wohnungen genutzt). Die Menge der Auswahleinheiten, aus denen eine Auswahl getroffen wird, heißt Auswahlgesamtheit.

Der Fall, in dem Erhebungseinheit und Auswahleinheit identisch sind, ist zwar theoretisch besonders einfach, setzt aber voraus, daß ein Erhebungsrahmen existiert, der alle Einheiten enthält und für die Zwecke der Auswahl verfügbar ist. Ein aktueller und vollständiger Rahmen steht aber meist nicht zur Verfügung und wäre überdies nicht praktikabel, weil die Auswahlprozedur einen viel zu großen Zeit- und Geldaufwand erfordern würde.

Diese technischen Probleme sind vermeidbar, wenn Auswahleinheiten verwendet werden, die aus mehreren Erhebungseinheiten bestehen. Bei der Klumpenauswahl (*cluster sampling*) werden Auswahleinheiten verwendet, die jeweils aus einer – ggf. künstlich abgegrenzten – Menge von Erhebungseinheiten, den Klumpen, zusammengesetzt sind. Aus der Menge dieser Klumpen wird eine vorgegebene Zahl von Klumpen nach dem Zufallsprinzip gezogen. Die so ausgewählten Klumpen werden mit allen darin enthaltenen Erhebungseinheiten in die Stichprobe genommen. Falls die zu Klumpen vereinigten Erhebungseinheiten untereinander vergleichsweise ähnlich sind, führt die Klumpenauswahl zu größeren Stichprobenfehlern als eine Zufallsstichprobe, für die die gleiche Zahl von Erhebungseinheiten unmittelbar ausgewählt wird. Diese Vergrößerung der Stichprobenfehler wird Klumpeneffekt genannt.

Bei der mehrstufigen Auswahl (*multistage sampling*) werden mehrere Auswahlprozeduren in Serie hintereinander geschaltet: Für die Auswahl erster Stufe werden – ebenso wie bei der Klumpenauswahl – Erhebungseinheiten zu Gruppen zusammengefaßt. Im Gegensatz zur Klumpenauswahl wird jedoch in den ausgewählten Einhei-

ten erster Stufe eine Unterauswahl durchgeführt, bei der Auswahleinheiten zweiter Stufe gezogen werden. In vielen Anwendungsfällen ist es möglich, alle Erhebungseinheiten in die Stichprobe einzubeziehen, die zu den ausgewählten Einheiten zweiter Stufe gehören, d.h. auf weitere Unterauswahlen zu verzichten (zweistufige Auswahl). Falls es die Sachlage erfordert, können jedoch im Anschluß an die Auswahl auf der zweiten Stufe weitere Unterauswahlen vorgesehen werden. Die Auswahleinheit einer Stufe besteht jeweils aus einer Gruppe von Auswahleinheiten der folgenden Stufe. Die Auswahl auf der $k$-ten Stufe ist jeweils eine Unterauswahl aus den Einheiten, die auf der $(k-1)$-ten Stufe ausgewählt worden sind.

Oft ist die mehrstufige Auswahl die einzige Methode, eine Stichprobe nach dem Zufallsprinzip zu ziehen. Die Ausgestaltung der mehrstufigen Auswahl hängt entscheidend von der Art der Unterlagen ab, die als Auswahlgrundlage genutzt werden können. Bei der Planung ist der Stufungseffekt zu beachten, d.h. die Vergrößerung der Stichprobenfehler der Ergebnisse aus einer mehrstufig gezogenen Stichprobe gegenüber den entsprechenden Fehlern einer hypothetischen Zufallsstichprobe, für die die gleiche Zahl von Erhebungseinheiten unmittelbar gezogen wird.

Die mehrphasige Auswahl (*multi-phase sampling*) schaltet ebenfalls mehrere Auswahlvorgänge hintereinander, verwendet dafür aber – im Gegensatz zur mehrstufigen Auswahl – stets die gleiche Auswahleinheit. In der ersten Phase wird eine Stichprobe gezogen; aus den dabei gezogenen Einheiten wird dann in der zweiten Phase eine Unterstichprobe ausgewählt. Dieses Verfahren kann ggf. mehrfach angewendet werden. Eine zweiphasige Auswahl kann z.B. dafür eingesetzt werden, Fehler infolge von Antwortausfällen zu quantifizieren: Aus den Erhebungseinheiten, die in der ersten Phase ausgewählt wurden, aber nicht antwortbereit sind, wird in der zweiten Phase eine Unterstichprobe gezogen und mit allen zu Gebote stehenden Mitteln versucht, Informationen über die Einheiten in dieser Unterstichprobe zu bekommen.

## Auswahltechniken

Für das Ziehen von $n$ Auswahleinheiten aus einer Auswahlgesamtheit von insgesamt $N$ Einheiten können mehrere Techniken verwendet werden. Sie stimmen darin überein, daß sie den gleichen Auswahlsatz $f = n/N$ realisieren, unterscheiden sich dagegen wesentlich in ihrer Qualität.

Eine echte Zufallsauswahl (*random sampling*) kann leicht verwirklicht werden, falls die Auswahlgesamtheit in einer maschinenlesbaren Form dokumentiert (oder verhältnismäßig klein) ist: Als Hilfsmittel dienen Zufallszahlen, die in tabellierter Form vorliegen oder mit einem Programm generiert werden. Falls eine Auswahl mit Zurücklegen vorgesehen ist, genügt es, $n$ Zufallszahlen im Bereich $(1, 2, ..., N)$ zu erzeugen. Für eine Auswahl ohne Zurücklegen werden alle mehrfach auftretenden Zufallszahlen getilgt und das Verfahren so lange fortgesetzt, bis $n$ verschiedene Zufallszahlen in dem genannten Bereich vorliegen. Durch Vergleich der Ordnungsnummer der Einheiten mit den gezogenen Zufallszahlen werden diejenigen Einheiten bestimmt, die zur Stichprobe gehören. Diese Auswahltechnik sichert, daß jede mögliche Stichprobe von $n$ aus $N$ Einheiten mit der gleichen Wahrscheinlichkeit realisiert wird und somit jede Einheit die gleiche Chance hat, in die Stichprobe zu kommen.

Eine Modifikation des Verfahrens erlaubt auch die Auswahl von Einheiten mit verschiedenen, z.B. entsprechend der Größe abgestuften, Wahrscheinlichkeiten.

Als Ersatz für die echte Zufallsauswahl wird häufig die systematische Auswahl (*systematic sampling*) mit Zufallsstart empfohlen. Bei diesem Verfahren wird ein ganzzahliger Auswahlabstand $a$ aus dem Verhältnis $N/n$ und eine Zufallszahl $k$ im Bereich $(1, 2, ..., a)$ bestimmt. Die Stichprobe besteht dann aus allen Einheiten mit den Ordnungsnummern $k, k+a, k+2a, ..., k+(n-1)a$. Die systematische Auswahl sichert zwar, daß alle Einheiten die gleiche Chance haben, in die Stichprobe zu gelangen, ist aber selbst dann der echten Zufallsauswahl nicht völlig äquivalent, wenn die Einheiten vor der Auswahl in eine Zufallsordnung gebracht werden: Die Stichprobe ist als ein einziger, zufällig ausgewählter künstlicher Klumpen von n Einheiten zu deuten. Infolgedessen können die Stichprobenfehler aus den Daten, die mit einer systematischen Stichprobe gewonnen worden sind, nicht zuverlässig abgeschätzt werden. Falls die Einheiten in der Auswahlgesamtheit periodisch geordnet sind, kann die systematische Auswahl schwerwiegende Fehler induzieren. Diese Auswahltechnik sollte deshalb grundsätzlich vermieden werden.

Methodisch zuverlässig ist dagegen die von Deming (1956) vorgeschlagene Auswahltechnik, die ein guter Ersatz für eine echte Zufallsauswahl ist und Anordnungseffekte (s.u.) bei der Auswahl der Stichprobe zu nutzen erlaubt. Die Technik besteht darin, die $N$ Einheiten der Auswahlgesamtheit entsprechend der gegebenen oder geschaffenen Ordnung in Zonen einzuteilen, aus denen jeweils $m$ Einheiten unabhängig voneinander nach dem Zufallsprinzip gezogen werden. Damit gerade $n$ Einheiten ausgewählt werden, müssen die Zonen jeweils $m \cdot N/n$ Einheiten enthalten. Die Vorteile dieser Auswahltechnik liegen u.a. darin, daß es in einfacher Weise Auswahleinheiten mit jeweils der gleichen Zahl von Erhebungseinheiten schafft und die Stichprobe automatisch entsprechend der Auswahlgesamtheit verteilt, die Nachteile der systematischen Auswahl dagegen vermeidet. Zudem ermöglicht die Technik u.a. auch eine feine Schichtung der Gesamtheit. Eine Einführung in diese Technik gibt Zindler (1957).

### Schichtung der Auswahleinheiten

Die Schichtung (*stratification*) der Auswahlgesamtheit ist eine Modifikation der Auswahl, mit der Informationen, die für alle Einheiten der Auswahlgesamtheit verfügbar sind, genutzt werden können, um bei gleichem Stichprobenumfang Ergebnisse mit kleineren Stichprobenfehlern zu erhalten. Eine Schichtung im stichprobenmethodischen Sinne unterteilt die Auswahlgesamtheit aufgrund der Werte eines Merkmals (oder mehrerer Merkmale) in strukturähnliche Teilgesamtheiten (Schichten), aus denen dann jeweils gesondert Zufallsstichproben gezogen werden. Mit der geschichteten Auswahl wird das Ziehen von ungünstigen, d.h. von der Zielgesamtheit strukturell stark abweichenden Stichproben verhindert, ohne das Zufallsprinzip bei der Auswahl zu verletzen.

Die Verkleinerung der Stichprobenfehler durch Anwendung des Schichtungsprinzips wird Schichtungseffekt genannt. Die Größe dieses Effektes hängt u.a. von den für die Schichtung verwendeten Merkmalen, der Zahl der Schichten und ihrer Abgrenzung sowie davon ab, wie der vorgesehene Stichprobenumfang auf die Schichten aufgeteilt wird. Falls mehrere Merkmale für die Schichtung verfügbar sind, sollte dasjenige präferiert werden, das eine hohe Korrelation mit den wichtigsten Erhebungsmerkmalen erwarten läßt.

Bei der Aufteilung des Stichprobenumfanges (bzw. der Festlegung der – nicht unbedingt gleichen – Auswahlsätze je Schicht) sind zwei Aufgabenstellungen zu unterscheiden:

- Die Statistik soll in erster Präferenz möglichst genaue Ergebnisse für die Gesamtheit erbringen, die Genauigkeit der Ergebnisse für Teilmengen ist dagegen von geringer Bedeutung. In diesem Falle können die Zahl der Schichten und ihre Grenzen nach methodischen Gesichtpunkten gewählt werden; vor allem aber kann der Stichprobenumfang so auf die Schichten aufgeteilt werden, daß zumindest für ein Merkmal Ergebnisse mit optimaler Genauigkeit zu erwarten sind.

- Die Statistik soll nicht nur Ergebnisse für die Gesamtheit, sondern auch hinreichend genaue Ergebnisse für konkret definierte Teile der Gesamtheit, die Studienbereiche (*domains of study*), erbringen. Bei dieser Aufgabenstellung kann zwar die Zahl der Schichten in geringem Maße variiert werden, die Grenzen der Schichten sind dagegen weitgehend durch die Studienbereiche bestimmt. Im Gegensatz zum vorangehenden Fall darf die Aufteilung des Stichprobenumfanges nicht auf eine möglichst hohe Genauigkeit des Gesamtergebnisses ausgerichtet werden, weil diese Aufteilung regelmäßig für einige Studienbereiche zu Ergebnissen führt, die nicht mehr ausreichend genau sind. Andererseits würde aber die Forderung, für alle Studienbereiche etwa gleich genaue Ergebnisse anzustreben, die Genauigkeit der Ergebnisse für die Gesamtheit vergleichsweise stark mindern. Zu empfehlen ist ein Kompromiß, nach dem die Genauigkeit der Ergebnisse für die Studienbereiche entsprechend ihrer sachlichen Bedeutung abgestuft werden.

### Anordnung der Auswahleinheiten

Die Stichprobenfehler können auch dadurch vermindert werden, daß die Einheiten der Auswahlgesamtheit nach den Werten eines Merkmals oder mehrerer Merkmale geordnet werden und anschließend eine Auswahl nach dem von Deming entwickelten Verfahren durchgeführt wird. Die Anordnung der Auswahleinheiten hat gegenüber der Schichtung den Nachteil, daß keine unterschiedlichen Auswahlsätze verwendet werden können, bietet andererseits aber den Vorteil, daß die Auswahl recht einfach ist und keine Materialgruppen gebildet werden, die bei der Auswertung der Daten getrennt zu halten sind. Der Anordnungseffekt, d.h. die Verminderung der Stichprobenfehler durch diese Modifikation der Auswahlprozedur, ist etwa gleich dem Schichtungseffekt, der mit einer vergleichbaren Schichtung erreicht werden kann, falls einheitliche Auswahlsätze verwendet werden.

### Kombinationen von Auswahlverfahren

Die hier vorgestellten Auswahlverfahren schließen einander nicht aus, sondern können weitgehend miteinander kombiniert werden. Zum Beispiel ist es häufig vorteilhaft, die Schichtung mit der Anordnung in der Weise zu kombinieren, daß die Auswahlgesamtheit zunächst nach einem Merkmal in Schichten unterteilt wird, die Auswahleinheiten dann je Schicht gesondert nach einem anderen Merkmal – etwa der

Regionalzugehörigkeit – geordnet und die Stichproben aus den Schichten nach dem Verfahren von Deming gezogen werden. Durch geschickte Kombination der Klumpenauswahl bzw. der mehrstufigen Auswahl mit der Schichtung und Anordnung ist es möglich, die Vergrößerung der Stichprobenfehler infolge von Stufungs- bzw. Klumpeneffekten zumindest teilweise durch Schichtungs- und Anordnungseffekte zu kompensieren.

## *1.2 Hochrechnung auf die Gesamtheit*

Die Auswahl der Stichprobe erzeugt ein verkleinertes Abbild der Auswahlgesamtheit. Dieses Bild muß bei der Aufbereitung der Daten wieder auf volle Größe gebracht werden. Die Umkehr aller Verkleinerungsprozeduren durch die Auswahl heißt Hochrechnung der Stichprobe auf die Gesamtheit. Dabei sind alle Besonderheiten der Auswahl, wie z.B. Mehrstufigkeit und Differenzierung der Auswahlsätze, genau, nur in umgekehrter Richtung, zu wiederholen.

Auch dann, wenn keine Totalwerte, sondern Anteilswerte oder Durchschnittswerte für die Gesamtheit anzugeben sind, ist in den meisten Fällen eine Umrechnung der Werte aus der Stichprobe erforderlich, z.B. um unterschiedliche Auswahlsätze oder differierende Auswahlwahrscheinlichkeiten zu berücksichtigen.

Die Hochrechnung und Umrechnung sollte nicht mit einer Gewichtung der Daten verwechselt werden, die verwendet wird, um Unterschiede in den Antwortquoten rein mechanisch auszugleichen, falls Informationen über die Struktur der Masse der nicht Antwortenden fehlen. Dieses Vorgehen ist äußerst problematisch, weil es vermeidbare Unkenntnis durch die – regelmäßig falsche – Unterstellung ersetzt, die Antwortbereiten würden die Nicht-Antwortbereiten repräsentieren. Im Gegensatz dazu beruhen Hochrechnung und Umrechnung nicht auf ungeprüften Annahmen, sondern auf Fakten, die durch die Auswahlprozedur determiniert sind.

### Freie Hochrechnung

Das einfachste Hochrechnungsverfahren ist die freie Hochrechnung, bei der alle bei einer Erhebungseinheit festgestellten Werte mit dem Kehrwert der Wahrscheinlichkeit multipliziert werden, mit der diese Einheit in die Stichprobe eingeschlossen worden ist. Bei diesem Verfahren werden also keine Informationen außer der Auswahlwahrscheinlichkeit für die Hochrechnung verwendet.

### Gebundene Hochrechnung

Im Gegensatz zur freien Hochrechnung werden bei der gebundenen Hochrechnung zusätzliche Angaben über die Gesamtheit benötigt, mit deren Hilfe die Genauigkeit der Stichprobenergebnisse verbessert werden kann. Während die zusätzlichen Informationen bei der Schichtung und Anordnung genutzt werden, um das Ziehen von ungünstigen Stichproben einzuschränken, werden sie bei der gebundenen Hochrechnung dafür eingesetzt, die Ergebnisse der vorliegenden Stichprobe zu justieren. Daraus folgt, daß die gebundene Hochrechnung z.B. mit der Schichtung kombiniert werden kann, sofern Informationen über mindestens zwei verschiedene Merkmale

verfügbar sind (erfahrungsgemäß lohnt es sich kaum, ein bereits für die Schichtung genutztes Merkmal nochmals für die Hochrechnung heranzuziehen).

Bei der gebundenen Hochrechnung werden stets zwei Merkmale betrachtet: Das Untersuchungsmerkmal, für das ein Totalwert der Gesamtheit aufgrund der Werte aus der Stichprobe ermittelt werden soll, und ein Basismerkmal, für das der Totalwert aus einer anderen Statistik bekannt ist. Für beide Merkmale ist die Summe der Werte in der Stichprobe dem Zufall unterworfen. Sofern jedoch die beiden Merkmale eng miteinander korreliert sind, ist zu erwarten, daß die Summe des Erhebungsmerkmals und die Summe des Basismerkmals in der Stichprobe beide zu groß oder beide zu klein sind. Dementsprechend hat z.B. der Quotient der Summen von Erhebungs- und Basismerkmal bei hoher Korrelation relativ kleinere Zufallsfehler als die Summen selbst.

Bei der besonders häufig angewandten Verhältnisschätzung wird der Quotient mit dem Totalwert des Basismerkmals multipliziert; wegen anderer Formen der gebundenen Hochrechnung wird auf die in Abschnitt 4 angegebene Fachliteratur verwiesen. Allgemein gilt, daß die gebundene Hochrechnung mehr Arbeitsaufwand erfordert, dafür aber auch genauere Ergebnisse liefern kann. Sofern das Untersuchungsmerkmal nicht eng mit dem Basismerkmal korreliert ist, lohnt sich der Mehraufwand nicht; bei schwacher Korrelation kann die gebundene Hochrechnung sogar ungenauer sein als die freie Hochrechnung. Es muß also für jedes Erhebungsmerkmal gesondert geprüft werden, ob eine gebundene Hochrechnung überhaupt einen wesentlichen Vorteil bietet und ggf. welches Verfahren zweckmäßig ist.

## *1.3 Abschätzung der Stichprobenfehler*

Die Ergebnisse einer Erhebung können nur dann objektiv beurteilt werden, wenn ihre Fehler zumindest näherungsweise bekannt sind. Für Ergebnisse von Stichprobenerhebungen, die auf Zufallsstichproben aufbauen, eröffnet die Stichprobentheorie einen Weg, wenigstens die Stichprobenfehler ihrer Größenordnung nach abzuschätzen. Dafür wird eine Maßzahl verwendet, der sog. Standardfehler (*standard error*) des Ergebnisses. Für diese Maßzahl gilt, daß der durch die Zufallsauswahl der Stichprobe bedingte Stichprobenfehler

in 683 von 1000 Fällen kleiner als der einfache Standardfehler,
in 955 von 1000 Fällen kleiner als der zweifache Standardfehler und
in 997 von 1000 Fällen kleiner als der dreifache Standardfehler ist.

Falls Erhebungseinheiten zugleich auch Auswahleinheiten sind und eine Stichprobe von $n$ Einheiten aus einer Auswahlgesamtheit von $N$ Einheiten nach dem Zufallsprinzip ohne Schichtung gezogen wird, kann der Standardfehler des frei hochgerechneten Totalwertes für das Merkmal $X$ in der Gesamtheit nach der Formel

$$\text{Standardfehler des frei hochgerechneten Totalwertes} = N \cdot \sqrt{\frac{s_X^2}{n}} \qquad (1)$$

geschätzt werden, falls $n$ klein gegen $N$ ist. Das Symbol $s_X^2$ steht für die erwartungstreu geschätzte Varianz der Einzelwerte $x_i$ des Merkmals $X$ um ihren Mittelwert $\bar{x}$

in der Stichprobe:

$$s_X^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \Big/ (n-1) \quad . \tag{2}$$

Die Standardfehler für Stichprobenpläne, die andere Auswahl- und Hochrechnungsverfahren vorsehen, können ebenfalls, allerdings nach zum Teil wesentlich komplexeren Formeln geschätzt werden, die der Fachliteratur zu entnehmen sind. Für eine pauschale Abschätzung genügt es meist, die fehlervergrößernden Einflüsse einer Klumpenauswahl bzw. mehrstufigen Auswahl durch Zuschlagsfaktoren $F_1 > 1$ und die fehlermindernden Effekte der Schichtung, Anordnung und gebundenen Hochrechnung durch Abschlagsfaktoren $F_2 < 1$ zu berücksichtigen. Durch Zusammenfassen der Faktoren $F = F_1 \cdot F_2$ erhält man für den Standardfehler des geschätzten Totalwertes $\widehat{X}$ in der Gesamtheit die Näherungsformel:

$$s_{\widehat{X}} \approx F \cdot N \cdot \sqrt{s_X^2 \cdot \frac{1}{n}} \tag{3}$$

Darin sollte der Faktor $F$ sicherheitshalber nicht kleiner als 1 gesetzt werden.

Die Größenordnung der Stichprobenfehler eines Ergebnisses läßt sich sachlich leichter beurteilen, wenn der Standardfehler als Bruchteil des Ergebniswertes ausgedrückt wird. Für den so definierten relativen Standardfehler (*relative standard error*) des Totalwertes in der Gesamtheit ergibt sich die Pauschalformel

$$\text{relativer Standardfehler des geschätzten Totalwertes} \approx F \cdot \sqrt{\frac{s_X^2}{(\bar{x})^2} \cdot \frac{1}{n}} \tag{4}$$

Wegen der Relativierung gilt diese Formel nicht nur für geschätzte Totalwerte, sondern auch für geschätzte Mittelwerte und geschätzte Anteilswerte in der Gesamtheit. In diesem Spezialfall ist $\bar{x}$ durch den Anteilswert $p_X$ und die Varianz $s_X^2$ durch $p_X(1-p_X)$ zu ersetzen.

Die Schätzung der Standardfehler von Totalwerten in Teilgesamtheiten erfordert dagegen, die Formeln (1), (3) und (4) jeweils um den Faktor $\sqrt{1 + (1-p_X)/(p_X \cdot n)}$ zu erweitern, wobei $p_X$ den Anteil der Teilgesamtheit an der Gesamtheit bezeichnet.

## 2 Systematische Fehler

Eine Stichprobenerhebung, die den Normen der Wissenschaft genügen will, muß sowohl die Stichprobenfehler als auch die systematischen Fehler der Erhebung angemessen beachten.

Die Berücksichtigung von systematischen Fehlern wird dadurch entscheidend erschwert, daß es für diese Fehler – im Gegensatz zu Stichprobenfehlern bei Zufallsauswahl – keine allgemeingültige Theorie, sondern nur eine Vielzahl von Hinweisen gibt, wie sie entstehen können und in welcher Weise ihre Wirkung beurteilt werden kann. Für eine objektive Abschätzung der systematischen Fehler reichen die Informationen aus der Stichprobe nicht aus, vielmehr werden dafür stets zusätzliche Unterlagen benötigt, deren Beschaffung meist viel Arbeit und hohen finanziellen Aufwand erfordert.

Bei vielen empirischen Untersuchungen wird leider noch immer die Gefährdung der wissenschaftlichen Erkenntnis durch systematische Fehler entweder gar nicht wahrgenommen, oder das Problem wird scheinbar dadurch gelöst, daß die Existenz von systematischen Fehlern zwar anerkannt, aber ohne weitere Prüfung des Sachverhaltes als irrelevant für die Ergebnisse dekretiert wird.

Methoden für die Messung und Reduktion von systematischen Fehlern werden in der Monographie von Lessler und Kalsbeek (1992) dargestellt; hier können nur einige wichtige Hinweise gegeben werden.

## *2.1 Fehler im Erhebungsrahmen*

Abweichungen zwischen der Erhebungsgesamtheit, die durch den verfügbaren Erhebungsrahmen definiert ist, und der intendierten Zielgesamtheit können Fehler in den Ergebnissen hervorrufen. Folgende Arten von Abweichungen sind zu unterscheiden:

1. Abweichungen im Umfang
   (a) Zur Zielgesamtheit gehörende Einheiten sind nicht in der Erhebungsgesamtheit enthalten (Untererfassung);
   (b) zur Zielgesamtheit gehörende Einheiten sind mehrfach in der Erhebungsgesamtheit enthalten (Doppelerfassung);
   (c) die Erhebungsgesamtheit enthält Einheiten, die nicht zur Zielgesamtheit gehören (Übererfassung).
2. Abweichungen im Inhalt
   (a) Die Erhebungsgrundlage enthält unkorrekte Angaben über planungsrelevante Hilfsmerkmale (z.B. das Alter einer Person);
   (b) in der Erhebungsgrundlage fehlen Angaben über planungsrelevante Hilfsmerkmale.

Die beiden Gruppen von Abweichungen unterscheiden sich wesentlich in ihrer Wirkung: Abweichungen im Umfang (Gruppe 1) führen sowohl bei Totalerhebungen als auch bei Stichprobenerhebungen zu systematischen Fehlern; dagegen vergrößern Abweichungen im Inhalt (Gruppe 2) bei Stichprobenerhebungen deren Stichprobenfehler, beeinflussen aber Totalerhebungen nicht oder nur schwach.

Besonders gravierende Abweichungen im Umfang entstehen, wenn eine Erhebungsgesamtheit – vielleicht aus triftigen Gründen – von vornherein auf eine geringe Zahl von Regionen (z.B. drei Großstädte und ihr Umfeld) beschränkt wird, die Ergebnisse der Erhebung dagegen für das ganze Land gelten sollen, d.h. für eine wesentlich größere Zielgesamtheit. Eine formale Möglichkeit, die Abweichung zwischen Zielgesamtheit und Erhebungsgesamtheit zu beseitigen, besteht darin, die Abweichung mitsamt den daraus resultierenden systematischen Fehlern wegzudefinieren, indem als Zielgesamtheit gerade die tatsächlich erfaßte Erhebungsgesamtheit vorgesehen wird. Dieser Ausweg ist aber nur dann vertretbar, wenn die Ergebnisse der Erhebung ausdrücklich als Aussagen über die neu definierte Zielgesamtheit dargestellt werden. Ein Transponieren der Ergebnisse auf die ursprünglich vorgesehene Zielgesamtheit ist wissenschaftlich nicht haltbar.

## *2.2 Fehler infolge von Antwortausfällen*

Bei Erhebungen, die auf eine freiwillige Beteiligung angewiesen sind, kann nicht damit gerechnet werden, daß es gelingt, für alle – oder zumindest die meisten – ausgewählten Erhebungseinheiten brauchbare Angaben einzuholen. In der Regel kann nur ein mehr oder minder kleiner Teil der befragten Personen dazu motiviert werden, die gestellten Fragen zu beantworten (bzw. die vorgesehene Testprozedur über sich ergehen zu lassen).

Falls die Gruppe der Antwortbereiten in ihrer Struktur mit der Gruppe der Antwortverweigerer übereinstimmen würde, hätte die Antwortverweigerung nur den Effekt, daß zwar die Stichprobenfehler infolge des Schrumpfens der Stichprobe vergrößert, aber keine systematischen Fehler auftreten würden. Erfahrungsgemäß ist diese Strukturgleichheit nicht gegeben, vielmehr verweigern z.B. zu Randgruppen gehörende Personen besonders häufig die Beteiligung an der Erhebung. Die Tatsache, daß die daraus resultierenden systematischen Fehler unsichtbar sind, berechtigt nicht zu dem Schluß, sie seien unbedeutend und dürften deshalb vernachlässigt werden: Je kleiner die Antwortquote ist und je größer die Strukturunterschiede sind, desto stärker können Fehler infolge von Antwortausfällen die Ergebnisse verfälschen.

Diese Fehler lassen sich nicht vermeiden, jedoch sollte ihre Größenordnung ermittelt und bei der Interpretation der Ergebnisse berücksichtigt werden. Es ist deshalb dringend zu empfehlen, bei jeder auf Freiwilligkeit beruhenden Stichprobenerhebung von vornherein eine – vergleichsweise kleine – Unterstichprobe einzuplanen und mit allen Mitteln zu versuchen, von möglichst vielen Einheiten in dieser Unterstichprobe die benötigten Informationen zu bekommen. Nur auf diesem Wege können Verzerrungen durch die selektiv wirksamen Antwortausfälle in wissenschaftlich einwandfreier Weise berücksichtigt werden.

## *2.3 Meßfehler*

Bei der Erhebung von Daten und bei ihrer Aufbereitung entstehen – wie bei jeder von Menschen ausgeübten Tätigkeit – Fehler. Ein Teil dieser Fehler ist von irregulärer Natur, tendiert also dazu, sich gegenseitig teilweise zu kompensieren. Es ist aber nicht auszuschließen, daß Mißverständnisse oder Nachlässigkeit einseitige Meßfehler bewirken, die zu systematischen Fehlern der Ergebnisse führen.

Meßfehler können nur durch Kontrollen der Erhebung und der Aufbereitung auf der Basis von Stichproben erfaßt werden. Dabei sind zwei Aufgabenstellungen zu unterscheiden:

- Deskriptive Kontrollen sind darauf ausgerichtet, die Größenordnung der Meßfehler zu quantifizieren, mit dem Ziel, die Güte der Ergebnisse zu beurteilen und entsprechend zu kommentieren.

- Operative Kontrollen haben dagegen die Aufgabe, systematische Fehler bei der Erhebung und Aufbereitung festzustellen, um die statistischen Ergebnisse zu verbessern.

Die Kontrollen liefern nicht nur Erkenntnisse über die Qualität der Statistik, sondern auch Hinweise auf ihre Fehlerquellen. Damit ist ein Weg eröffnet, vorbeugende

Maßnahmen zur Einschränkung von Fehlerursachen bei der Planung von künftigen Erhebungen zu treffen.

Bei der Erhebung sollte grundsätzlich die Reliabilität des Meßverfahrens und die Stringenz seiner Anwendung anhand einer Stichprobe untersucht werden, auch wenn diese Prüfung vergleichsweise hohen Aufwand erfordert. Falls für eine Erhebung Interviewer eingesetzt werden, ist eine Untersuchung angezeigt, ob Interviewer systematisch differierende Einflüsse ausüben. Für Kontrollen der Aufbereitung sind Methoden der statistischen Qualitätssicherung (vgl. Rinne & Mittag, 1991) angezeigt.

## 3 Integrierte Planung

Der Stichprobenplan für eine Erhebung kann nur dann als methodologisch korrekt gelten, falls er

1. alle wesentlichen Fehlerquellen bei der Erhebung der Daten und ihrer Aufbereitung aufgrund theoretischer Kenntnisse und empirischer Erfahrungen bei der Planung berücksichtigt;

2. darauf abgestellt ist, bei vorgegebenem Finanzvolumen die zu erwartenden Erhebungsfehler für spezielle Merkmale in konkret definierten Teilen der Gesamtheit zu minimieren (bzw. solche Ergebnisse mit vorgeschriebener Genauigkeit unter Einsatz von möglichst kleinen finanziellen Mitteln zu realisieren);

3. die erforderlichen Vorkehrungen zur Abschätzung der Erhebungsfehler bei der Auswertung der Daten trifft und

4. in angemessenem Detail angibt, wie diese Informationen bei der Analyse der Daten und bei der Interpretation der Ergebnisse zu berücksichtigen sind.

Die unter Ziffer 2 genannte Aufgabe setzt Kenntnisse über die Größenordnung der Quellen von Erhebungsfehlern voraus, die mit Hilfe einer Probeerhebung beschafft werden müssen, sofern sie aus vorliegenden Erfahrungen nicht mit hinreichender Sicherheit erschlossen werden können. Planungsunterlagen werden sowohl über die Stichprobenfehler als auch über die systematischen Fehler der relevanten Ergebnisse benötigt, weil der Nutzen eines Stichprobenergebnisses nicht allein von dem (vergleichsweise einfach abschätzbaren) Stichprobenfehler, sondern von dem Aggregat aus dem Stichprobenfehler und der Summe aller systematischen Fehler abhängt.

Der mittlere quadratische Gesamtfehler $g_{\widehat{X}}$ des geschätzten Totalwertes $\widehat{X}$ kann nach der Formel

$$g_{\widehat{X}}^2 = s_{\widehat{X}}^2 + \left(\sum b_r\right)^2 \tag{5}$$

aus dem Standardfehler $s_{\widehat{X}}$ (vgl. Formel 3) und der geschätzten Summe $\sum b_r$ aller systematischen Fehler $b_r$ berechnet werden. Diese Maßzahl ermöglicht Wahrscheinlichkeitsaussagen über den Bereich des gesamten Erhebungsfehlers eines Ergebnisses in der gleichen Weise, in der Standardfehler Aussagen über die Größenordnung von Stichprobenfehlern erlauben.

Die integrierte Planung der Stichprobenerhebung (vgl. Kish, 1965) hat von dieser Größe auszugehen. Dabei ist zu berücksichtigen, daß die systematischen Fehler $b_r$ nicht wesentlich von der Zahl $n$ der ausgewählten Erhebungseinheiten abhängen, sich also anders verhalten als die Standardfehler $s_{\widehat{X}}$ des geschätzten Totalwertes $\widehat{X}$, für die nach Formel (3)

$$s_{\widehat{X}} = \frac{A}{\sqrt{n}} \tag{6}$$

gilt, wenn $A = F \cdot N \cdot \sqrt{s_X^2}$ gesetzt wird.

Wenn $C_r$ die Kosten bezeichnet, die für die Reduktion des $r$-ten systematischen Fehlers auf das Niveau $b_r$ entstehen, und $c$ der Kostensatz je Erhebungseinheit in der Stichprobe ist, folgt für die variablen Kosten $C$ der Erhebung das Modell

$$C = c \cdot n + \sum C_r \tag{7}$$

Die bei der Planung zu lösende Aufgabe lautet also, bei vorgegebenen Kosten $C$ ein Bündel von Techniken zu finden, die zu einem Minimum der Funktion

$$g_{\widehat{X}}^2 = \frac{A^2}{n} + \left(\sum b_r\right)^2 \tag{8}$$

führen.

## 4 Weiterführende Literatur

Einen umfassenden Überblick über Theorie und Anwendungen des Stichprobenverfahrens bietet das zweibändige Werk von Hansen, Hurwitz und Madow (1953a, 1953b). Als Einstieg eignet sich das Lehrbuch von Cochran (1963), das die Grundzüge des Stichprobenverfahrens knapp und klar darstellt. Etwas anspruchsvoller ist die von Stenger (1986) – in deutscher Sprache – gegebene Einführung, die auch auf die wichtigsten theoretischen Fortschritte der letzten Jahrzehnte eingeht.

## Literaturverzeichnis

Cochran, W. G. (1963). *Sampling techniques* (2nd ed.). New York: Wiley.

Deming, W. E. (1956). On simplifications of sampling designs through replication with equal probabilities and without stages. *Journal of the American Statistical Association, 51*, 24–53.

Hansen, M. H., Hurwitz, W. N. & Madow, W. G. (1953a). *Sample survey methods and theory. Volume I: Methods and applications.* New York: Wiley.

Hansen, M. H., Hurwitz, W. N. & Madow, W. G. (1953b). *Sample survey methods and theory. Volume II: Theory.* New York: Wiley.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

Lessler, J. T. & Kalsbeek, W. D. (1992). *Nonsampling errors in surveys.* New York: Wiley.

Stenger, H. (1986). *Stichproben.* Heidelberg: Physica.

Rinne, H. & Mittag, H.-J. (1991). *Statistische Methoden der Qualitätssicherung* (2. Aufl.). München: Hanser.

Zindler, H.-J. (1957). Über einige Aspekte des Demingplanes. *Mitteilungsblatt für Mathematische Statistik, 9*, 55–72.