Zur Datenqualität primärstatistischer Erhebungen

Heinrich Strecker

Emer. o. Professor Dr. rer. nat. Heinrich Strecker, Universität Tübingen und
Honorarprofessor der Ludwig-Maximilians-Universität München, Rosenstr. 11,
D 82319 Starnberg bei München

1. Allgemeine Einführung und Begriffe

Die nachfolgenden Ausführungen sind eine Fortsetzung und Ergänzung früherer Abhandlungen (siehe Literaturverzeichnis). Angeregt durch einen Studienaufenthalt im US-Bureau of the Census, Washington, D.C., habe ich mich in den letzten Jahrzehnten mit der Methodik statistischer Erhebungen und deren Datenqualität befasst. Diese Fragen haben mich bereits während meiner Tätigkeit im ehemaligen Bayerischen Statistischen Landesamt beschäftigt. Viele neue Erkenntnisse habe ich später während meiner Berater-Tätigkeit in Belgien (Institut national de statistique) gewonnen und diese dort auch praktisch angewandt.

Im Zeitalter von zunehmender Computerisierung und Vernetzung von Daten nimmt der Anteil primärstatistischer Erhebungen am Gesamtprogramm eher ab. Dennoch muß die Qualität ihrer Ergebnisse nach wie vor berücksichtigt werden. Es ist sehr schwierig wegen der unterschiedlichen Standards statistischer Dienste, einheitliche Richtlinien für ein Qualitätskriterium auf internationaler Ebene zu vereinbaren. Deshalb haben in den letzten 20 Jahren internationale Institutionen wie Eurostat und Internationaler Währungsfond sowie nationale statistische Ämter, z.B. Statistics Canada, Statistics Sweden, Statistics UK und andere, große Anstrengungen unternommen, um einheitliche Kriterien zur Verbesserung der Datenqualität aufzustellen und zu erreichen (P. Biemer/L.Lyberg 2003). Dazu gehören Wichtigkeit (relevance), Genauigkeit (accuracy), Vergleichbarkeit (comparability), Kohärenz (coherence), Aktualität und Pünktlichkeit (timeliness and punctuality). Ziel war und ist, die internationale Vergleichbarkeit von Volkswirtschaften und Wirtschaftsräumen zu ermöglichen.

Was die Genauigkeit betrifft, so erkannte man in den 50er Jahren, dass es nicht ausreicht, allein auf die Effizienz der Schätzverfahren bei Stichproben zu achten, sondern dass auch der Nicht-Stichprobenfehler sowie die Variabilität der Antworten von Befragten berücksichtigt werden müssen.

In den 60er Jahren wurden deshalb entsprechende Modelle mit (Wiederholungs-)-Zählungen von den US-Statistikern, insbesondere M. Hansen, W. Hurwitz, M.

Bershad, L. Prizker, W.G. Cochran und B. Bailar, entwickelt, die jedoch auf dem Gebiet der Erhebungsstatistik in Deutschland noch wenig beachtet worden sind.

Die Durchführung einer Erhebung bedarf einer umfassenden Vorbereitung. Es wird zunächst von einer Zielgröße Z ausgegangen, für die Ergebnisse vom Konsumenten der Erhebung gewünscht werden. Z ist nicht immer eindeutig, sondern zwangsläufig häufig nur annähernd zu definieren. Man kann in einer Zählung auch nicht für alle Merkmale der Gesamtheit Ergebnisse ermitteln. Das wäre mit einem zu großen Arbeitsaufwand verbunden und sicherlich wären einige Daten für den Konsumenten der Statistik von geringem Interesse. Es muss eine erste Auswahl der Merkmale und damit eine Simplifizierung vorgenommen werden. Weiterhin ist die Dauer der Erfassung und Aufbereitung festzulegen sowie auf die Probleme bei der Feldarbeit einzugehen. Kostenrestriktionen spielen keine untergeordnete Rolle. Von großer Bedeutung sind vor allem die mit der Adäquation von gewünschten Zielgrößen zu einer operablen Erfassung der Merkmale verbundenen Arbeitsgänge. Unter der statistischen Adäquation wird der Prozess bezeichnet, in dem idealtypische oder andere theoretische Definitionen in statistisch operable Größen zwecks Datenerfassung umgewandelt werden. Dieser Prozess sowie der geplante zeitliche Ablauf der Erhebung und die Erhebungstechnik bilden aufeinander bezogene Elemente, die im so genannten ARBEITSSYSTEM G (Generalized conditions) festgelegt werden müssen:

a) Zielsetzung der Erhebung, Kosten- und Zeitplan
b) Definitionen von statistischen Einheiten, Merkmalen und Massen (Gesamtheiten)
c) Tabellenprogramm
d) Organisation der Feldarbeit
e) sofern die Erhebung als Stichprobe durchgeführt wird: Wahl des Stichprobenplans, des Auswahlrahmens, Punkt- und Intervallschätzung
f) Sammeln der Daten, operative und deskriptive Kontrollen
g) Verarbeitung der Daten
h) Prüfung der Zuverlässigkeit der Daten
i) Tabellierung
j) Veröffentlichung der Ergebnisse oder Weitergabe an Benutzer (Auftraggeber)

Das Arbeitssystem kann sehr unterschiedlich gestaltet sein, wie mit Selbstausfüllung der Fragebogen mit oder ohne Einfluss eines Zählers, Befragung durch Interviewer, Telefoninterviews usw., Durchführung als Vollerhebung oder als Stichprobe. Diese Varianten führen zu verschiedenen Ausformungen des Arbeitssystems G, G', G" etc. bei gleichen Zielen der Erhebung.

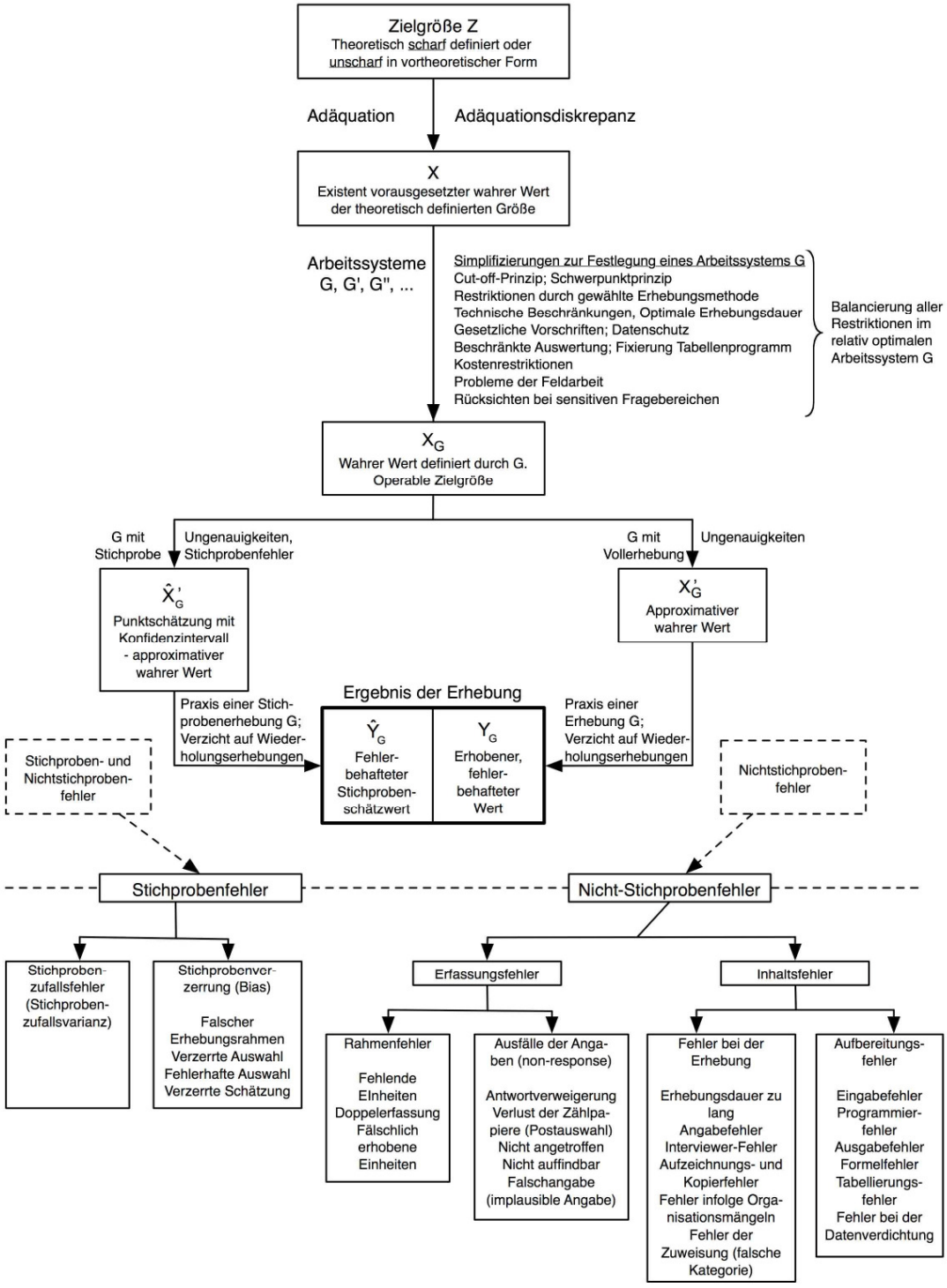
Die Ergebnisse einer Erhebung werden dann als wahre Werte definiert, wenn das festgelegte Arbeitssystem korrekt und sorgfältig realisiert wurde. Die Qualität statistischer Ergebnisse ist somit abhängig von der Durchführung des Arbeitssystems. Bei den Bemühungen, die wahren Werte X_G der Merkmale zu ermitteln, treten infolge menschlicher Unzulänglichkeit trotz aller Sorgfalt Ungenauigkeiten auf, d.h. bestenfalls kann der wahre Wert X_G nur als ein approximativer Wert X'_G erhoben werden. Bei allen Arbeitsgängen können diese Abweichungen vom Arbeitssystem auftreten, die zu verschiedenen Fehlern führen.

In der folgenden Übersicht sind die möglichen Fehlerarten, die die Ergebnisse der Erhebung Y_G bzw. $\hat{Y}_G$ bei Stichprobenerhebungen beeinflussen, aufgeführt (Übersicht 1).

In der Übersicht 2 sind nochmals die hier beim Arbeitssystem erörterten Begriffe und Definitionen in einer anderen Form dargestellt. Beide Schemata zusammen geben einen aufschlussreichen Überblick über die Aufgaben und Probleme, die sich bei einer primärstatistischen Erhebung stellen. In diesem Zusammenhang sei noch besonders auf die instruktive Übersicht über Fehler und Genauigkeitskontrollen in P. von der Lippe 1996, 42 und 43 hingewiesen.

Übersicht 1

Schema: Interaktion zwischen Begriffsbildung, Adäquation, Simplifizierung und Fehler bei primärstatistischen Erhebungen.

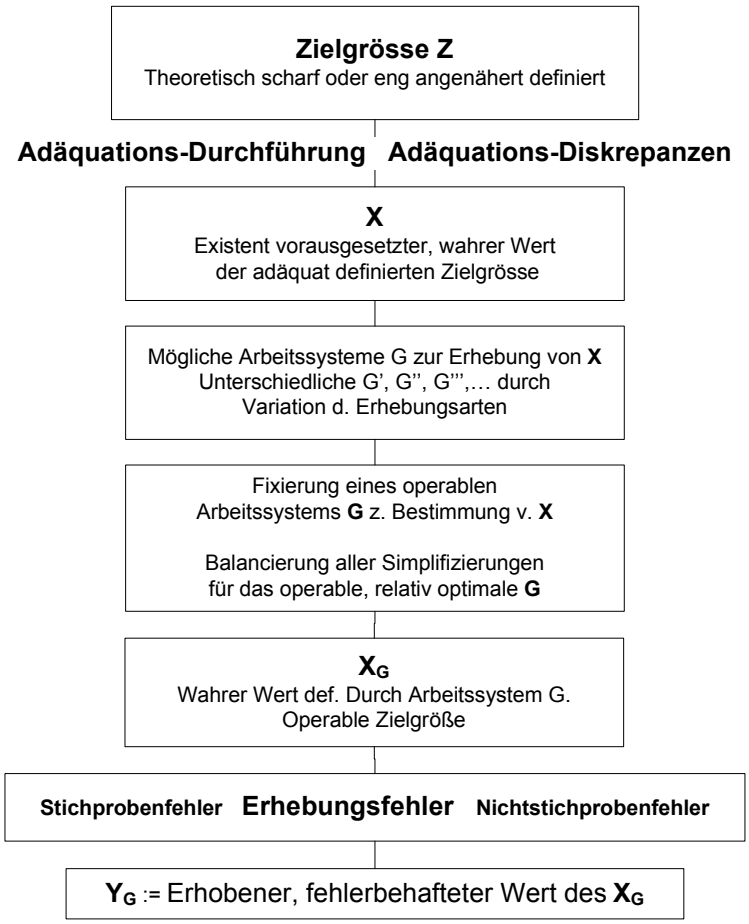


Bezogen auf eine Vollerhebung würde sich das Schema vereinfachen, es sei denn man deutet jede Vollerhebung als Stichprobe aus einer "Gesamtheit höherer Ordnung".

Eine gesonderte Übersicht hierzu soll hier nicht erstellt werden.

Übersicht 2

Schema zur Begriffsbildung, Adäquation, Vereinfachung und zu Fehlern bei primärstatistischen Erhebungen



Quelle: H.Strecker/R.Wiegert 2006, 314

2. Methoden der Qualitätsanalyse

Zur Untersuchung und Verbesserung der Datenqualität sind viele unterschiedliche Methoden entwickelt worden. Sie alle zu behandeln, soll nicht Aufgabe dieses Beitrags sein. Ein wichtiges Verfahren ist die Schätzung der Antwortvariabilität mit Hilfe von (Wiederholungs-)Zählungen. Wird eine Erhebung im nahen zeitlichen Abstand mindestens zweimal oder mehrmals nach dem gleichen Arbeitssystem durchgeführt, so zeigt sich, dass die Antworten der Befragten in der Regel variieren können. Diese Variabilitäten können zufallsbedingt oder in seltenen Fällen auch systematischer Natur sein.

Wenn bei Bestandsmassen die Werte des Merkmals zwischen den Stichtagen der Haupterhebung und den Wiederholungszählungen erfahrungsgemäß konstant bleiben, z.B. die Landwirtschaftliche Nutzfläche der Betriebe, ist es nicht notwendig, die Angaben für den Stichtag der Wiederholungszählung auf den Stichtag der Haupterhebung zu adjustieren. Man kann dann ohne weiteres die Variabilitäten der Antworten für das Merkmal entsprechend einem Fehlermodell schätzen. Wenn die Werte des Merkmals sich zwischen den Stichtagen ändern, dann muss eine Adjustierung der Angaben aus der Wiederholungszählung auf den Stichtag der Haupterhebung erfolgen (H.Strecker/R.Wiegert 1986, 99 - 130, P. von der Lippe 1996, 42-44).

Bei Daten für einen Referenz-Zeitraum von Bewegungsmassen können ebenfalls Antwortvariabilitäten auftreten, die dann wieder entsprechend einem Fehlermodell für denselben Zeitraum zu adjustieren sind.

Um für ein Fehlermodell, das als Grundlage für die Analyse und Schätzung der Antwortvariabilität dienen kann, die erforderlichen quantitativen Größen zu erhalten, bedarf es der Angaben aus mindestens zwei (Wiederholungs-)Zählungen und einer Kontrollerhebung. Die Kontrollerhebung ist eine Wiederholungszählung, die nach dem gleichen Arbeitssystem, jedoch mit allergrößter Sorgfalt, möglichst von besonders geschultem Personal und mit einer Überprüfung der Angaben vor Ort durchgeführt wird, wie z.B. in Belgien von Bediensteten des Institut national de statistique (Moniteure). In der Kontrollerhebung müssen, zumindest approximativ, die "wahren" Veränderungen der individuellen Merkmalswerte zwischen dem Zeitpunkt der Haupterhebung und dem Tag der Kontrollerhebung, d.h. die Zu- und Abgänge, erfasst werden. Dann erfolgt eine Rückrechnung auf die wahren Werte am Stichtag der Haupterhebung.

Die Differenz zwischen den ursprünglich erhobenen Werten und den zurückgerechneten Werten sind die individuellen approximativen Angabefehler in der Haupterhebung (H.Strecker/ R.Wiegert 1983, 90-101).

Das Fehlermodell ist auf dem quantitativen individuellen Angabewert y_{it} der Erhebungseinheit i und den (Wiederholungs-)Zählungen t (t = l, 2, 3, …, k, ...) aufgebaut. Eine Aggregation dieser individuellen Werte aus einer Vollerhebung über

alle Einheiten N ($\sum_{i=1}^{N} y_i$, wobei $y_i = E_t(y_{it})$ ist) liefert die Gesamtergebnisse Y_N für diese Zählung, bzw. bei einer Stichprobe über die Einheiten n ($\hat{Y}_n$). Der individuelle quantitative Angabewert y_{it} der Erhebungseinheiten i wird zunächst formal in seine Komponenten zerlegt: Wahrer Wert x_i, systematischer Fehler e_i und Zufallsfehler ε_{it}. Nachfolgend ist in einem Schema diese Zerlegung veranschaulicht.

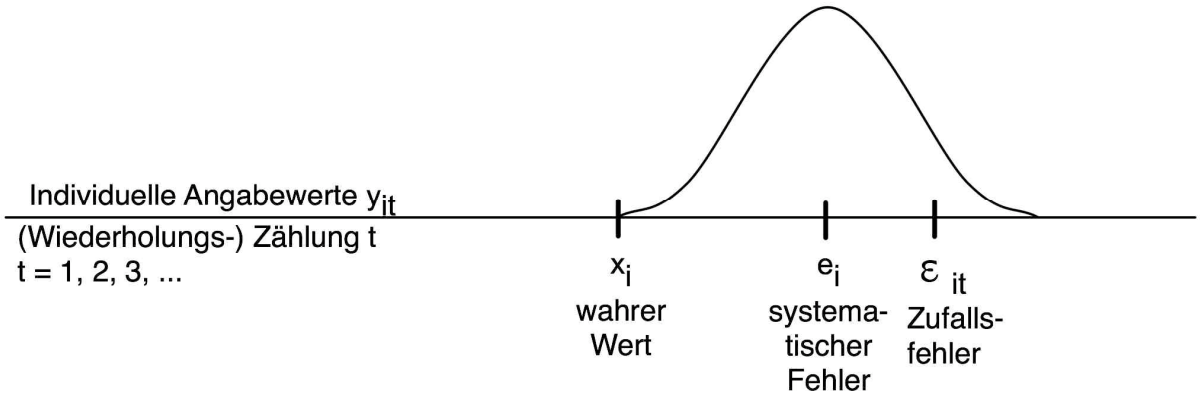
Schema

Zerlegung des individuellen quantitativen Angabewertes y_{it} der Erhebungseinheit i in seine Komponenten: wahrer Wert x_i, systematischer Fehler e_i und Zufallsfehler ε_{it}.

Lineares Fehlermodell

(Wiederholungs-)Zählung t

Individueller Angabewert $y_{it} = x_i + e_i + \varepsilon_{it}$, t = 1,2,3...

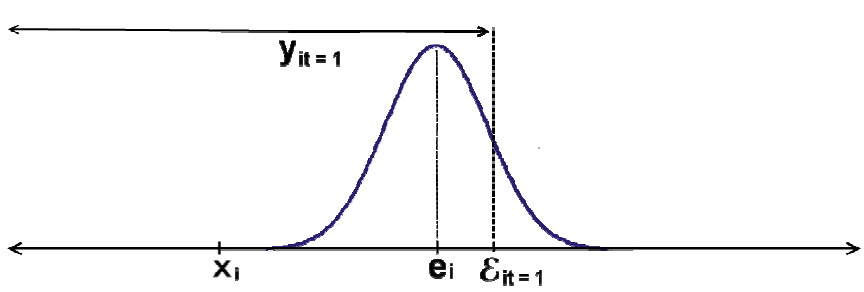


Für die Bestimmung und Schätzung der Fehlerkomponenten der wahren Werte und der Antwortvariabilitäten sind die individuellen Daten einer Haupterhebung (t = l), einer Wiederholungszählung (t = 2) und einer Kontrollerhebung (t = 0) notwendig.

Angabewert $y_{it=1} = x_i + e_i + \varepsilon_{it=1}$

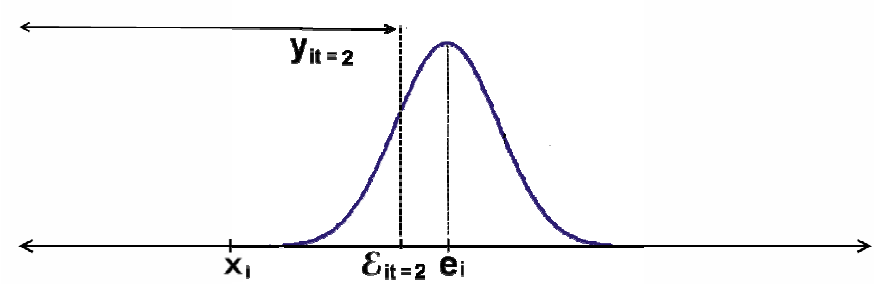
Haupter-

Hebung (t = 1)



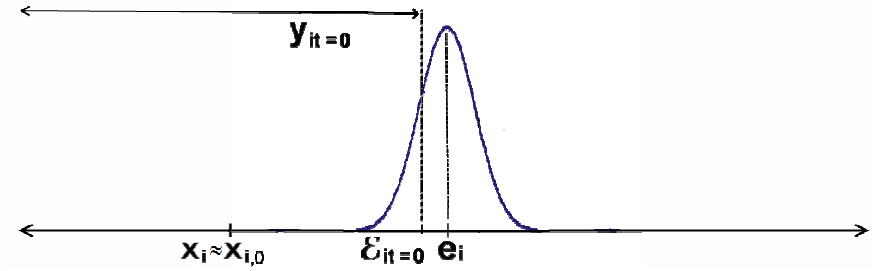
Angabewert $y_{it=2} = x_i + e_i + \varepsilon_{it=2}$

Wiederholungs-Zählung (t = 2)



Kontrollwert $y_{it=0} = x_{it=0} + e_{it=0} + \varepsilon_{it=0} \approx x_{i,0}$

Kontroll-Erhebung (t = 0)



"wahrer" Näherungswert $x_{i,0}$ laut Kontrollerhebung

Hinweis:

Bei Vorliegen individueller Angabewerte y_{it} aus mehreren (Wiederholungs-)-Zählungen ist das Schema entsprechend zu erweitern (k>2). Für qualitative individuelle Erhebungsmerkmale ist in Analogie das Zerlegungs-Schema zu modifizieren. Schätzwerte für die Gesamtheit erhält man durch Aggregation der individuellen Werte - bei einer Stichprobe $\sum_{i=1}^{n} y_{it}$, bei einer Vollerhebung $\sum_{i=1}^{N} y_{it}$. Ein Fehlermodell für qualitative Merkmale kann man analog ohne Schwierigkeiten konzipieren. In der Praxis handelt es sich jedoch in der Regel um die Kontrolle quantitativer Merkmale.

Es ist zu beachten, dass die wahren Werte x_i in allen Erhebungen stets dieselben sind. Die systematischen Fehler e_i sind in der Haupterhebung und in den Wiederholungszählungen dieselben, in der Kontrollerhebung aber wesentlich kleiner. Die zufälligen Fehler ε_{it} haben in der Haupterhebung und in den Wiederholungszählungen die gleiche Gauß'sche Fehlerverteilung, während in der Kontrollerhebung die Gauß'sche Verteilung eine erheblich kleinere Streuung aufweist (kleinere Streuung der Zufallsfehler).

Die individuellen Antwortvariabilitäten wirken sich auch auf die Einordnung der Erhebungseinheiten in Größenklassen und damit auf die nach Größenklassen

gegliederten Erhebungsergebnisse aus. Ein Maß für die Antwortvariabilitäten in nach Größenklassen gegliederten Erhebungsergebnissen ist der sog. Inkonsistenz-Index. Er misst die Unzuverlässigkeit oder Inkonsistenz der Antworten R_i der Einheiten i.

In den vorhergehenden Ausführungen und vorgelegten Schemata wurde gezeigt, wie man den individuellen wahren Wert approximativ auf Grund eines Fehlermodells mit Kontrollerhebung schätzen kann. Der individuelle Zufallsfehler ε_{it} kann stochastisch bestimmt werden, wie im Folgenden gezeigt wird mit Hilfe der Variate Difference Methode. Der systematische Fehler kann dann als Differenz zwischen Angabewert einerseits und wahrer Wert + Zufallsfehler andererseits geschätzt werden.

Da sowohl der Inkonsistenz-Index wie auch die Schätzung des individuellen Zufallsfehlers ε_{it} mit Hilfe der Variate Difference-Methode in der Praxis noch wenig bekannt sind, sollen diese beiden Verfahren in den Abschnitten 3 und 4 etwas ausführlicher erläutert werden.

Besonders sei noch darauf hingewiesen, dass in der Survey Statistik die Methoden des Operations Research bisher kaum beachtet worden sind. Bei der Erhebung von statistischen Daten kann z.B. das Problem des regional und zeitlich zu optimierenden Einsatzes von Zählern oder Interviewern entstehen. Es ist aus Gründen der Zeit- und Kostenersparnis wünschenswert, die Interviewer oder Kontrollbeamten einer Kontrollerhebung den Gemeinden bzw. Betrieben oder Haushalten optimal zuzuordnen. Ein derartiges Modell wurde in Belgien für den Einsatz von Moniteuren (Beamte des Institut national de statistique im Außendienst) bei einer Nachprüfung entwickelt. Der berechnete Zeitaufwand für die Durchführung der Kontrollerhebung reduzierte sich von acht auf etwa fünf Arbeitstage. Das bedeutete, daß die Feldarbeit innerhalb einer Woche abgeschlossen werden kann und die Ergebnisse schneller vorliegen könnten (R.Wiegert/ K.Kafka/ H.Strecker/ R.Steylaerts 1976, 428-463).

3. Inkonsistenz-Index

Der Inkonsistenz-Index, berechnet für die nach Größenklassen h (h=1,2,3,...,L) gegliederten Erhebungsergebnisse, nimmt Werte zwischen 0 und 1 an oder in Zeichen $0 \le I_{Rh} \le 1$.

Wenn die individuellen Angaben von den Erhebungseinheiten aus zwei oder mehr (Wiederholungs-)Zählungen keine Variabilität der Antworten bezüglich der Einordnung der Einheiten i in einzelne Größenklassen aufweisen, ist $I_{Rh} = 0$, bei totaler Unzuverlässigkeit ("zufällige" Verschiedenheiten der Antworten) ergibt sich der Index-Wert von 1. Der Gesamtindex I_R für alle Größenklassen ist dann das Mittel der I_{Rh}.

Die Schätzung von Inkonsistenz-Indices für verschiedene Größenklassen h und der Antwortvarianzen setzt die Kenntnis von mindestens zwei oder mehr Angabewerten (Antworten) für jede Einheit i, gewonnen aus einer Haupterhebung mit N Einheiten und einer oder mehrerer Wiederholungszählungen, voraus. Jede zusätzliche Erhebung zur Haupterhebung ist in der Regel mit einem hohen Kosten- und Organisationsaufwand sowie einer stärkeren Belastung der Befragten verbunden. Dieses Verfahren sollte dennoch bei wichtigen Zählungen angewandt werden, da es insbesondere dazu dient, die Qualität statistischer, nach Größenklassen gegliederten Daten aus Erhebungen zu beurteilen, deren Ergebnisse die Grundlage für weitere Arbeiten der Wirtschafts-, Markt- und Meinungsforschung bilden.

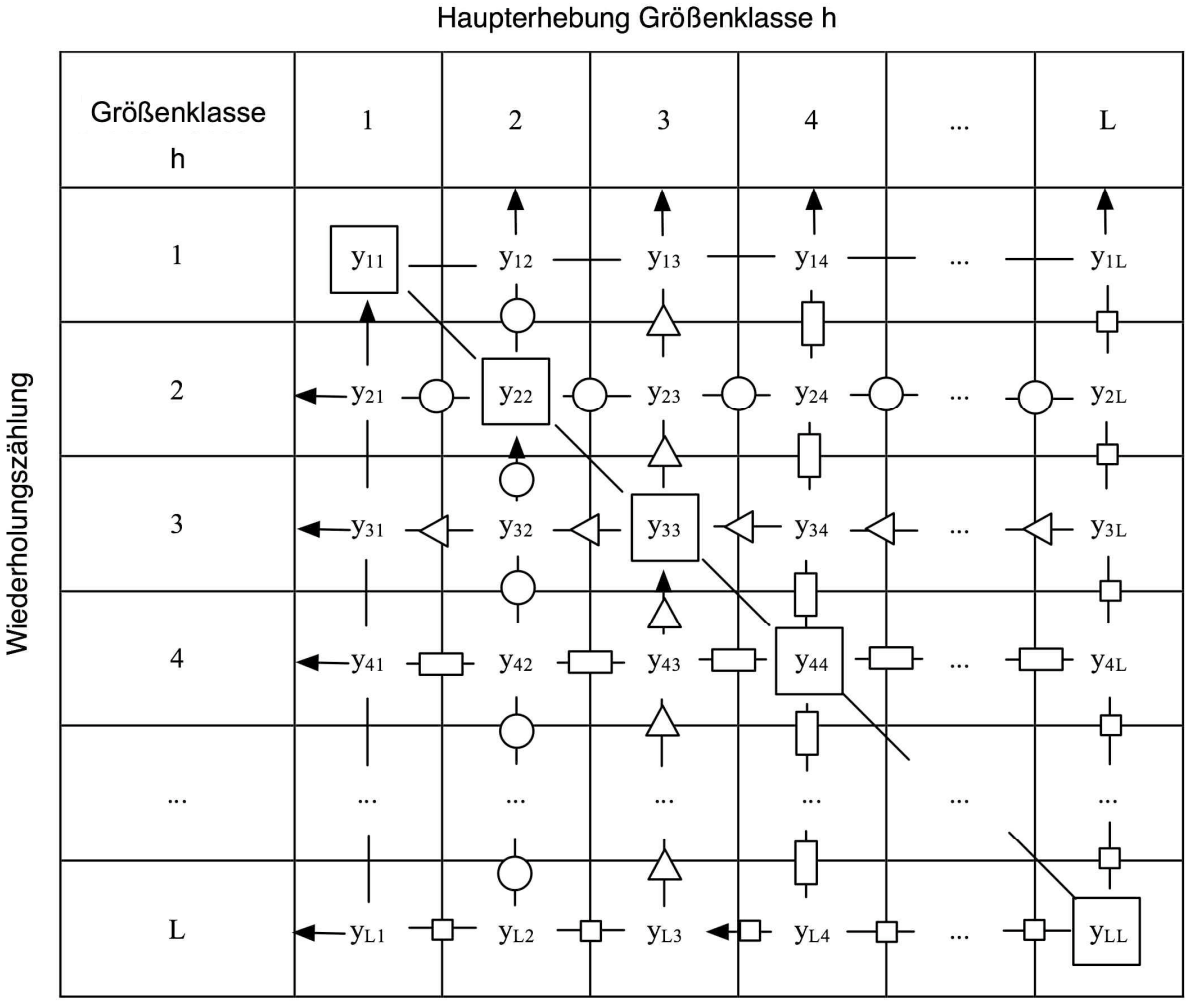
Es ist verständlich, dass in der Praxis fast ausschließlich nur eine zusätzliche Erhebung zur Haupterhebung (t = 2) und diese meist nur als Stichprobe mit dem Umfang n durchgeführt werden kann. Ist die Haupterhebung selbst eine Stichprobe, so wird empfohlen, die Wiederholungszählung als Unterstichprobe vorzunehmen. Hier wird die Schätzformel für den Inkonsistenz-Index nur für den Spezialfall mit zwei Antworten $y_{it=1G}$, $y_{it=2G}$ angegeben.

Die Werte der Zufallsvariablen y_{i1G}, y_{i2G} sind die Antworten der Einheiten i in der Haupterhebung (t = l) und in der Wiederholungszählung (t = 2) , welche beide nach dem gleichen Arbeitssystem, jedoch in der Regel mit verschiedenen Stichtagen, durchgeführt wurden.

Wenn die Merkmalswerte von Bestandsmassen zwischen den beiden Stichtagen nicht konstant bleiben (z.B. Schweinebestände in landwirtschaftlichen Betrieben), dann müssen im Gegensatz zu zeitweise konstanten Merkmalen (z.B. Landwirtschaftlichen Nutzflächen) die individuellen Angaben auf einen einheitlichen Stichtag, meist den der Haupterhebung, adjustiert werden und seien dann einfach mit y_{it} (t = 1, 2, usw.) bezeichnet. Wenn die Angaben aus Bewegungsmassen gewonnen wurden, müssen diese Werte auf denselben Referenz-Zeitraum bezogen

sein. Die Grundlage für die Schätzung liefern die nach Größenklassen gegliederten Erhebungsergebnisse, die in der nachfolgenden Kontingenz-Tabelle dargestellt sind.

Kontingenz-Tabelle



Erklärung zur Kontingenz-Tabelle:

Die Zahlen $y_{h,h}$ in den Größenklassen h (Haupterhebung) und h (Wiederholungszählung) (h = 1,2,3,...,L; h = 1,2,3,...,L) sind Häufigkeiten von Paaren der individuellen Angaben y_{it} aus den beiden (Wiederholungs-)Zählungen (1) und (2).

Beispiel: $y_{h=1,h=1} = y_{11}$ ist die Häufigkeit der Zahlenpaare $y_{it=1}$, $y_{it=2}$, deren Angaben in der Haupterhebung der Größenklasse h=1 und in der Wiederholungszählung wieder h=1 zugeordnet sind. In einem anderen Beispiel mögen die individuellen Angaben in der Haupterhebung $y_{it=1}$ und in der Wiederholungszählung $y_{it=2}$ zur Zuordnung in die Größenklasse 1 (Haupterhebung) bzw. h=3 (Wiederholungszählung) führen und damit zu dem Paar y_{13}.

Die Zuordnung der Angabewerte y_{it} zu den verschiedenen Größenklassen erfolgt nun, indem diese Werte mit Hilfe der Dichotomisierung z_{i1} und z_{i2} - Variablen als (0,1) - Variable nach folgender Vorschrift zugeordnet werden:

$$z_{it} = \begin{cases} 1, \text{ wenn die gegebene Antwort } y_{it} \text{ der Einheit i fur ein Merkmal eine bestimmte Eigenschaft A in der (Wiederholungs-)Zählung t aufweist, hier: ist die Einheit i der Großenklasse h zuzuordnen.} \\ 0, \text{ wenn die gegebene Antwort } y_{it} \text{ der Einheit i eine andere Merkmalsausprägung (nicht A, in Zeichen A) hat, hier: ist die Einheit i nicht der Großenklasse h zuzuordnen.} \end{cases}$$

Im vorliegenden Fall hat t die Werte 1 für die Haupterhebung oder 2 für die Wiederholungszählung.

Da die Aufgabe ist, für Proportionen (Anteile von einem qualitativen Merkmal mit der Eigenschaft A = zugehörig zur Größenklasse h bzw. nicht A ($\overline{A}$ = nicht zugehörig zur Größenklasse h) Inkonsistenz-Indices zu schätzen, sei von einer Stichprobe mit n Elementen und folgendem Schema ausgegangen; insgesamt mögen L Größenklassen ausgewiesen sein.

Vier-Felder-Schema

Wiederholungszählung oder Repetition	Haupt- oder Originalerhebung		Σ
	Größenklasse h A $z_{i1=1}$	Rest $\overline{A}$ $z_{i1=0}$	
Größenklasse h A $z_{i2=1}$	a_h	b_h	$a_h + b_h$
Rest $\overline{A}$ $z_{i2=0}$	c_h	d_h	$c_h + d_h$
Σ	$a_h + c_h$	$b_h + d_h$	n

Zuordnung der Einheiten (Betriebe) i: "Der Größenklasse h zugehörig - Eigenschaft A " bzw. "nicht der Größenklasse h zugehörig - Eigenschaft $\overline{A}$ (Rest)"

Die Werte a_h, b_h, c_h und d_h in diesem Schema sind die Häufigkeiten von Paaren aus beiden Erhebungen (Haupterhebung (1) und Wiederholungszählung (2) und der Stichprobenumfang ist n).

Für jede Größenklasse h = l, 2, ..., L wird ein solches Vier-Felder-Schema aufgestellt. Die insgesamt L Vier-Felder-Schemata erhält man aus der vorstehenden Kontingenz-Tabelle durch Streichung von einzelnen Zeilen und Kolonnen ähnlich wie man den Wert einer Determinante aus den Unterdeterminanten durch Streichen von Zeilen und Kolonnen bestimmt. Die Vier-Felder-Schemata für die einzelnen Größenklassen h haben die folgenden Werte:

h = 1

$a_1 = y_{11}$

b_1 Man streicht die erste Zeile y_{12}, y_{13}, y_{14}, ..., y_{1L} ↔

c_1 Man streicht die erste Kolonne y_{21}, y_{31}, y_{41}, ..., y_{L1} ↕

d_1 Der Wert ergibt sich aus $n-(a_1+b_1+c_1)$

Vier-Felder-Schema

h = 1, y_{11}

$a_1 = y_{11}$	$b_1 = y_{12}+y_{13}+y_{14}+...+y_{1L}$
$c_1 = y_{21}+y_{31}+y_{41}+...+y_{L1}$	$d_1 = n-(a_1+b_1+c_1)$

Linie

Analog erhält man für die übrigen Größenklassen h=2, h=3, h=4, ..., h=L die weiteren Vier-Felder-Schemata.

h = 2, y_{22}

$a_2 = y_{22}$	$b_2 = y_{21}+y_{23}+y_{24}+...+y_{2L}$
$c_2 = y_{12}+y_{32}+y_{42}+...+y_{L2}$	$d_2 = n-(a_2+b_2+c_2)$

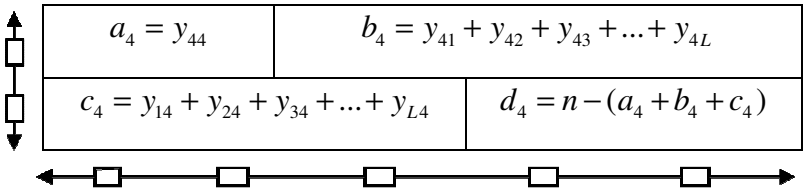
Kreis

h = 3, y_{33}

$a_3 = y_{33}$	$b_3 = y_{31}+y_{32}+y_{34}+...+y_{3L}$
$c_3 = y_{13}+y_{23}+y_{43}+...+y_{L3}$	$d_3 = n-(a_3+b_3+c_3)$

Dreieck

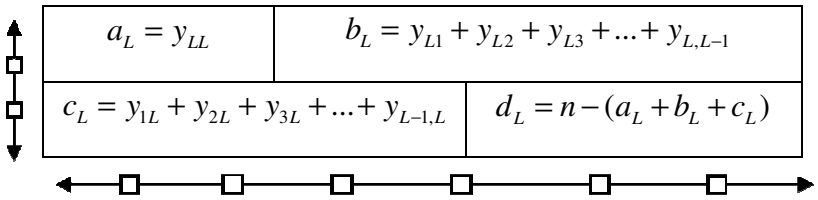
h = 4, y_{44}



Rechteck

...

h = L, y_{LL}



Quadrat

Für jede Einheit i wird die individuelle Antwortvarianz ($s^2_{R_i}$) aus den individuellen Angabewerten ((0,1)-Variablen der k=2 (Wiederholungs-)Zählungen)) gebildet. Bei einer Vollerhebung ergeben sich N Antwortvarianzen (i = l, 2, ..., N), bei einer Stichprobe n Antwortvarianzen (i = l, 2, ..., n). Das Mittel aller individuellen Antwortvarianzen ist dann die Einfache Antwortvarianz (s^2_R). Sie wird ins Verhältnis zur einfachen Binomialvarianz ($P_G(1-P_G)$) der Binomialvariablen (0,1), üblicherweise zitiert als Stichprobenzufallsvarianz einer einelementigen Stichprobe, gesetzt. Diesen Quotienten $s^2_R / P_G(1-P_G)$ nennt man dann den Inkonsistenz-Index. Er ist ein Maß für die Unzuverlässigkeit oder Inkonsistenz der Antworten der Einheiten i - hier der Zuordnung der Angaben zu verschiedenen Größenklassen h.

Anhand des vorstehenden Vier-Felder-Schemas kann man ohne große Schwierigkeiten zuverlässige Schätzwerte für die Einfache Antwortvarianz σ^2_R und die Binomialvarianz ($P_G(1-P_G)$) bestimmen und in die Inkonsistenz-Index-Formeln einsetzen. Als Schätzformel erhält man nach einigen Rechnungen

$$I_{Rh} \approx \hat{I}_{Rh} = \frac{b_h + c_h}{\frac{(a_h + c_h)(b_h + d_h) + (a_h + b_h)(c_h + d_h)}{n}},$$

wobei $n = a_h + b_h + c_h + d_h$ ist.

Dieser Schätzer $\hat{I}_{Rh}$ (≤ 1) gilt nur, wenn $b_h^2 + c_h^2 \leq 2a_h d_h$ ist, was in der Praxis in der Regel erfüllt ist. Zufallsbedingt ist, dass die Zufallsvariable $\hat{I}_{Rh}$ auch einmal einen Wert etwas größer als 1 (>1) annehmen kann, aber die Wahrscheinlichkeit dafür ist sehr gering.

Für jede der Größenklassen (h = I, 2, ..., L) ist ein solcher $\hat{I}_{Rh}$-Wert zu berechnen. Der Gesamtindex ist dann das gewogene arithmetische oder geometrische Mittel aus diesen Größenklassen Indices $\hat{I}_{Rh}$ mit den Gewichten N_h = Zahl der Einheiten i in der Größenklasse h. Näheres hierzu siehe Strecker 1999, 359-361 mit ausführlichen Literaturangaben.

Notwendig für eine weitere Analyse auf diesem Gebiet wäre eine genaue Kenntnis der Verteilung der $\hat{I}_{Rh}$ bzw. auch der $\hat{I}_R$ und davon, wie sich die Indexwerte bei einer Zusammenfassung oder einer weiteren Untergliederung der Größenklassen ändern würden. Hier besteht Handlungsbedarf!

Beispiel

Nachfolgend sollen die Ergebnisse einer Schätzung des Inkonsistenz-Index aus der Amtlichen Statistik Belgiens mitgeteilt werden. In sehr kurzem zeitlichen Abstand fanden am 15. Mai 1985 die Haupterhebung (Landwirtschafts- und Gartenbauzählung - G) als Vollerhebung und am 4. Mai 1985 eine Wiederholungszählung ex ante (Enquete Pilote - G') als Stichprobe nach annähernd gleichem Arbeitssystem (G≈G') statt. Die Analyse beschränkte sich auf die Angaben über die Landwirtschaftliche Nutzfläche der hauptberuflichen Betriebsinhaber (Kategorie 1) von insgesamt N≈61 000 Betrieben in der Haupterhebung und n≈2 350 Betrieben in der als Wiederholungszählung durchgeführten Stichprobe. Näheres zu den Erhebungsmethoden und Fragebogen siehe H.Strecker/R.Wiegert avec la collaboration de J.Peeters, L.Anciaux, E.Draelants, 2000.

In Belgien war es, zu dieser Vorbereitungszeit, technisch und organisatorisch nur möglich, als aktuellen Rahmen für die Auswahl der Betriebe i in der Stichprobe die frühere Landwirtschafts- und Gartenbauzählung vom 15. Mai 1983 heranzuziehen. Da 1985 einige der 1983 erfassten Betriebe nicht mehr bestanden und in der Zwischenzeit Neuzugänge erfolgten, musste eine Trennung der untersuchten Betriebe in zwei Schichten, Alter Bestand (N_{ABh}) und Neuer Bestand (N_{NBh}) vorgenommen werden. Vom alten Bestand wurde eine Stichprobe von 1 873 Betrieben ausgewählt, die Neuzugänge von 477 Betrieben wurden total erfasst. Die Ergebnisse dieser Befragung sind in der nachfolgenden Kontingenz-Tabelle aufgeführt. Anhand dieser Übersicht wurden dann für jede Größenklasse h, getrennt nach Altem und Neuem Bestand, jeweils Inkonsistenz-Indices berechnet. Die einzelnen Berechnungen erfolgten auf Grund der Vier-Felder-Schemata. Aus dieser Kontingenz-Tabelle wurden, getrennt nach Altem und Neuem Bestand, jeweils vier Vier-Felder-Schemata durch Streichen von Zeilen und Kolonnen aufgestellt - so exemplarisch für die Größenklasse h = I: 0.01 - < 10 ha Landwirtschaftlicher Nutzfläche die Tabellen a) und b) - Tabellen 1 und 2.

Tabelle 1: Kontingenz-Tabelle

Landwirtschafts- und Gartenbauzählung in Belgien am 15. Mai 1985
- Hauptberufliche Betriebsinhaber (Kategorie 1) -

Kontingenztabelle für die Zuordnung der Alten und Neuen Betriebe nach den Angabewerten der Landwirtschaftlichen Nutzfläche in der Haupterhebung (15.Mai 1985) und der Enquête Pilote (4. Mai 1985) zu den Größenklassen h der Landwirtschaftlichen Nutzfläche (LN)

Alter Bestand (AB) - Stichprobe **Neuer Bestand (NB) - Vollerhenung**

Enquête Pilote Größenklasse LN ha h		Haupterhebung Größenklasse LN ha, h									
		0.01 - < 10		10 - < 30		30 - < 100		100 - < +		Σ	
		1		2		3		4			
		AB	NB	AB	NB	AB	NB	AB	NB	AB	NB
0.01 - < 10	1	410	213	36	7	0	2	0	0	446	222
10 - < 30	2	12	9	767	173	27	8	0	0	806	190
30 - < 100	3	1	0	17	12	533	47	8	1	559	60
100 - < +	4	0	0	0	0	3	1	59	4	62	5
Σ		423	222	820	192	563	58	67	5	1 873	477

Tabelle 2: Vier-Felder-Schema für die erste Größenklasse (h=l)

0.01 -< 10 ha Landwirtschaftliche Nutzfläche, a_1, b_1, c_1, d_1

a) Alter Bestand (AB)

Enquête Pilote y_{i2}	Haupterhebung y_{i1}		
	$h=1$ A $z_{i1}=1$	Rest $\overline{A}$ $z_{i1}=0$	**Σ**
$h=1$ A $z_{i2}=1$	a_1 = 410	b_1 = 36	a_1+b_1 = 446
Rest $\overline{A}$ $z_{i2}=0$	c_1 = 13	d_1 = 1 414	c_1+d_1 = 1 427
Σ	a_1+c_1 = 423	b_1+d_1 = 1 450	$n_{AB}=a_1+b_1+c_1+d_1$ = 1 873

a) Neuer Bestand (NB)

Enquête Pilote y_{i2}	Haupterhebung y_{i1}		
	$h=1$ A $z_{i1}=1$	Rest $\bar{A}$ $z_{i1}=0$	**Σ**
$h=1$ A $z_{i2}=1$	$a_1=213$	$b_1=9$	$a_1+b_1=222$
Rest $\bar{A}$ $z_{i2}=0$	$c_1=9$	$d_1=246$	$c_1+d_1=255$
Σ	$a_1+c_1=222$	$b_1+d_1=255$	$n_{NB}=a_1+b_1+c_1+d_1=477$

Es sind

h = 1: 0.01 - < 10 ha; Rest: 10 ha - < +

$$\left.\begin{array}{l} \text{A} = 1 \text{ für } 0.01 \le \text{y}_{\text{i..}} < 10 \\ \bar{\text{A}} = 0 \text{ Rest - 10 ha} \le \text{y}_{\text{i..}} < + \end{array}\right\} \text{ Landwirtschaftliche Nutzfläche}$$

Die Schätzung des Index für eine Größenklasse (h = 1) ergibt, getrennt nach altem Bestand (AB) und Neuzugängen (NB) die beiden Schätzwerte:

$$\hat{I}_{ABh=1} = \frac{49}{\frac{423\cdot 1450 + 446\cdot 1427}{1873}} = 0.0734 \approx 7,34\%$$

$$\hat{I}_{*h=1} = \frac{(b_{*1}+c_{*1})}{\frac{(a_{*1}+c_{*1})(b_{*1}+d_{*1})+(a_{*1}+b_{*1})(c_{*1}+d_{*1})}{n_*}}$$

(für * = AB bzw. NB und für n_* ist einmal $n_{AB}=1873$ und zum anderen $n_{NB}=477$ zu setzen)

$$\hat{I}_{NBh=1} = \frac{18}{\frac{222\cdot 255 + 222\cdot 255}{477}} = 0.07583 \approx 7,58\%$$

Die Schätzung der Indices für jede Größenklasse beruht auf den Zahlen für den Alten und den Neuen Bestand, siehe Tabellen 1 und 3.

Tabelle 3: Zahl der Betriebe nach der Landwirtschafts- und Gartenbauzählung am 15. Mai 1985 der Kategorie 1 in Belgien

Größenklasse der Landwirtschaftlichen Nutzflächen ha, h	N_{ABh}	N_{NBh}	N_h $(= N_{ABh} + N_{NBh})$	$N_{h\|85}$	Landwirtschaftliche Nutzfläche Y_h ha 15. Mai 1985
0.01 - < 10	19 017	222	19 239	19 284	105 371,89
10 - < 30	29 622	192	29 814	29 844	538 202,54
30 - < 100	11 489	58	11 547	11 558	537 080,05
100 - < +	629	5	634	635	85 176,63
Insgesamt (Σ)	60 757	477	$N_{NB} =$ 61 234	$N_{h\|85} =$ 61 321	$Y_N =$ 1 265 831.11

Die für die jeweilige Größenklasse und für die Gesamtheit geschätzten Indices sind in der folgenden Tabelle mitgeteilt. Die Mittel sind gewogene Mittelwerte mit den Gewichten N_{ABh} und N_{NBh} - einmal geometrische Mittel $\hat{I}_{gh}$ und zum anderen arithmetische Mittel $\hat{\bar{I}}_h$, gewonnen aus der Zahl der Betriebe im Alten Bestand und den Meldungen der Gemeinden an Neuzugängen. Diese N_h stimmten gut mit den später in der Haupterhebung festgestellten Zahlen der Betriebe $N_{h|85}$ überein (siehe Tabelle 4).

Tabelle 4: Die Schätzwerte des Inkonsistenz-Index - in %

h	$\hat{I}_{ABh}$	$\hat{I}_{NBh}$	$\hat{I}_{gh}$	$\hat{\bar{I}}_h$
1	7.34	7.58	7.35	7.35
2	10.00	15.72	10.03	10.04
3	7.13	23.21	7.17	7.21
4	8.83	20.21	8.89	8.92
			$\hat{i}_g = 8.53$	$\hat{\bar{i}}_g = 8.63$

Erläuterung

Der Index I_h für die Größenklasse 1 (0,01 - < 10 ha LN) (Alter Bestand + Neuzugänge) ist das gewogene Mittel mit den Gewichten $N_{ABh=1} = 19017$ und $N_{NBh=1} = 222$.

So ergab sich ein Schätzwert für das geometrische Mittel $\hat{I}_{gh=1} = 0,0735 \approx 7,35\%$ und für das arithmetische Mittel $\hat{\bar{I}}_{h=1}$ $0,0735 \approx 7,35\%$. In der Tabelle 4 sind die für vier Größenklassen berechneten Indices aufgeführt. Der Gesamt-Index $\hat{I}_g$ bzw. $\hat{\bar{I}}$ wurde wieder als gewogenes Mittel mit den Gewichten N_h: N_1, N_2, N_3, N_4 errechnet (Formeln für die Schätzung siehe Strecker 1997, 654 und 1998, 184).

Der Gesamt-Index betrug: $\hat{\bar{I}} = 0,0863 \approx 8,63\%$, $\hat{I}_g = 0.053\%$.

Für alle Schätzwerte ist die Bedingung $b_{*h}^2 + c_{*h}^2 \leq 2a_{*h}d_{*h}$ ($* = AB$ bzw. NB) erfüllt. Die Ergebnisse für das Merkmal Landwirtschaftliche Nutzfläche Y_h, Y weisen also Antwortvariabilitäten $\hat{I}_{gh}$, $\hat{\bar{I}}_h$, $\hat{I}_g$, $\hat{\bar{I}}$ auf, wie sie in der Tabelle angegeben sind. Die geschätzten Werte für die Inkonsistenz-Indices zeigen, dass der Einfluss der Variabilitäten der Antworten auf nach Größenklassen aufgegliederten Ergebnissen nicht vernachlässigt werden darf.

Wegen eines vermutlich bestehenden Erinnerungseffektes bei den Befragten ist eine positive Korrelation zwischen den individuellen Antwortabweichungen von Zählung zu Zählung (y_{i1G} und y_{i2G}) vorhanden. Daher sind die vorstehenden Inkonsistenz-Indices etwas nach unten verzerrt; es liegt möglicherweise eine Unterschätzung vor (vgl. Strecker 1997, 645-646). Die Größe der Verzerrung ist kaum zu bestimmen, da die Korrelation schwer zu berechnen ist. Die Folgerung, die sich daraus ergibt: Die angegebenen Index-Werte sind untere Schranken für das Ausmaß des Einflusses der Antwortvariabilitäten auf die Besetzungszahlen in den Größenklassen.

4. Schätzung der individuellen glatten Komponente und der jeweiligen Zufallskomponente der individuellen Angabewerte mit Hilfe der Variate Difference-Methode

Man kann die individuelle glatte Komponente m_i der Einheit i mit Hilfe der Variate Difference-Methode schätzen. Da dieses Schätzverfahren in diesem Zusammenhang kaum bekannt ist, wird darüber nachfolgend etwas ausführlicher berichtet.

Zuvor wird ein kurzer Abriss der Methode als Zeitreihen-Analyseverfahren gegeben, um diese dann in einem Spezialfall auf die gestellte Aufgabe anzuwenden.

Die Variate Difference-Methode ist ein iteratives Verfahren, um die Beobachtungswerte y_t einer statistischen Zeitreihe in die permanente oder glatte Komponente m_t und in die stochastische oder Zufallskomponente ε_t zu zerlegen. (Näheres hierzu siehe Strecker u.a., vol.9, 1988, 484-488 mit ausführlichen Literaturangaben).

Die Zufallskomponente wird mit Hilfe sukzessiver Differenzen von der glatten Komponente separiert, so dass dann die glatte und damit auch die Zufalls-komponente geschätzt werden können.

Es wird ausgegangen von einer Zeitreihe zum Zeitpunkt t mit N Beobachtungswerten

y_t für t = 1, 2, ..., k, ..., N zu verschiedenen Zeiten t.

Die Angabewerte mögen additiv aus der glatten Komponente m_t und der stochastischen Komponente ε_t zusammengesetzt sein (lineare Verbundenheit):

$y_t = m_t + \varepsilon_t$ (t = l, 2, ..., k, ..., s, ..., N) N = Zahl der gesamten Beobachtungen

Als glatte oder permanente Komponente soll z.B. ein Polynom n-ten Grades oder andere Funktionen, wie z.B. die logarithmischen Funktionen oder Exponentialfunktionen usw. angenommen werden.

Ferner wird vorausgesetzt, dass

$E_t(\varepsilon_t) = 0$, $E_t(\varepsilon_t^2) = \sigma^2$ (Homoskedastizität)

$E_{t,s}(\varepsilon_t \cdot \varepsilon_s) = 0$ (wenn $t \neq s$ ist) $E_{t,s}(m_s \cdot \varepsilon_t) = 0$

(keine Korrelation mit jedem anderen Wert).

Diese Annahmen sind plausibel und entsprechen grundsätzlich der Realität. Es werden sukzessive Differenzen gebildet und zwar:

$\Delta(y_t) = y_{t+1} - y_t$,

$\Delta^2(y_t) = \Delta(\Delta(y_t)) = \Delta(y_{t+1} - y_t) = \Delta(y_{t+1}) - \Delta(y_t)$

...

$$\Delta^l(y_t) = \Delta(\Delta^{l-1}(y_t)) = \Delta^{l-1}(y_{t+1} - y_t) = \Delta^{l-1}(y_{t+1}) - \Delta^{l-1}(y_t)$$

und es ist

$$\Delta(y_t) = \Delta(m_t) + \Delta(\varepsilon_t)\,,$$

$$\Delta^2(y_t) = \Delta^2(m_t) + \Delta^2(\varepsilon_t)$$

... usw. ...

$$\Delta^l(y_t) = \Delta^l(m_t) + \Delta^l(\varepsilon_t)\,, \qquad \text{usw., usw.}$$

Es zeigt sich, dass z.B. die oben erwähnten Funktionen der glatten Komponente mit zunehmender Zahl der Differenzen Δ immer kleiner und kleiner und die Differenzen Δ der Zufallskomponente immer größer und größer werden. So ist z.B. die n-te Differenz eines Polynom n-ten Grades $\Delta^n(y_t)$ konstant und die (n+1) Differenz hat den Wert 0. Entsprechendes gilt auch z.B. für andere oben erwähnte Funktionen. So werden die höheren Differenzen der zu untersuchenden Daten nur noch Werte der stochastischen Komponente enthalten. Man kann eine Testfunktion R_K bilden, bestehend aus empirischen Varianzen verbunden mit Bionomial-Koeffizienten, die approximativ bei höheren Differenzen normal verteilt mit Mittelwert 0 und Varianz 1 ist. Wenn $|R_K| < 3$ ist, widerspricht nichts der Annahme, dass die glatte oder permanente Komponente m_t nach $l = l_0$ Differenzen ganz oder bis auf kleine Reste eliminiert ist.

Den Schätzwert für m_t erhält man mit Hilfe einer linearen Annäherungsfunktion F_t als Aggregat mit konstanten Koeffizienten b_1, b_2, b_3 usw.:

$$m_t \approx \hat{m}_t = y_t + F_t\,,$$

wobei $F_t = b_1\Delta^{2n}(\varepsilon_{t-n}) + b_2\Delta^{2n+2}(\varepsilon_{t-n-1}) + ...$

usw. ist. Aus Gründen der Symmetrie wird $l_0 = 2n$ für l_0 eine gerade und $l_0 + 1 = 2n$ für l_0 eine ungerade Zahl gesetzt. n ist hier ein Laufindex und nicht der Umfang einer Stichprobe oder der Grad eines Polynoms. m_t wird man umso effizienter schätzen, je größer die Zahl der Koeffizienten b_* ist.

Die Koeffizienten b_1, b_2 usw. werden dann so ausgewählt, dass

$$D = E_t(\varepsilon_t + F_t)^2 = Minimum$$

wird. Setzt man die so erhaltenen b_1, b_2 usw. in F_t ein, so erhält man mit Hilfe von $y_t + F_t$ die relativ besten Schätzwerte $\hat{m}_t$ für die Komponente der glatten Werte m_t. Führt man diese Rechnungen aus, dann erhält man in erster Näherung

$$m_t \approx \hat{m}_{t(1/3)} = \frac{1}{3}(y_{t-1} + y_t + y_{t+1}), \qquad \text{wobei t = 1, 2, 3, ..., k, ... usw.}$$

Weitere Näherungen sowie Literaturangaben siehe H. Strecker in: M. Beckmann/ R.Wiegert 1987, 279-280 sowie 323-325. Übrigens sind die vorstehend angeführten Annahmen und Hypothesen, wie man leicht zeigen kann, in der Regel erfüllt und entsprechen der Realität. Einen Schätzwert der stochastischen Komponente ε_t zum Zeitpunkt t erhält man z.B. durch Substraktion

$$\varepsilon_t \approx \hat{\varepsilon}_{t(1/3)} = y_t - \hat{m}_{t(1/3)} \qquad (t = 1, 2, 3, ..., k, ...)$$

Diese Methode der Schätzung der glatten Komponente soll nun auf die hier gestellte Aufgabe übertragen werden, die individuellen glatten Komponenten in der Survey Statistik zu schätzen. Es wird von der linearen Verbundenheit, hier in der Erhebungseinheit i, ausgegangen:

$$y_{it} = m_i + \varepsilon_{it} \qquad i = 1, 2, 3, ..., N \text{ (bzw. n, } t = 1, 2, 3, ..., k, ...)$$

In den Einheiten i sind die glatten Komponenten m_i jeweils eine Konstante, etwa $m_i = a_i$ und ε_{it} die stochastische Komponente einer Zeitreihe in der Einheit i (t =1, 2, 3, ..., k). Die Annahmen und Hypothesen mögen die gleichen bleiben wie bei der Variate Difference-Methode als Zeitreihen-Zerlegungsmethode. Da von der Annahme ausgegangen wird, dass die (Wiederholungs-)Zählungen t = l, 2, 3, ..., k unabhängig sind und die gleiche Streuung aufweisen (Homoskedastizität), gilt $E_t(\varepsilon_{it}) = 0$, $E_t(\varepsilon_{it}^2) = \sigma_i^2$, $E_{t,s}(\varepsilon_t \cdot \varepsilon_s) = 0 \quad (t \neq s)$ (keine Autokorrelation).

Laut Modellaufbau sind diese Annahmen erfüllt. Die Variate Difference-Methode kann also angewandt werden und es gilt dann:

$$y_{it} = m_i + \varepsilon_{it} \qquad \Delta(y_{it}) \qquad (t = 1, 2, 3, ..., k, ...)$$

$$\text{z.B. k} = 4 \begin{cases} \mathrm{y_{i1}} = m_i + \varepsilon_{i1} \\ \mathrm{y_{i2}} = m_i + \varepsilon_{i2} \\ \mathrm{y_{i3}} = m_i + \varepsilon_{i3} \\ \mathrm{y_{i4}} = m_i + \varepsilon_{i4} \end{cases} \rightarrow \begin{cases} y_{i2} - y_{i1} = (m_i + \varepsilon_{i2}) - (m_i + \varepsilon_{i1}) = \varepsilon_{i2} - \varepsilon_{i1} \\ y_{i3} - y_{i2} = (m_i + \varepsilon_{i3}) - (m_i + \varepsilon_{i2}) = \varepsilon_{i3} - \varepsilon_{i2} \\ y_{i4} - y_{i3} = (m_i + \varepsilon_{i4}) - (m_i + \varepsilon_{i3}) = \varepsilon_{i4} - \varepsilon_{i3} \end{cases}$$

usw. $\qquad \Delta^2(y_{it})$

$$(y_{i3} - y_{i2}) - (y_{i2} - y_{i1}) =$$

$$((m_i + \varepsilon_{i3}) - (m_i + \varepsilon_{i2})) - ((m_i + \varepsilon_{i2}) - (m_i + \varepsilon_{i1})) = \varepsilon_{i3} - 2\varepsilon_{i2} + \varepsilon_{i1}$$

$$(y_{i4} - y_{i3}) - (y_{i3} - y_{i2}) =$$

$$((m_i + \varepsilon_{i4}) - (m_i + \varepsilon_{i3})) - ((m_i + \varepsilon_{i3}) - (m_i + \varepsilon_{i2})) = \varepsilon_{i4} - 2\varepsilon_{i3} + \varepsilon_{i2}$$

usw. i = 1, 2, 3, ..., N bzw. n

usw., usw.

Hieraus sieht man, dass die glatte Komponente m_i der Einheit i als Konstante bereits nach der ersten Differenz der Angabewerte ($\Delta^1(y_{it})$) eliminiert wurde; das ist ein Spezialfall der Variate Difference-Methode. Man kann direkt die glatte Komponente m_i z.B. mittels 3er Durchschnitte schätzen und zwar ist bei drei (Wiederholungs-)-Zählungen

$$m_i \approx \hat{m}_{i(1/3)} = \frac{1}{3}(y_{it-1} + y_{it} + y_{it+1}) = \frac{1}{3}(y_{i1} + y_{i2} + y_{i3}) .$$

Die individuellen Zufallsfehler der drei (Wiederholungs-)Zählungen in der Einheit i sind dann als Schätzwerte

$$\varepsilon_{it} = y_{it} - m_i \approx \hat{\varepsilon}_{it} = y_{it} - \hat{m}_{i(1/3)} \qquad (t = 1, 2, 3).$$

(Siehe hierzu Strecker, 2004, 198-230)

Erschwerend für die Schätzung ist hierbei, dass die individuellen Angaben der Einheit i aus drei (Wiederholungs-)Zählungen bekannt sein müssen. Das ist in der Praxis wegen der Kosten und des in der Realisierung der Zählungen nicht unerheblichen Arbeitsaufwandes schwierig. Dabei muss ferner untersucht werden, ob die individuellen Angabewerte y_{it} zeitlich schnellen Änderungen (wie z.B. Schweinebestände) unterliegen, oder ob sie für eine kürzere Zeit, etwa 2 bis 3 Monate, konstant bleiben, z.B. Weizenfläche von der Aussaat bis zur Ernte. Wenn bei variablen Werten und verschiedenen Stichtagen bzw. Referenz-Zeiträumen unterschiedliche Angaben gemacht wurden, muss eine Adjustierung der Daten auf einen einheitlichen Stichtag bzw. Berichts-Zeitraum vorgenommen werden. Bei zeitweise konstanten Merkmalswerten können die (Wiederholungs-)Zählungen auf einen Zeitraum verteilt werden und eine Adjustierung ist in diesem Fall nicht nötig. Weitere Einzelheiten siehe H.Strecker/R.Wiegert 1986, 99-130.

Da es zur Zeit nicht möglich ist, drei Wiederholungszählungen in einem kurzen Zeitraum durchzuführen, stützt sich das folgende Beispiel auf simulierte individuelle Daten. So wurden die individuellen Angabewerte y_{it} sowie ihre Komponenten glatter Wert m_i, Zufallsfehler ε_{it} und wahrer Wert x_i für drei unabhängige (Wiederholungs-)-Zählungen (i=1,2,3) anhand von Daten aus der belgischen Landwirtschaftsstatistik

simuliert. Der individuelle Zufallsfehler ε_{it} wurde der hier behandelten Aufgabenstellung entsprechend nach der Variate Difference-Methode geschätzt. Auf eine Schätzung der wahren Werte x_i wird hier nicht eingegangen.

Die Resultate für eine Gesamtheit N = 10, 15 und 20 Einheiten wurden in den folgenden Tabellen 5 und 6 zusammengefasst: Die Schätzwerte der individuellen Zufallsfehler $\hat{\varepsilon}_{it(1/3)}$ ergaben sich aus der Differenz zwischen Angabewerten y_{it} und der glatten Komponente m_i $\varepsilon_{it} \approx \hat{\varepsilon}_{it(1/3)} = y_{it} - \hat{m}_{i,t(1/3)}$. Die durchschnittlichen Zufallsfehler $\hat{\bar{\varepsilon}}_{Nt(1/3)}$ betrugen z.B. für t = 1 und N = 10, 15 und 20 (absolut) $|-0.97|$, $|-0.85|$, $|-0.62|$.

Tabelle 5: Simulationsbeispiel Einheit i

Die individuellen Angabewerte y_{it} sowie ihre Komponenten wahrer Wert x_i, systematischer Fehler e_i und Zufallsfehler ε_{it} für drei unabhängige (Wiederholungs-) Zählungen (t = l, 2, 3) und die Schätzwerte der glatten Komponente mit 3er Durchschnitten

	Glatte Komponente m_i, x_i, e_i			Zufallsfehler Auswahl t = 1, 2, 3 ε_{it}			Angabewert y_{it}			Schätzung 3er Durchschnitte
i	m_i	x_i	e_i	ε_{i1}	ε_{i2}	ε_{i3}	y_{i1}	y_{i2}	y_{i3}	$\hat{m}_{i(1/3)}$
1	75	66	9	-4	-3	-1	71	72	74	72.33
2	131	120	11	-2	0	4	129	131	135	131.67
3	310	288	12	-9	-9	-9	301	301	301	301.00
4	79	62	17	0	3	-3	79	82	76	79.00
5	50	38	12	2	-2	0	52	48	50	50.00
6	106	99	7	-8	2	2	98	108	108	104.67
7	36	31	5	3	-3	0	39	33	36	36.00
8	113	107	6	0	0	6	113	113	119	115.00
9	29	21	8	-2	3	-1	27	32	28	29.00
10	54	50	4	1	0	2	55	54	56	55.00
$\sum_{i=1}^{10}$	983	892	91	-19	-9	0	964	974	983	973.67
11	85	77	8	4	0	-4	89	85	81	85.00
12	78	68	10	0	-1	-3	78	77	75	76.67
13	42	47	-5	-1	1	1	41	43	43	42.33
14	159	143	16	-6	6	-2	153	165	157	26.67
15	25	25	0	0	2	3	25	27	28	
$\sum_{i=1}^{15}$	1 372	1 252	120	-22	-1	-5	1 350	1 371	1 367	1 362.67

16	100	87	13	2	6	-4	102	106	96	101.33
17	73	68	5	0	2	-4	73	75	69	72.33
18	21	21	0	0	3	-2	21	24	19	21.33
19	54	58	-4	-2	1	-1	52	55	53	53.33
20	82	88	-6	3	1	3	85	83	85	84.33
$\sum_{i=1}^{20}$	1 702	1 574	128	-19	12	-13	1 683	1 714	1 689	1 695.35

Zahl der Differenzen Δ zwischen den geschätzten (Mittel-)Werten $\hat{\bar{m}}_{N(1/3)}$ und den Mitteln der "wahren" glatten Komponenten $\bar{m}_N$, für die Gesamtheiten N = 10, 15, 20

Größenklasse	$\lvert\Delta(\hat{\bar{m}}_{N(1/3)})\rvert$ N = 10	N = 15	N = 20
0.00 - < 0.25			
0.25 - < 0.50			1
0.50 - < 0.75		1	**Σ** 1
0.75 - < 1.00	1	**Σ** 1	
1.00 - < 1.25	**Σ** 1		
1.25 - < 1.50			
1.50 - < 1.75			
1.75 - < 2.00			

Erläuterung:

$$\Delta(\hat{\bar{m}}_{N(1/3)}) = \frac{1}{N}\sum_{i=1}^{N}\hat{m}_{i(1/3)} - \bar{m}_N$$

$$\bar{m}_N = \frac{1}{N}\sum_{i=1}^{N} m_i$$

Man sieht, dass mit zunehmendem N die geschätzten durchschnittlichen Zufallsfehler kleiner werden (H. Strecker 2004, 587 Tabelle: Schätzwerte der Zufallsfehler). Weiterhin wurde noch die Zahl der Differenzen $\Delta(\hat{\bar{m}}_{N(1/3)}) = \frac{1}{N}\sum_{i=1}^{N}\hat{m}_{i(1/3)} - \bar{m}_N$ zwischen den durchschnittlichen geschätzten Komponenten-Werten $\hat{\bar{m}}_{N(1/3)}$ $(= \frac{1}{N}\sum_{i=1}^{N}\hat{m}_{i(1/3)})$ und den Mitteln der wahren glatten Komponenten $\bar{m}_N$ für die Gesamtheiten = 10, 15, 20 berechnet. Dabei zeigte sich, dass mit zunehmendem Umfang der Gesamtheiten N die Differenzen kleiner werden, d.h. die Mittel der geschätzten glatten Komponenten näheren sich den Mitteln der wahren glatten Werte. Daraus ergibt sich, dass die Variate Difference-Methode ein geeignetes Verfahren ist, um die individuellen Zufallsfehler in den Angaben der Einheiten i zu schätzen. Die Zufallsfehler in den Gesamtergebnissen einer Erhebung werden umso kleiner je größer N ist.

Mit der Maßzahl "Antwortvariabilität" kann man noch Aussagen über die Variabilität der Zufallsfehler in den Gesamt-Ergebnissen mit N Einheiten i vornehmen. Diese gibt Auskunft darüber, wie weit zufallsbedingte Gesamt-Ergebnisse aus t = 2 und mehr (Wiederholungs-)-Zählungen variieren können.

Wenn die Angabewerte y_{it} der Einheit i in der (Wiederholungs-)Zählung t mit t = 1, 2, 3, ... bezeichnet sind und insbesondere bei zwei Erhebungen t = 1, 2 sind, dann ist die individuelle Varianz der Zufallsfehler der Einheit i

$$\sigma_{R_i}^2 = E_t(y_{it} - E_t(y_{it}))^2 \approx s^2 = \frac{1}{2-1}\left[\left(y_{it} - \frac{y_{i1}+y_{i2}}{2}\right)^2 + \left(y_{i2} - \frac{y_{i1}+y_{i2}}{2}\right)^2\right] = \frac{1}{2}(y_{i1} - y_{i2})^2$$

$$E_t(y_{it} - y_i)^2 = E_t(x_i + e_i + \varepsilon_{it} - x_i - e_i)^2 = E_t(\varepsilon_{it}^2) = \sigma_{R_i}^2$$

Die Gesamtvarianz der Zufallsfehler ist dann

$$\sigma_R^2 = E_i(\sigma_{R_i}^2) \approx s_R^2 = \frac{1}{N}\sum_{i=1}^{N} s_{R_i}^2$$

- N Gesamtzahl der Einheiten i

Tabelle 6: Mittelwert der Zufallsfehler $\bar{\varepsilon}_{Nt}$ und ihre Schätzwerte $\hat{\bar{\varepsilon}}_{Nt(1/3)}$ sowie die Differenzen Δ zwischen Schätzwerten und wahren Mittelwerten der Zufallsfehler der $\bar{\varepsilon}_{Nt}$ (t = 1, 2, 3) in der Abfolge für verschiedene Gesamtheiten N

Auswahl

Zufallsfehler	Gesamtheiten N			Zufallsfehler	Gesamtheiten N		
	10	15	20		10	15	20
$\lvert\bar{\varepsilon}_{N1}\rvert$	$\lvert-1.90\rvert > \lvert-1.47\rvert > \lvert-0.95\rvert$			$\lvert\bar{\varepsilon}_{N3}\rvert$	$\lvert0.00\rvert <^* \lvert-0.33\rvert <^* \lvert-0.65\rvert$		
$\lvert\hat{\bar{\varepsilon}}_{N1(1/3)}\rvert$	$\lvert-0.97\rvert > \lvert-0.85\rvert > \lvert-0.62\rvert$			$\lvert\hat{\bar{\varepsilon}}_{N3(1/3)}\rvert$	$\lvert0.93\rvert > \lvert0.29\rvert <^* \lvert-0.32\rvert$		
$\lvert\Delta(\hat{\bar{\varepsilon}}_{N1(1/3)})\rvert$	$\lvert0.93\rvert > \lvert0.62\rvert > \lvert0.33\rvert$			$\lvert\Delta(\hat{\bar{\varepsilon}}_{N3(1/3)})\rvert$	$\lvert0.93\rvert > \lvert0.05\rvert <^* \lvert0.33\rvert$		
$\lvert\Delta(\hat{\bar{\varepsilon}}_{N1(1/3)}) / \bar{m}_N\rvert$	$\lvert0.95\%\rvert > \lvert0.68\%\rvert > \lvert0.39\%\rvert$			$\lvert\Delta(\hat{\bar{\varepsilon}}_{N3(1/3)}) / \bar{m}_N\rvert$	$\lvert0.95\%\rvert > \lvert0.05\%\rvert <^* \lvert0.39\%\rvert$		
$\lvert\bar{\varepsilon}_{N2}\rvert$	$\lvert-0.90\rvert > \lvert-0.07\rvert <^* \lvert0.60\rvert$			(*) stochastische Störung Erläuterung: $\bar{\varepsilon}_{Nt} = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_{it}$, $\hat{\bar{\varepsilon}}_{Nt(1/3)} = \frac{1}{N}\sum_{i=1}^{N}\hat{\varepsilon}_{it(1/3)}$ $\Delta(\hat{\bar{\varepsilon}}_{Nt(1/3)}) = \hat{\bar{\varepsilon}}_{Nt(1/3)} - \bar{\varepsilon}_{Nt}$, $\bar{m}_N = \frac{1}{N}\sum_{i=1}^{N} m_i$ t = 1, 2, 3		
$\lvert\hat{\bar{\varepsilon}}_{N2(1/3)}\rvert$	$\lvert0.03\rvert <^* \lvert0.55\rvert > \lvert0.33\rvert$						
$\lvert\Delta(\hat{\bar{\varepsilon}}_{N2(1/3)})\rvert$	$\lvert0.87\rvert > \lvert0.60\rvert > \lvert0.33\rvert$						
$\lvert\Delta(\hat{\bar{\varepsilon}}_{N2(1/3)}) / \bar{m}_N\rvert$	$\lvert0.85\%\rvert > \lvert0.66\%\rvert > \lvert0.39\%\rvert$						

Antwortvarianz und Antwortvariabilität

Bei zwei Zählungen ergeben sich dann als Schätzwerte für die Gesamt-Antwortvarianz $s_R^2 = \frac{1}{N}\sum_{i=1}^{N} s_{R_i}^2$ und für die Antwortvariabilität $\hat{V}_{RN} = \frac{s_R^2}{2N}$ oder auch $\hat{V}_{RN}^{1/2} = \frac{s_R}{\sqrt{2N}}$.

Da bei der Antwortvariabilität in jeder Größenklasse die Mittel der jeweiligen Zufallsfehler nach dem Zentralen Grenzwertsatz normal verteilt sind, sind auch die geschätzten Mittelwerte normal verteilt (Reproduktivität). Demnach sind bei einer Gesamtheit von N Einheiten die Schätzwerte der Gesamtmittel approximativ normal verteilt und man kann schon bei zwei Zählungen mit $V_{RN}^{1/2}$ bzw. $\hat{V}_{RN}^{1/2}$ Konfidenzintervalle für die Zufallsfehler in den Gesamt-Ergebnissen der Erhebungen angeben.

Ergänzend wird noch ein numerisches Beispiel für die Schätzung der Antwortvarianz und der Antwortvariabilität bei einem konstanten Merkmalswert (keine Adjustierung auf einen einheitlichen Stichtag notwendig) gegeben werden. Im Mai 1985 wurden in Belgien kurzfristig nach einem annähernd gleichen Arbeitssystem zwei Erhebungen - Landwirtschafts- und Gartenbauzählung als Haupterhebung am 15. Mai und Enquête Pilote als Wiederholungszählung ex ante am 4.Mai durchgeführt. Auf Grund dieser Daten wurden die Antwortvarianzen und Antwortvariabilitäten für Anbauflächen verschiedener Fruchtarten (Winterweizen, Wintergerste, Zuckerrüben, Handelsgewächse) und der Landwirtschaftlichen Nutzfläche insgesamt bestimmt. Das Beispiel beschränkt sich auf die Angaben für die Weizenanbaufläche der Betriebe (i) der Kategorie 5: Betriebe, die pflanzliche Produkte erzeugen und zugleich Dienstleistungen für andere Betriebe anbieten. Die Zahl N dieser Einheiten i betrug 450, alle wurden zweimal befragt (y_{i1}, y_{i2}), es gab somit keine Stichprobenzufallsfehler, was für die Schätzung von Vorteil war. Da die Wiederholungszählung auf freiwilliger Grundlage erfolgte, beteiligten sich bei dieser Enquête Pilote N = 378 Betriebe. Die Response Quote von 378/450 = 84 % war noch sehr zufriedenstellend. Die Ergebnisse dieser Untersuchung werden in der Tabelle 7 mitgeteilt. Unter "effektiv" werden hier die Betriebe bezeichnet, die in beiden Erhebungen - Haupterhebung und Enquête Pilote - individuelle Angaben über ihre Anbauflächen gegeben haben. Agrarstatistische Fachleute haben die Schätzwerte der Antwortvarianzen und insbesondere der Antwortvariabilitäten als plausibel anerkannt.

Tabelle 7: Belgien - Landwirtschafts- und Gartenbauzählung Mai 1985

Kategorie 5 - Vollerhebung - Merkmal: Winterweizenfläche Antwortvarianz und Antwortvariabilität

Die Antwortvarianzen s_{Rh}^2 in den Größenklassen h (Schichten) ,Gesamt-Antwortvarianzen s_R^2 und die Antwortvariabilitäten $\hat{V}_{Rh}$ und $\hat{V}_{RN}$							
Größenklasse der landwirtschaftlichen Nutzfläche in ha, h	Zahl der Betriebe (effektiv) $N_{h(85)}$	Gesamtwert det Angaben (effektiv) in a		Gesamt mittel $\frac{Y_{2h}+Y_{1h}}{2N_{h(85)}}$ in a	Antwortvarianz $s_{Rh}^2 = \frac{1}{N_{h(85)}} \sum_{i=1}^{N_{h(85)}} s_{Rih}^2$	Antwortvariabilität $\hat{V}_{Rh}$, $\hat{V}_{RN}$	
		4. Mai Y_{2h}	15. Mai Y_{1h}			$\hat{V}_{Rh} = \frac{S_{Rh}^2}{2N_{h(85)}}$	Wurzel %
0,01 - < 10	266	10 987	11 120	41.553	207.103	0.389	1.5
10 - < 30	87	26 003	25 621	296.690	8 181.023	47.017	2.3
30 - < 100	24	19 355	20 699	834.458	42 663.458	888.822	3.6
100 - < +	1	4 500	4 500	4 500	0	0	0.0
insgesamt	$N_{(85)} = 378$ Zahl der Betriebe der Kategorie 5	$Y_2 = \sum_h Y_{2h}$ = 60 845	$Y_1 = \sum_h Y_{1h}$ = 61 940	$\frac{Y_2 + Y_1}{2N_{h(85)}}$ = 162.414 in a	s_R^2 = 4 737.464	$\hat{V}_{RN} = \frac{s_R^2}{2N_{(85)}}$ = 6.266	1.5
		Gesamtweizenfläche (effektiv) in a					

Es bedeutet $Y_{2h} = \sum_{i=1}^{N_{h(85)}} y_{i2}$, $Y_{1h} = \sum_{i=1}^{N_{h(85)}} y_{i1}$

Antwortvarianz der Einheit i in der Größenklasse h und i = 1, 2, 3, ..., N_h

$$\sigma_{R_i h}^2 \approx s_{R_i h}^2 = \frac{1}{2}(y_{i1h} - y_{i2h})^2$$

Antwortvarianz der Größenklasse h

$$\sigma_{Rh}^2 \approx s_{Rh}^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} s_{R_i h}^2$$

Gesamt-Antwortvarianz

$$\sigma_R^2 \approx s_R^2 = \frac{1}{N} \sum_{h=1}^{L} N_h \cdot s_{R_h}^2$$

Antwortvariabilität der Größenklasse h

$$V_{Rh} \approx \hat{V}_{Rh} = \frac{s_{Rh}^2}{2N_h}$$

Gesamt-Antwortvariabilität

$$V_{RN} \approx \hat{V}_{RN} = \frac{s_R^2}{2N}$$

Haupterhebung und Wiederholungserhebung (Enquête Pilote) wurden als Vollerhebungen durchgeführt (n = N) - kein Stichprobeneffekt im engeren Sinne der Stichprobentheorie!

(L = Zahl der Größenklasse h, im Beispiel L=4) Zahl der effektiven Betriebe N_h, N der Kategorie 5 nach Zählung Mai 1985.

5. Fehlermodell

Für die theoretischen Betrachtungen sei zunächst von einem deduktiven Fehlermodell ausgegangen, während für die Berichtigung von Zählungsergebnissen der induktive Fehlerbegriff der adäquate ist. Der deduktive Angabe- oder Antwortfehler ist für den i-ten Befragten definiert: Fehler = Angabewert - wahrer Wert, der induktive Fehler wird umgekehrt erklärt. Die deduktive Betrachtungsweise ist für Analysen von theoretischen Zusammenhängen (Zerlegung des Angabewertes in die Komponenten: Wahrer Wert, systematischer Fehler, Zufallsfehler) geeignet, während der induktive Fehlerbegriff zugrunde gelegt werden soll, wenn die Aufgabe gestellt ist, diejenige Größe (Verbesserung) zu finden, die dem Angabewert hinzugefügt werden muss, um einen berichtigten Wert (wahren Wert) zu erhalten. Dieser Gedanke findet z.B. Anwendung in der Geodäsie.

Nachfolgend soll ein Fehlermodell mit einer Untergliederung des individuellen Angabewertes y_{it} in wahren Wert x_i, systematischen Fehler e_i und Zufallsfehler ε_{it} dargelegt werden.

Weiter soll das Fehlermodell in Verbindung mit einer Vollerhebung ohne Zählereffekt behandelt werden. In Westeuropa werden meist die individuellen Daten durch Selbstausfüllung der Fragebogen ohne Beeinflussung durch Zähler gewonnen. Das Fehlermodell einer Vollerhebung erhält man, indem man in den Formeln den Stichprobenumfang n durch den Gesamtumfang N ersetz.

Begriffe und Symbole

Es möge sein:

N Zahl der Befragten in der Grundgesamtheit - bei allen nachfolgenden Überlegungen als vollständig erfasst vorausgesetzt

n Zahl der Befragten in der Stichprobe

y_{it} individueller angegebener Wert (Angabewert) der Einheit i (des Befragten i) in der (Wiederholungs-)Zählung t und zwar innerhalb kurzer Zeit für denselben Stichtag oder Berichtszeitraum mehrmals befragt (t = l,2,3,... und i = 1,2,...,N bzw. i = 1,2,...,n) (Antwortvariabilität)

x_i Wahrer Wert des Merkmals beim i-ten Befragten (i = 1,2,...,N, bzw. i = 1, 2, ..., n)

e_i individueller systematischer Fehler des Befragten i (Einheit i)

ε_{it} individueller Zufallsfehler des Befragen i (Einheit i) in der (Wiederholungs-)-Zählung t (t = 1, 2, 3, ...) - zur Vereinfachung wurden die Zufallsfehler als voneinander unabhängig angenommen, was eine plausible Arbeitshypothese ist

Die Verbundenheit der Fehlerkomponenten soll linear sein, daher gilt

$y_{it} = x_i + e_i + \varepsilon_{it}$ (hier angenommen, dass es keinen Zählereffekt gibt!)

Demzufolge hat man

$E_t(y_{it} \mid i) = x_i + e_i + E_t(\varepsilon_{it} \mid i) = x_i + e_i$, - in Anlehnung an die Zeitreihenanalyse als "glatter Wert" (m_i) bezeichnet

$E_t(\varepsilon_{it} \mid i) = 0$, - nach Definition des Zufalls

$e_i + \varepsilon_{it}$ individueller Angabefehler des Befragten (Einheit i) in der (Wiederholungs-)Zählung t

Demnach ist

$\varepsilon_{it} = y_{it} - (e_i + x_i)$ individueller Zufallsfehler

$E_t(e_i + \varepsilon_{it}) = e_i$ individueller systematischer Fehler des Befragten i (Einheit i)

Mittelwert und Varianz

Hier soll zur Vereinfachung von einem Stichprobenmodell einer Einfachen uneingeschränkten Zufallsstichprobe ohne Zurücklegen und ohne Zählereffekt ausgegangen werden. Die Werte für eine Vollerhebung erhält man ohne Schwierigkeit, indem man statt des Stichprobenumfanges n den Gesamtumfang N einsetzt.

Die nachfolgenden Schätzformeln der Stichprobe können auch in die höheren Stichprobenmodelle subsumiert werden, denn alle weiterführenden Stichprobenmodelle sind nach dem "Baukastenprinzip" aus dem Modell einer uneingeschränkten Stichprobe aufgebaut.

Es gilt für die Einheit i mit k (Wiederholungs-)Zählungen (t = 1, 2, 3, ..., k), die Stichprobe mit Umfang n und die Gesamtheit N

Einheit i

$$y_i = E_t(y_{it} \mid i) \approx \hat{y}_i = \frac{1}{k}\sum_{t=1}^{k} y_{it} = \frac{1}{k}\sum_{t=1}^{k}(x_i + e_i + \varepsilon_{it})$$

Stichprobe n

Mittel $$\hat{\bar{y}}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = ... = \frac{1}{n}\sum_{i=1}^{n}(x_i + e_i + \frac{1}{k}\sum_{t=1}^{k}\varepsilon_{it}) = \hat{\bar{x}}_n + \hat{\bar{e}}_n + \hat{\bar{\varepsilon}}_{n(k)} = \hat{\bar{y}}_{(k)n}$$

$$\hat{\bar{x}}_n = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad \hat{\bar{e}}_n = \frac{1}{n}\sum_{i=1}^{n} e_i\,, \quad \hat{\bar{\varepsilon}}_{n(k)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k}\sum_{t=1}^{k}\varepsilon_{it}$$

Gesamtheit N

Mittel $\bar{y}_N \approx \frac{1}{N}\sum_{i=1}^{N}(x_i + e_i + \frac{1}{k}\sum_{t=1}^{k}\varepsilon_{it}) = \bar{x}_N + \bar{e}_N + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{k}\sum_{t=1}^{k}\varepsilon_{it}$

Wert der Gesamtheit N: $Y_N = \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} e_i + \sum_{i=1}^{N}\frac{1}{k}\sum_{t=1}^{k}\varepsilon_{it} = x_N + e_N + \hat{\varepsilon}_{(k)\dot{N}}$

Es führt zu weit, für die vorstehenden Schätzformeln die einschlägigen Formeln der Stichprobenzufallsvarianzen aufzuzeigen. Man kann diese nach den in der Literatur behandelten üblichen Stichprobenmodellen entwickeln.

Von großer Bedeutung für die Praxis ist zweifellos die folgende Punktschätzung des durchschnittlichen Angabefehlers und die Stichproben-Varianzschätzung.

Wenn man den individuellen Angabefehler

$e_{it}(\varepsilon_{it}) = y_{it} - x_i$ (systematischer Fehler e_i + Zufallsfehler ε_{it})

in der (Wiederholungs-)Zählung t so einführt und definiert, dann ergeben sich bei einer (Wiederholungs-)Zählung als Vollerhebung mit N Einheiten und einer Kontrollerhebung als Stichprobe vom Umfang n (n<N) mit ohne Zurücklegen als Schätzwerte für den Mittelwert der individuellen Angabefehler einer Erhebung bei k = l mit einer Haupterhebung als Mittelwert der individuellen Angabefehler

$$\hat{\bar{e}}_{(\varepsilon_{it=1})n} = \frac{1}{n}\sum_{i=1}^{n} e_{it=1(\varepsilon_{it=1})}$$

und als Stichprobenzufallsvarianz

$$\sigma^2_{\hat{\bar{e}}(\varepsilon_{it=1})n} \approx s^2_{\hat{\bar{e}}(\varepsilon_{it=1})n} = \frac{N-n}{N-1}\cdot\frac{s^2}{n}, \quad \text{mit } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(e_{it=1} - \hat{\bar{e}}_{nt=1})^2$$

und bei k (Wiederholungs-)Zählungen als Mittelwert

$$\hat{\bar{e}}_{(\varepsilon_{it})n} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k}\sum_{t=1}^{k} e_{it(\varepsilon_{it})}$$

und als Stichprobenzufallsvarianz

$$V_{(\hat{\bar{e}}_{(\varepsilon_{it})n})} = \frac{N-n}{N-1}\cdot\frac{\sigma_b^2}{n} + \frac{\sigma_w^2}{nk}, \quad \text{(+Korrelationen)}$$

hierbei sind

$$\sigma_b^2 = \underset{i}{E}\left[e_{i(\varepsilon_{it})} - \underset{i}{E}\, e_{i(\varepsilon_{it})} \right]^2$$

und

$$\sigma_w^2 = \underset{i}{E}\left[(\underset{t}{E}\, e_{i(\varepsilon_{it})} - e_{i(\varepsilon_{it})})^2 \right]$$

Bei nur k=2 Zählungen (Haupterhebung und einer Wiederholungszählung) ist in den Formeln k=2 zu setzen. Wenn die Haupterhebung selbst eine Stichprobe mit Umfang n ist, dann kann die Kontrollerhebung entweder eine Unterstichprobe mit einem Umfang n' (n'<n) oder es werden alle Angaben aus der Stichprobe auf ihre Fehlerhaftigkeit (n'=n) überprüft. Die Formeln für die Punkt- und Varianzschätzung sind dann nach den Regeln der Stichprobentheorie und der Schätzung entsprechend zu gestalten. Schließlich kann auch die Kontrollerhebung bei den Einheiten i in vollem Umfang wie in der Haupterhebung durchgeführt werden (n=N). Die Schätzformel für das Mittel der individuellen Fehler ist dann $\hat{\bar{e}}_{(\varepsilon_{it})N} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{k}\sum_{t=1}^{k} e_{it(\varepsilon_{it})}$ und die Varianz ist dann einfach nach dem Stichproben-Modell zu bestimmen.

In diesem Mittel werden öfters auch die individuellen Fehler $e_{it(\varepsilon_{it})}$ je nach der Richtung des Einzelfehlers unterschiedliche Vorzeichen (+ oder -) haben und daher bezeichnet man diese Gesamtmittel als Mittel der Netto-Fehler. Die Qualität der individuellen statistischen Daten aus einer Erhebung, deren wahren Werte x_i z.B. aus einer Kontrollerhebung als Stichprobe vom Umfang n gewonnen werden, kann man bekanntlich nur mit Hilfe des Mittelwertes der Brutto-Angabefehler gebührend beurteilen. Dieser wird aus den Absolutbeträgen der individuellen Angabefehler $\left|e_{it(\varepsilon_{it})}\right|$ geschätzt. In die Stichprobenformeln ist jeweils der Wert $\left|e_{it(\varepsilon_{it})}\right|$ einzusetzen, d.h. $\hat{\bar{e}}^{*}_{(\varepsilon_{it})n} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k}\sum_{t=1}^{k}\left|e_{it(\varepsilon_{it})}\right|$.

Es gibt Erhebungen, in denen sich die individuellen Netto-Fehler um 0 symmetrisch verteilen und der Mittelwert der Netto-Fehler exakt oder ungefähr 0 (≈ 0) ist, während der Mittelwert der Brutto-Fehler $\hat{\bar{e}}^{*}_{(\varepsilon_{it})n}$ größer oder sogar erheblich größer als 0 (> 0) ist. Die Erhebung weist also unter Umständen große fehlerhafte Daten auf (Strecker 1995, 402-424).

Abschließend soll hier noch auf das wichtige Fehlermaß Mittlerer quadratischer Fehler "Mean Square Error (MSE)" hingewiesen werden. Dieser gibt eine umfassende Auskunft über die Qualität der Daten. Er lässt sich in folgende Komponenten zerlegen:

MSE	=	Gesamt(Stichproben)-Antwortvarianz	+	(Stichproben-zufallsvarianz)	+	Kovarinaz (Wechselwirkungen)	+	Bias-Quadrat

Nach umfangreichen Umformungen (Strecker 1980, 385-420) erhält man allgemein für eine Stichprobe n entsprechend der vorstehenden Symbolik die Mean Square Error-Zerlegung:

$$MSE(\hat{\bar{y}}_{(k)n}) = \underset{i,t}{E}(\hat{\bar{y}}_{(k)n} - \bar{x}_g)^2 \approx \underset{i,t}{E}(\hat{\bar{y}}_{(k)n} - \bar{x}'_g)^2 =$$

$$= \frac{\sigma_R^2}{nk}\left[1 + (k-1)\rho_t + (n-1)\rho_R + (n-1)(k-1)\rho'_R\right] + V(\hat{\bar{y}}_{(k)n}) + B^2$$

hierbei sind:

n Umfang der Stichprobe

k Zahl der (Wiederholungs-)Zählungen

$\frac{\sigma_R^2}{nk}$ Antwortvarianz des Stichprobenmittels oder (Gesamt-)Antwortvariabilität

$B = (Y_N - X_{GN}) \approx B' = (Y_N - X'_N)$ Bias (Gesamt-Angabefehler)

ρ_t Mittlere Autokorrelation der Angabewerte y_{it} innerhalb derselben Einheiten i bei verschiedenen (Wiederholungs-)Zählungen t, t' (i bzw. $i = j,\ t \neq t'$)

ρ_R Korrelation zwischen den Angabewerten verschiedener Einheiten (ij) in derselben (Wiederholungs-)Zählung t ($i \neq j,\ t = t'$)

ρ'_R Korrelationen zwischen den Angaben verschiedener Einheiten (i, j) bei verschiedenen (Wiederholungs-)Zählen t, t' ($i \neq j,\ t \neq t'$)

$V(\hat{\bar{y}}_{(k)n}) = \underset{i,t}{E}(\hat{\bar{y}}_{(k)n} - \bar{Y}_N)^2$ Stichprobenzufallsvarianz

Für die Schätzung des Mean Square Error (MSE) sind dann die entsprechenden Schätzwerte für $\sigma_R^2,\ \rho_t,\ \rho_R,\ \rho'_R,\ \hat{V}(\hat{\bar{y}}_{(k)n})$ in die Formel einzusetzen.

Für k=2 kann man Abschätzungen des MSE angeben:

$$V(\hat{\bar{y}}_{2n}) + B'^2 \leq \widehat{MSE}(\hat{\bar{y}}_{2n}) \leq \frac{\sigma_R^2}{n} + V(\hat{\bar{y}}_{2n}) + B'^2$$

Unterer Schätzwert des MSE ohne Berücksichtigur der Antwortvarianz

Oberer Schätzwert des MSE mit Berücksichtigung der Antwortvarianz

Beispiel

In der amtlichen Agrarstatistik Belgiens werden seit 1965 in mehrjährigen Abständen deskriptive Kontrollerhebungen durchgeführt. Beamte des Statistischen Außendienstes (Moniteure) im Institut national de statistique nehmen diese Überprüfung der individuellen Daten vor, um statistische Fehler in den Angabewerten festzustellen.

Das erfolgt, indem diese Bediensteten im Rahmen einer Stichprobe vom Umfang n eine nochmalige Zählung genau nach den Vorschriften und Richtlinien unter Zuhilfenahme von Unterlagen in den Betrieben (z.B. Stallbüchern, Tierärztlichen Bescheinigungen, Verkaufs-Quittungen usw.) durchführen.

Als Grundlage für diese Bestimmung und Schätzung des Mittels der Angabefehler usw. diente das Fehler-Modell, wie es vorstehend dargelegt wurde. Auf Einzelheiten dieser Nachprüfungen soll hier nicht weiter eingegangen werden - in zahlreichen einschlägigen Veröffentlichungen (siehe z.B. Strecker, H., Wiegert, R., 1994, insbesondere S. 1 bis 98) wurden die Methoden und Ergebnisse dieser Kontrollen dargelegt.

Die Überprüfung erfolgte an die für die Ernährung wichtigen natürlichen Merkmale - hier Zahl der Schweine. Diese Merkmalsbegriffe ändern sich in der Zeit nicht und die Zählungs- und Kontrollergebnisse sind ohne Einschränkung voll vergleichbar.

In der Tabelle 8 sind im Auszug nur einige Maßzahlen und Ergebnisse von zwei Erhebungen und Kontrollen im Mai 1965 und Mai 1971 aufgeführt. Erwähnt sei hier, dass neben den Mittelwerten der Angabefehler je Betrieb (Netto-Fehler) auch die entsprechenden Werte der Brutto-Fehler mitgeteilt werden. Wie nicht anders zu erwarten, sind die Mittel der Netto-Fehler stets kleiner als die der Brutto-Fehler. Da es in der Wirtschaftsstatistik nur wenige Angaben über den MSE gibt, wurden hier noch bewusst Schätzwerte über den Mean Square Error aufgeführt. Die Größenordnungen der Mittel der Angabewerte und des Mean Square Error veranlassen keine Änderungen in der Organisation und im Ablauf des Arbeitssystems der Landwirtschafts- und Gartenbauzählungen in Belgien. Die durchschnittlichen Angabefehler unterscheiden sich nicht wesentlich von denen seinerzeit in der Bundesrepublik Deutschland festegestellten (1960, $\hat{\bar{e}}_n = 7.0\%$ und $\hat{V}^{1/2}_{(\hat{\bar{e}}_n)} = 1.2\%$ - Statistisches Bundesamt).

Tabelle 8: Belgien

Ergebnisse der Landwirtschafts- und Gartenbauzählungen am 15. Mai 1965 und 15. Mai 1971 sowie zweier Kontrollerhebungen als Stichproben in jeweils n = 240 Auswahlgemeinden mit insgesamt 960 Betrieben in je 4 Betrieben in jeder Auswahlgemeinde.

Merkmal: Zahl der Schweine

Zählung und Kontrollerhebung	Zahl der		Umfang der Stichprobe. Zahl der (Kontroll-)		Netto-Fehler			
					Mittel des Angabefehlers je Betrieb		Stichproben-zufallsvarianz (Wurzel)	
	Gemeinden	Betriebe	Gemeinden	Betriebe	$\hat{\bar{e}}_{(\varepsilon_{it})n}$	%	$\hat{V}^{1/2}_{(\hat{\bar{e}}_{(\varepsilon_{it})n})}$	%
15. Mai 1965	2 586	104 089	240	960	1.55	8.85	0.345	1.97
15. Mai 1971	2 381	75 963	240	960	2.77	5.36	0.345	0.67

Zählung	Zahl der Schweine Y_N	Mittelzahl der Schweine je Betrieb $\bar{Y}_N$	Mittlere Quadratische Fehler Mean Square Error ($\widehat{MSE}$)		Brutto-Fehler			
					Mittel des Angabefehlers je Betrieb		Stichproben-Zufallsvarianz (Wurzel)	
			Unterer Schätzwert	Oberer Schätzwert	$\hat{\bar{e}}^{*}_{(\varepsilon_{it})n}$	%	$\hat{V}^{1/2}_{(\hat{\bar{e}}^{*}_{(\varepsilon_{it})n})}$	%
15. Mai 1965	1 823 756	17.52	2.28	2.40	2.39	13.64	0.345	1.97
15. Mai 1971	3 916 702	51.56	7.55	7.67	3.67	7.12	0.335	0.65

Landwirtschafts- und Gartenbauzählung: Mittelwert und Varianz Schätzwert für den "wahren" Wert der Zahl der Schweine $\hat{X}'$ Berichtigtes Zählungsergebnis (Zahlen gerundet)				
Zählung	Zahl der Schweine Y_N	Schätzwert des Gesamt-Angabefehlers $\hat{E}_n$	Varianz (Wurzel) $\hat{V}^{1/2}_{(\hat{E}_n)}$	Berichtigtes Zählungsergebnis Zahl der Schweine $X' \approx \hat{X}' = Y_N + \hat{E}_n$
15. Mai 1965	1 823 756	161 442	35 911	1 985 189
15. Mai 1971	3 916 702	210 114	26 179	4 126 816

Weitere Hinweise und Kommentare zu diesem Thema sollen hier nicht gegeben werden; das Wesentliche wurde mitgeteilt. (Näheres siehe Strecker 1980, 385-420.)

Schlussbemerkungen

In dem vorstehenden Beitrag wurde gezeigt, wie eine Planung für eine statistische Erhebung zu erfolgen hat und welche Fehler in den Zählungsergebnissen enthalten sind sowie man diese bestimmen und in Maßzahlen (Mittelwerte der Fehler, Antwortvarianzen, Antwortvariabilitäten usw.) darstellen kann. Vieles zu dieser Problematik wäre noch anzumerken und auszuführen, aber hier sei nur auf einschlägige Veröffentlichungen verwiesen. Interessant wäre es, noch einige Grafiken anzufertigen, wie z.B. u.a. die Verteilung der individuellen Angabefehler oder Punkt- und Streuungsdiagramme der individuellen "wahren" Werte und Angabewerte mit den Koordinaten-Achsen "wahrer" Wert und Angabewert, oder auch, wenn die individuellen Angabewerte es erlauben, einen Vergleich der Endziffern der individuellen Angabewerte mit den Endziffern der durch die Kontrollen festgestellten einzelnen Angaben, wie z.B. dargestellt in einer Agrarstatistik Belgiens (Landwirtschafts- und Gartenbauzählung vom 15. Mai 1965, H. Strecker/R. Wiegert, u.a., 1983, 36-52, 85-88).

Die hier dargelegten Verfahren wurden in der Landwirtschaftsstatistik angewandt. Sie können auch ohne weiteres auf Erhebungen in der Bevölkerungs- und Wirtschaftsstatistik übertragen werden.

Die Werte der Maßzahlen wie Mittel der individuellen Netto- und Brutto-Angabefehler, der Antwortvariabilitäten, Inkonsistenz-Indices usw. müssen von den Konsumenten der Statistiken überprüft werden, ob die jeweiligen Erhebungsergebnisse für ihre Analysen und Auswertungen noch tolerierbar sind. Wenn nicht, dann muss das Arbeitssystem so geändert werden können, dass akzeptable Resultate gewonnen werden können.

Grundsätzlich wird sich in der Survey Statistik in nicht allzu ferner Zeit manches ändern, wenn viele Daten in Computern gespeichert werden und es aus Gründen der Kostenersparnis und Rationalisierung keine statistischen Ergebnisse gewonnen aus einer bzw. mehreren primärstatistischen (Wiederholungs-)Zählungen im klassischen Sinn geben wird. Schätzwerte wie Antwortvariabilität oder Inkonsistenz-Indices ect. können dann nicht mehr berechnet werden und damit fehlen wichtige Hinweise auf die Qualität statistischer Daten. Das bedeutet eine erhebliche Simplifizierung in der statistischen Datenanalyse.

Die Gesamt-Ergebnisse Y_G einer Zählung (t = I, Haupterhebung) setzen sich in der herkömmlichen Statistik aus den individuellen Angabewerten zusammen $Y_G = \sum_{i=1}^{N} y_{it=1}$ und weiter aufgegliedert $\sum_{i=1}^{N}(x_i + e_i + \varepsilon_{it=1}) = X + E + \sum_{i=1}^{N}\varepsilon_{it=1}$. $E = \sum_{i=1}^{N} e_i$ = systematischer Fehler bzw. bei einer Stichprobe die Summation von i von 1 bis n - $(\sum_{i=1}^{n})$.

Im Allgemeinen wird im Computer nur noch Y_G gespeichert. Würde man trotzdem die Daten für eine Wiederholungszählung erhalten wollen, dann ergäben sich dieselben Werte, nämlich $y_{it=2} = y_{it=1}$ und das Gesamt-Ergebnis wäre wieder dasselbe. Die Maßzahl Antwortvariabilität hätte somit ihre Grundlage verloren, da die Wiederholungszählung nicht unabhängig ist. Den Zufallsfehler gibt es nur einmal in dem Gesamt-Ergebnis der Haupterhebung und kann nicht numerisch geschätzt werden. Daher ist eine weitere Qualitätsanalyse nicht möglich.

Wenn von den Befragten nur ein Teil mit ihren Angaben im Computer gespeichert ist, und der andere Teil im herkömmlichen Sinne erhoben und aufbereitet wird, dann muss die Gesamtheit in zwei Schichten aufgeteilt werden und man muss für jede Teil-Gesamtheit getrennt besondere Analysen vornehmen, einmal mit Hilfe der hier beschriebenen Methoden, zum andern mit neu zu entwickelnden Methoden.

Bei der Aufbereitung und Analyse der Daten von Erhebungseinheiten, die keine Computer bei der Beantwortung benutzen, sind die Qualitätskriterien anzuwenden, wie sie u.a. in den vorstehenden Ausführungen beschrieben wurden. Trotz des allgemeinen Trends zur Computerisierung wird es auch in Zukunft Erhebungen geben müssen, deren Daten im klassischen Sinn gewonnen werden. Mit einem Spruch von dem griechischen Philosophen Heraklit

Πάντα ῥεῖ = Alles fließt

mögen hier die Ausführungen zu diesem Thema abgeschlossen werden!

Literatur-Verzeichnis

- Auswahl -

- Biemer, P., Lyberg, L., Introduction to Survey Quality, 2003, 402 S.
- Brachinger, H.W., Hamerle, A., Münnich, R., Schweitzer, W., Wirtschaftsstatistik, 2006, 382 S.
- Cochran, W.G., Statistical Survey Techniques, 3rd Edition, 1977, 428 S.
- Jessen, R.J., Statistical Survey Techniques, 1978, 520 S.
- Krug, W., Nourney, M., Schmidt, J., Wirtschafts- und Sozialstatistik, 6. Aufl. 2001, 431 S.
- von der Lippe, P., Wirtschaftsstatistik, 5. Aufl. 1996, 516 S.
- Münnich, R., Data Quality in Complex Surveys, in: Eurostat (Hrsg.) NTTS & ETK 2001 Conference proceedings, http:/webfarm.jrc.cec.eu.int. /ETK
- Pokropp, F. Stichproben, Theorie und Verfahren, 2.Aufl. 1996, 247 S.
- Strecker, H., Model for the Decomposition of Errors in Statistical Data into Components and the Ascertainment of Respondent Errors by Means of Accuracy Checks, in: Jahrbücher für Nationalökonomie und Statistik, Bd. 195, 1980, 385-420
- Strecker, H., Wiegert, R., Peeters, J., Kafka, K., Messung der Antwortvariabilität auf Grund von Erhebungsmodellen mit Wiederholungszählungen, in: Angewandte Statistik und Ökonometrie, Heft 25, 1983, 112 S.
- Strecker, H., L'enquête contrôle, un instrument permettant de déceler les erreurs dans les declarations des effectifs et de déterminer la variance des responses, Recensement agricole et horticole au 15 mai 1979 en Belgique, Etudes No.75, Institut national de statistique, Bruxelles, 1985, 56 S., auch in niederländischer Übersetzung
- Strecker, H., Wiegert, R., Die Antwortvariabilität bei statistischen Erhebungen. Wiederholungszählungen und Antwortvarianz, in: Österreichische Zeitschrift für Statistik und Information, 16.Jg. 1986, 99-130
- Strecker, H., in: Beckmann, M./Wiegert, R., Statistische Erhebungen: Methoden und Ergebnisse, Schriftenreihe Angewandte Statistik und Ökonometrie, Heft 30, 1987, 279-280 und 323-325
- Strecker, H., Wiegert, R., Geschichtete Stichproben und Messung der Antwortvariabilität, in: Allgemeines Statistisches Archiv, Bd. 76, 1992, 240-267

- Strecker, H., Wiegert R., Stichproben, Erhebungsfehler, Datenqualität, in: Angewandte Statistik und Ökonometrie, Heft 36, 1994, 246 S., mit ausführlichen Literaturangaben
- Strecker H., Der Netto- und Bruttofehler sowie Beispiele für besondere Fehlerursachen, in: Allgemeines Statistisches Archiv, Bd. 79, 1995, 402-424
- Strecker, H., Inconsistency, Strecker's Index of, in: Encyclopedia of Statistical Sciences, Update vol. 3, 1999, 359-361
- Strecker, H., Zur Schätzung des Inkonsistenz-Index als Maßzahl zur Fehlermessung bei klassierten Erhebungsergebnissen, in: Jahrbücher für Nationalökonomie und Statistik, Bd. 220, 2000, 777-792, mit ausführlichen Literaturangaben.
- Strecker, H., Wiegert, R. avec collaboration de Peeters, J., Anciaux, L., Draelants, E., La variabilité des réponses dand les enqutes statistiques, Estimation théoretique et pratique, Etudes statistiques, No. 106, Institut National de statistique, Bruxelles, 2000, 87 + 14 S., auch in niederländischer Übersetzung
- Strecker, H., Tintner, G., Wörgötter, A., Wörgötter, G., Variate Difference Method, in: Encyclopedia of Statistical Sciences, vol. 9, 1988, 484-488, mit ausführlichen Literaturangaben
- Strecker, H., Der Nicht-Stichprobenfehler und seine Zerlegung in die beiden Komponenten „glatter Wert“ – wahrer Wert plus systematischer Fehler – und „Zufallsfehler“ (Schätzung mit Hilfe der Variate Difference-Methode), in: Jahrbücher für Nationalökonomie und Statistik, Bd. 224, 2004, 198-230
- Strecker, H., Der Zufallsfehler in Gesamt-Ergebnissen statistischer Erhebungen - Vertikale Aggregation der Angabewerte von Erhebungseinheiten, in: Jahrbücher für Nationalökonomie und Statistik, Bd. 224, 2004, 579-611
- Strecker, H., Wiegert, R., Adäquation und neue Systematiken zur Datenqualität in der Statistik, in: Wirtschaftsstatistik, hrsg. von Brachinger, H.W., Hamerle, A., Münnich, R., Schweitzer, W., 2006, 381 S.
- Wiegert, R., Kafka, K., Strecker, H., Steylaerts, R., Über eine optimale Zuordnung von Interviewern zu Erhebungsgemeinden - dargestellt an einem Beispiel aus der Agrarstatistik Belgiens, Bulletin de Statistique, No. 1/2, Jan./Febr. 1977, in: Jahrbücher für Nationalökonomie und Statistik, Bd. 190, 1976, 428-463 - auch erschienen in: Bulletin de statistique, No. 1/2, Jan./Febr.1977, Institut national de statistique, Bruxelles, in französischer und niederländischer Übersetzung
- Tintner, G., Rao, J.N.K., Strecker, H., New Results in the Variate Difference Method, in: Angewandte Statistik und Ökonometrie, Heft 11, 1978, 102 S.