AUSTRIAN JOURNAL OF STATISTICS
Volume 31 (2002), Number 1, 41-53

Bemerkungen zum Zusammenhang von amtlicher und methodischer Statistik

Wilfried Grossmann
Institut für Statistik und Decision Support Systems, Universität Wien

Gerhart Bruckmann zum 70. Geburtstag gewidmet

Zusammenfassung: Die Datenproduktion in statistischen Institutionen ist ein komplexer Vorgang, der die problemgerechte Kombination von Statistik, Substanzwissenschaft und Informatik erfordert. Dieses Zusammenspiel wird am Beispiel der Aggregation von Daten erläutert.

Abstract: The production process in statistical institutions is a rather complex process, which needs a combination of statistical knowledge, substantive knowledge and computer science. The need for such an integrated view is demonstrated for the case of data aggregation.

Schlüsselwörter: Aggregation, Tabellenmanipulation, Amtliche Statistik.

1 Einleitung

Gerhart Bruckmann hat sich stets um eine integrale Sicht der Statistik bemüht und in der Tradition des Institutes an der Universität Wien methodische Statistik und amtliche Statistik als sich gegenseitig beeinflussende Zweige einer Statistik gesehen. Da nicht nur Ämter, sondern auch andere nationale und supranationale Institutionen (z.B. Interessenvertretungen, OECD, Vereinte Nationen) derartige Statistiken produzieren und veröffentlichen, wollen wir im folgenden neben dem Begriff "Amtliche Statistik" synonym die Bezeichnung "Institutionelle Statistik" verwenden. Vielfach entsteht heute der Eindruck, dass methodische und amtliche Statistik eher nebeneinander als miteinander tätig sind und - bedingt durch die Bedeutung der Informatik für die Organisation und Verfügbarmachung der Daten - die institutionelle Statistikproduktion der Datenmodellierung im Sinne der Informatik näher steht als der methodischen Statistik. In den folgenden Ausführungen wollen wir anhand des Beispiels der Tabellenproduktion zeigen, dass nur eine Integration der drei Betrachtungsweisen zu einer sinnvollen Lösung führen kann. In Abschnitt 2 gehen wir kurz auf den Arbeitsprozess der institutionellen Statistik ein und zeigen die Schnittstellen zwischen methodischer Statistik, institutioneller Statistik und informatischer Datenmodellierung auf. Im dritten Abschnitt wird die Frage der Tabellenproduktion behandelt. Die Ausführungen erheben keinerlei Anspruch auf Originalität und stellen nur eine Verbindung von bekannten einfachen Tatsachen dar. Die Erfahrungen des Autors in der Diskussion mit Kollegen aus der institutionellen Statistik im Rahmen einer Reihe von internationalen Projektkooperationen zeigen aber, dass vielfach solche elementaren Überlegungen nicht angestellt werden und offensichtliche Zusammenhänge in Vergessenheit geraten sind.

2 Statistikproduktion in der institutionellen Statistik

In der methodischen Statistik werden Daten meist sehr allgemein als Realisation von Zufallsvariablen beschrieben. Die Struktur der Grundgesamtheit, für die diese Zufallsvariablen definiert sind, spielt dabei keine entscheidende Rolle; wesentlich ist nur die durch diese Abbildung generierte Verteilung der Zufallsvariablen. Dieser Ansatz hat den wesentlichen Vorteil, dass Modellierungstechniken und Analyseverfahren für Verteilungen unabhängig von einem spezifischen Anwendungskontext entwickelt werden können. Dadurch wird die statistische Methodik ein wesentliches Hilfsmittel zur Beantwortung vieler substanzwissenschaftlicher Fragestellungen. Die zur Anwendung notwendigen operativen Aspekte der Datengewinnung und der Berechnung werden üblicherweise ebenfalls "anwendungsneutral" im Rahmen der Stichprobentheorie und der computationalen Statistik behandelt.

Im Gegensatz dazu dominiert in der institutionellen Statistik die inhaltliche Analyse der Daten und die adäquate Repräsentation des Anwendungsaspektes. Traditionellerweise ist der Arbeitsprozess in Statistikinstitutionen durch folgenden Ablauf gekennzeichnet:

1. Ziel der Datenproduktion und Abgrenzung des Untersuchungsbereiches
   Die inhaltliche Definition von statistischen Einheiten, die einer Untersuchung zugrunde gelegt werden, ist eine wesentliche und vielfach auch schwierige Aufgabe der institutionellen Statistik, die im Rahmen der Zusammenführung von verschiedenen Datenquellen immer mehr an Bedeutung gewinnt. Als Beispiel für solche Bemühungen auf internationaler Ebene sei etwa das von EUROSTAT geförderte und kürzlich beendete Projekt CLAMOUR genannt, das sich mit einer detaillierten Analyse des Begriffs Unternehmen beschäftigte (van der Hoeven et al., 2001). Aufbauend auf solchen statistischen Einheiten werden endliche Grundgesamtheiten definiert, deren Beschreibung der eigentliche Gegenstand der institutionellen Statistik ist. Wichtig ist neben dieser inhaltlichen Spezifikation aber auch die operationale Bereitstellung von Grundgesamtheiten, etwa in Form eines Registers.

2. Inhaltliche Definition der zu messenden Zufallsvariablen
   Auch hier spielen substanzwissenschaftliche Überlegungen eine zentrale Rolle, sowohl bei der Bestimmung des Messkonzeptes (z.B. die Definition des Einkommens einer Person) als auch bei der Bestimmung des Wertebereiches der Zufallsvariablen vom inhaltlichen Standpunkt. Letztere Frage ist vor allem deshalb von Interesse, weil die Messgenauigkeit von statistischen Größen immer im Kontext der Problemstellung gesehen werden muss. So einfach und klar diese Frage im Bereich von metrischen Größen zu formulieren ist, so komplex wird sie vielfach für qualitative Variable und erfordert die Arbeit von Experten im Bereich der Entwicklung von Terminologien und Klassifikationen. Die Verfügbarmachung und Dokumentation solcher Klassifikationssysteme ist wieder ein Problem der Informatik. (Vergleiche dazu etwa den RAMON-Server, der Information über internationale Standardklassifikationen bietet: `http://www.un.org/depts/unsd/class/`).

3. Auswahl der in die Untersuchung aufzunehmenden Einheiten
   Traditionell ist dies der Bereich der Stichprobentheorie, die entscheidend auf dem

Prinzip der Randomisierung beruht. Die klassische Betrachtung ist dabei, dass man das Auswahlverfahren als den die Verteilung generierenden Mechanismus ansieht und die Werte der Variablen für die statistischen Einheiten als hochdimensionale Parameter interpretiert. Die starke Betonung dieses Prinzips hat zweifelsohne seine Meriten und bietet viele Vorteile in der Praxis, andererseits hat es dazu geführt, dass die Stichprobentheorie zwar in der statistischen Praxis eine zentrale Rolle spielt, im Rahmen der methodischen Statistik aber vielfach ohne unmittelbaren Bezug zu den anderen Bereichen ist. Auch in der Praxis der institutionellen Statistik scheint die Verbindung zwischen Stichprobentheorie und methodischer Statistik, etwa im Rahmen eines modellbasierten Ansatzes, nicht sehr weit verbreitet (vgl. etwa Valliant et al., 2000).

4. Datenerhebung
   Die Datenerhebung ist heute stark von den technischen Entwicklungen im Bereich der Kommunikationsmedien geprägt. Ausgehend vom traditionellen Fragebogen wurden unterschiedliche computergestützte Erhebungsinstrumente (Computer Aided Survey Information Collection Tools, CASIC Tools) entwickelt, die durch den Einsatz neuer Technologien sowohl den Respondenten als auch den Statistikern eine Reihe von Vorteilen bieten. Ein gutes Beispiel für eine derartige Entwicklung ist der elektronische Fragebogen von Statistik Austria (Koller und Zettel, 2001). Neben diesen Entwicklungen zur Erfassung von neuen Daten hat in letzter Zeit auch der Bereich der Datenkombination zur Informationsgewinnung aus bereits vorhandenen Datenquellen entscheidend an Bedeutung gewonnen. Wesentliche Vorteile dieser Sekundärnutzung von Daten sind vor allem Kostenersparnis und eine Verringerung der Belastung von Bürgern mit Umfragen.

5. Datenanalyse
   Zentrale Aufgaben der Datenanalyse sind das Editing (Entdecken und Korrigieren von fehlenden Werten) und die Behandlung von Nonresponse. Aufbauend auf der Arbeit von Fellegi und Holt konzentrierte sich das Editing primär auf das Auffinden von logisch nicht plausiblen Werten. In neuerer Zeit spielen auch Methoden der explorativen Datenanalyse eine immer größere Rolle (vgl. etwa Bethlehem und van de Pool, 1998). Die Frage der Behandlung von fehlenden Antworten (Datenimputation) wurde vor allem durch das von Rubin entwickelte Modell der multiplen Imputation vom Standpunkt der methodischen Statistik behandelt (Rubin, 1987).

6. Berechnung von Ergebnissen und Analysen
   Die wesentlichen Ergebnisse der institutionellen Statistik sind die Tabelle und die Zeitreihe, wobei die Angabe der statistischen "Qualität" der Daten in Form von statistischen Maßzahlen wie Varianz oder Bias immer mehr an Bedeutung gewinnt. Die Dokumentation dieser Maßzahlen ist dabei meist weit umfangreicher als die Daten selbst (vgl. etwa Kalton et al., 2000, für ein gutes Beispiel einer derartigen Dokumentation aus dem Bereich der Schulstatistik in den USA).

7. Veröffentlichung und Verfügbarmachung der Daten
   Bedingt durch den steigenden Bedarf nach statistischer Information bei wirtschaftlichen und politischen Entscheidungsträgern gewinnt dieser Bereich in der insti-

tutionellen Statistik immer mehr an Bedeutung. Statistische Informationssysteme sollen diesem Bedürfnis Rechnung tragen und die Ergebnisse der Statistikproduktion möglichst rasch und bedarfsgerecht zur Verfügung stellen. Dies bedingt vielfach eine Reorganisation der Arbeitsabläufe innerhalb statistischer Organisationen (Colledge, 1998). Eine neue Qualität der Veröffentlichung ist durch das Internet hinzugekommen.

Diese kurze Beschreibung macht deutlich, dass die Durchführung der einzelnen Arbeitsschritte eine Analyse von verschiedenen Gesichtpunkten erfordert. Im wesentlichen können dabei vier verschiedene Aspekte identifiziert werden:

- *Statistisch-methodologischer Aspekt:*
  Dies ist der Bereich der methodischen Statistik, wie er in der Einleitung kurz dargestellt wurde. Er definiert das statistische Modell für die Daten und damit auch die Logik der Auswertung und ist insbesondere in den Arbeitsschritten 3, 5 und 6 von zentraler Bedeutung.

- *Substanzwissenschaftlicher Aspekt:*
  Dieser Aspekt ist notwendig, um die Brücke zwischen statistischer Methodik und Anwendung herzustellen. Insbesondere in den Arbeitsschritten 1 und 2 ist er von essentieller Bedeutung.

- *Algorithmische und Repräsentationsaspekte:*
  Diese Aspekte schaffen insbesondere in den Arbeitsschritten 4, 6 und 7 die Grundlage für eine effiziente und computergestützte Durchführung.

- *Administrative Aspekte:*
  Diese Aspekte sind für das Management von statistischen Daten und Prozessen notwendig und bestimmen den institutionellen Rahmen der Durchführung.

In den letzten Jahren hat es in den für die amtliche Statistik wesentlichen methodischen Bereichen zweifelsohne Fortschritte gegeben und statistische Modelle haben an Bedeutung gewonnen. Trotz dieser Fortschritte fällt aber auf, dass im Gesamtablauf der Arbeit in statistischen Organisationen der statistisch methodologische Aspekt vielfach in den Hintergrund tritt, und besonders im internationalen Bereich die substanzwissenschaftlichen Aspekte und die Präsentation der Ergebnisse in Form von statistischen Informationssystemen im Vordergrund stehen. Die Entwicklung derartiger Informationssysteme ist primär Aufgabe von Informatikern und statistische Methoden werden vielfach als kodifizierte Bausteine innerhalb des Systems verwendet. Diese Entwicklung führt dazu, dass oft Probleme in Informationssystemen auftauchen, die bei der Behandlung der Frage vom statistischen Standpunkt leicht vermeidbar wären. Ein klassisches Beispiel bei der Produktion von Tabellen wollen wir im folgenden etwas näher betrachten.

3 Das Modell der Aggregation

Statistische Tabellen stellen einen wesentlichen Output der Datenproduktion in der institutionellen Statistik dar. Grundsätzlich versteht man unter einer statistischen Tabelle die

Darstellung einer oder mehrerer Zielgrößen, die nach einer Reihe von Variablen gegliedert werden. Sundgren (1993) hat für die Tabellenproduktion im Rahmen seines *infologischen* Zuganges das sogenannte $\alpha\beta\gamma\tau$-Modell entwickelt, das ein allgemeines Format für statistische Tabellen definiert. Jede Tabelle wird dabei durch die folgenden vier Komponenten beschrieben:

- *Objekt-Komponente (α-Komponente)*
  Diese gibt die interessierende Grundgesamtheit von Objekten an, die auch durch Selektion aus einer größeren Grundgesamtheit eingeschränkt werden kann.

- *Eigenschafts-Komponente (β-Komponente)*
  Diese Komponente definiert einen Parameter oder eine statistische Charakteristik, die sowohl für die Grundgesamtheit als Ganzes als auch für Unterbereiche geschätzt wird. Üblicherweise wird dieser Parameter durch einen Aggregationsoperator bestimmt (z.B. Summe, Durchschnitt, Anzahl), der auf Variable, die für die statistischen Einheiten gemessen wurden, angewendet wird.

- *Kreuzklassifikations-Komponente (γ-component)*
  Diese teilt die Grundgesamtheit in eine Reihe von Subbereichen auf, die üblicherweise durch Variablen definiert werden, deren Werte eine Partition für die Grundgesamtheit definieren.

- *Zeit-Komponente (τ-component)*
  Diese Komponente legt die Zeit fest, zu der (oder während der) die Grundgesamtheit und ihre Subbereiche existierten und die Parameter die angegebenen Werte hatten.

Anhand folgender Tabelle (Ausschnitt aus einer Tabelle im Statistischen Taschenbuch der Stadt Wien, 2001) soll diese Terminologie demonstriert werden.

Tabelle 1: Unselbständig Beschäftigte in Wien, 2001

Wirtschaftstätigkeit	männlich	weiblich
Sachgütererzeugung	61.600	31.839
Bauwesen	48.098	5.650

Die Beschreibung in der $\alpha\beta\gamma\tau$-Terminologie lautet folgendermaßen:
α-Komponente: Personen in Wien
(Register des Hauptverbandes der österreichischen Sozialversicherungsträger)
β-Komponente: Jahresdurchschnitt der unselbständig Beschäftigten
γ-Komponente: Wirtschaftstätigkeit nach ÖNACE 95, Ebene 1 und Geschlecht
τ-Komponente: 2001

Im Sinne der methodischen Statistik beschreibt dieses Modell offensichtlich folgende Situation: Es gibt eine statistisch relevante endliche Grundgesamtheit Ω vom Umfang N. Für eine Teilmenge dieser Grundgesamtheit wird eine vektorwertige Zufallsvariable

(Y, C) bestimmt, wobei $Y = (Y_1, Y_2, \ldots, Y_p)$ metrische Größen oder Indikatorvariable sind und $C = (C_1, C_2, \ldots, C_k)$ ein Vektor von Zufallsvariablen ist, der die Messungen von qualitativen Größen beschreibt. Für die Verteilung der Variablen Y sollen nun sowohl für die Grundgesamtheit als auch für die durch C definierte Partition Parameter geschätzt werden, die mit θ und θ_C bezeichnet werden. Für jede Komponente von Y wird der Schätzer als lineare Funktion der Merkmalsausprägungen definiert und die Berechnung des Parameters als *Aggregation* bezeichnet. Den Operator für die Schätzung des Parameters θ bezeichnen wir mit $\mathcal{A}(Y)$ und den Operator für die Schätzung von θ_C mit $\mathcal{A}_C(Y, C)$. Dabei muss üblicherweise die Verträglichkeitsbedingung $\mathcal{A}(\mathcal{A}_C(Y)) = \mathcal{A}(Y)$ gelten, die der Randsummenbildung in der Tabelle entspricht. In der Praxis sind dies meist Erwartungswerte oder Summen, also Schätzungen von $\theta = \mathbb{E}[Y]$ beziehungsweise $N\theta$ und $\theta_C = \mathbb{E}[Y|C]$ beziehungsweise $N\theta_C$, die in natürlicher Art und Weise diese Verträglichkeitsbedingung erfüllen. Der Erwartungswert kann dabei entweder im Sinne einer durch das Designs generierten Verteilung oder aber im Sinne einer durch ein Modell definierten Verteilung verstanden werden. Falls Y eine eindimensionale Indikatorvariable ist, so entspricht dies der Schätzung der absoluten oder relativen Häufigkeiten in der durch die Klassen gegebenen Partition der Grundgesamtheit.

Im Sinne dieser Beschreibung einer Tabelle ist es auch sinnvoll von *Caselevel-Daten* und *Summary-Daten* zu sprechen:

- Unter *Caselevel-Daten* wollen wir Daten verstehen, bei denen für bestimmte durch die Stichprobe definierte Einheiten der Grundgesamtheit Ω die Werte der Zufallsvariablen (Y, C) bekannt sind.

- Unter *Summary-Daten* verstehen wir die Angabe von Schätzungen von Parametern für eine Grundgesamtheit Ω, die durch eine lineare Funktion definiert sind.

Diese Definition der Tabelle ist eine eher statische, das heißt die Tabelle wird als das Endprodukt einer geschlossenen Produktionskette gesehen. Im Zusammenhang mit statistischen Informationssystemen ist man aber häufig an der Manipulation von Tabellen interessiert. Im Beispiel von Tabelle 1 könnte man etwa an einer Tabelle interessiert sein, die eine feinere Gliederung nach Wirtschaftstätigkeiten enthält, zum Beispiel ÖNACE 95 auf Ebene 3. Hier stellt die Informatik im Rahmen der Datenmodellierung eine Reihe von wesentlichen Hilfsmitteln zur Verfügung. Grundsätzlich unterscheidet die Informatik in diesem Zusammenhang zwischen *relationalen Daten* und *dimensionalen Daten*:

- *Relationale Daten* sind solche, die strukturell durch ein relationales Datenmodell repräsentiert werden können. Im einfachsten Fall ist dies eine Tabelle deren Spalten die Variablen (Attribute) definieren und deren Zeilen durch die beobachteten Tupel gegeben sind. Diese Tupel können durch Schlüsselattribute identifiziert werden. In dieser Datenstruktur können die Operationen der Relationenalgebra angewendet werden, insbesondere Kombination von Daten mit gleichem relationalen Schema, Selektion von Tupel mit bestimmten Eigenschaften, Projektion auf einen Teilraum, der durch weniger Variable bestimmt ist, oder Hinzufügen von neuen Variablen.

- *Dimensionale Daten* sind Daten, bei denen zwischen *Dimensionen* und *Maßzahlen* unterschieden wird: Die Dimensionen sind durch Variable (Attribute) mit einem endlichen Wertebereich definiert, deren kartesisches Produkt einen *Würfel* oder

Cube bildet. Die Maßzahlen definieren aufgrund einer wohldefinierten *Summary-function* ein summarisches Maß für jedes durch die Dimensionen definierte Tupel (Zelle im Würfel). Die Operationen für solche dimensionale Daten ergeben sich im wesentlichen aus den Mengenoperationen für kartesische Produkte, also Vereinigung und Durchschnitt und eine Reihe von weiteren Funktionen, die als OLAP-Funktionen bekannt sind. Beispielsweise kann man in jeder Dimension bestimmte Werte zu neuen Werten zusammenfassen (*roll up*) oder den Wertebereich der Dimensionen verfeinern (*drill down*).

Die strukturellen Korrespondenzen zwischen Caselevel-Daten und relationalen Daten beziehungsweise zwischen Summary-Daten und dimensionale Daten erlauben es, eine Reihe von wichtigen Operationen der statistischen Praxis als formale Datenbankoperationen durchzuführen. Dabei ist das relationale Modell das grundlegende Modell für alle statistischen Programmpakete, während das dimensionale Modell heute allgemein als die Grundlage für statistische Informationssysteme angesehen wird (vgl. dazu auch Shoshani, 1998 für einen detaillierte Analyse des Zusammenhanges zwischen statistischen Tabellen und dimensionalen Daten). Von entscheidendem Vorteil sind die OLAP Operationen dann, wenn die Dimensionen durch Variable erzeugt werden, deren Wertebereiche aus einer Klassifikationshierarchie stammen. Damit wird die Produktion von Tabellen auf unterschiedlichen Hierarchieniveaus zu einem automatisierbaren Vorgang - vorausgesetzt, es handelt sich dabei um additive Summenfunktionen und die Information zur Mittelbildung ist vorhanden. Wir wollen dies an Hand von Tabelle 1 kurz erläutern. Der Würfel wird offensichtlich von den Variablen "Wirtschaftstätigkeit" und "Geschlecht" gebildet und die Werte der Variablen "Wirtschaftstätigkeit" entstammen der Klassifikation ÖNACE 95, Ebene 1. Eine Darstellung nach Wirtschaftstätigkeiten auf einer feineren Ebene (etwa ÖNACE 95, Ebene 3) entspricht genau einem drill down.

Ein derart flexibles Informationsangebot ist für den Endbenutzer zweifelsohne von großem Nutzen, allerdings ist darauf zu achten, dass diese Möglichkeiten immer im Rahmen des zugrundeliegenden statistischen Modells durchgeführt werden. Dieses wird aber im dimensionalen Modell nicht betrachtet, da im Gegensatz zum $\alpha\beta\gamma\tau$-Modell statistisch relevante Größen wie Grundgesamtheit, Verteilung in der Grundgesamtheit und Stichprobe nicht adäquat repräsentiert werden. Welche Konsequenzen dies bei der Manipulation von Tabellen haben kann, wollen wir anhand einer Erweiterung unseres Tabellenbeispiels zeigen.

Gegeben sei neben Tabelle 1 noch die folgende Tabelle 2 über Arbeitsstätten in Wien (Quelle: Statistisches Taschenbuch der Stadt Wien, 2001).

Tabelle 2: Arbeitsstätten in Wien für zwei Wirtschaftstätigkeiten, 2001

Wirtschaftstätigkeit	Anzahl der Arbeitsstätten
Sachgütererzeugung	5.381
Bauwesen	2.884

Die zugehörige Spezifikation im $\alpha\beta\gamma\tau$-Modell sieht folgendermaßen aus:

α-Komponente: Arbeitsstätten in Wien
(aus einem Unternehmensregister oder einer Betriebszählung)
β-Komponente: Anzahl der Arbeitsstätten (an einem Stichtag)
γ-Komponente: Wirtschaftstätigkeit nach ÖNACE 95, Ebene 1
τ-Komponente: 2001

Aus diesen beiden Tabellen lässt sich vom formalen Standpunkt mit dem Operationskalkül für dimensionale Daten problemlos folgende Tabelle erzeugen:

Tabelle 3: Strukturdaten für Wirtschaftstätigkeiten

Wirtschaftstätigkeit	Anzahl der Arbeitsstätten	Unselbständig Beschäftigte
Sachgütererzeugung	5.381	93.439
Bauwesen	2.884	53.748

Die Beschreibung von Tabelle 3 im $\alpha\beta\gamma\tau$-Modell ist bezüglich der β- und τ-Komponenten offensichtlich: die β-Komponente ergibt sich als zweidimensionales Merkmal aus den beiden β-Komponenten von Tabelle 1 und Tabelle 2 und die τ-Komponente ist durch das Jahr 2001 gegeben. Schwieriger wird es mit der α-Komponente, da ja Tabellen von zwei unterschiedlichen Grundgesamtheiten zusammengeführt wurden. Aufgrund der Spezifikation in den Tabellen 1 und 2 scheint eine mögliche Lösung die zu sein, dass wir die Grundgesamtheit, auf die sich diese Tabelle bezieht durch die statistische Einheit "Wirtschaftstätigkeit" definieren. Je nach der gewählten Genauigkeit für die Definition der statistischen Einheit "Wirtschaftstätigkeit" handelt es sich dann bei Tabelle 3 um Summary-Daten für eine Grundgesamtheit (z.B. "Wirtschaftstätigkeiten nach ÖNACE 95, Ebene 3") mit einer γ-Komponente, die durch ÖNACE 95, Ebene 1 gegeben ist, oder aber um Caselevel-Daten, falls wir die Grundgesamtheit durch "Wirtschaftstätigkeiten nach ÖNACE 95, Ebene 1" definiert haben (die γ-Komponente ist in diesem Falle nicht existent). In jedem Falle wurde im Sinne der Terminologie von Sundgren aus einem Teil der γ-Komponente von Tabelle 1 und der γ-Komponente von Tabelle 2 die α-Komponente der neuen Tabelle 3. Dies ist auch sinnvoll, wenn wir Tabelle 3 als Indikatoren für die Wirtschaftstätigkeiten in Wien interpretieren.

Anders liegt die Situation, wenn wir in Tabelle 1 die Daten nicht aus personenbezogenen Caselevel-Daten, sondern ebenfalls aus einer Erhebung an Arbeitsstätten produziert hätten (die α-Komponente in Tabelle 1 wären also ebenfalls die "Arbeitsstätten"). In diesem Falle wäre es auch für Tabelle 3 sinnvoll, die α-Komponente durch "Arbeitsstätten" zu definieren.

Der beschriebene mögliche Wechsel zwischen α-Komponente und γ-Komponente bei der Manipulation von Tabellen wird vielfach als Problem in statistischen Informationssystemen angesehen, der offensichtlich auf der rein operativen Ebene des dimensionalen Kalküls nicht gelöst werden kann. Natürlich könnte man einwenden, dass dieses Problem artifiziell ist und man bereits bei der Planung von Erhebungen so vorgehen sollte, dass nur Tabellen aus gleichen Grundgesamtheiten kombiniert werden müssen (in unserem Beispiel also die Daten von Tabelle 1 auf einer Erhebung für die statistische Einheit

"Arbeitsstätte" beruhen). In der in Abschnitt 2 beschriebenen Realität der institutionellen Statistik ist dies aber kaum möglich: bei Daten, die der Öffentlichkeit zur Verfügung gestellt werden, kann man mit vertretbarem Aufwand nicht alle möglichen Verwendungen vorhersehen. Nebenbei sei angemerkt, dass die oben beschriebene Vorgangsweise der Kombination von Tabellen mit unterschiedlichen Grundgesamtheiten ein Produktionsstandard bei EUROSTAT und anderen supranationalen Statistikorganisationen ist.

Als Ausweg aus diesem Dilemma schlagen viel Praktiker im Zusammenhang mit der Konstruktion von Output-orientierten statistischen Informationssystemen vor, eine neutrale statistische Einheit zu wählen (in unserem Falle wäre dies vermutlich die Wirtschaftstätigkeit), oder vom sachlogischen her zu entscheiden, welche Einheit die bessere ist (vgl. Willebordse et al., 2001). Dieser Vorgang scheint aber wenig befriedigend, weil ja offensichtlich mögliche Unterschiede in den Ausgangssituationen nicht erfasst werden können. Vernünftiger scheint es, das operative Modell mit einem statistischen Modell zu kombinieren, also die Beschreibung im Sinne des $\alpha\beta\gamma\tau$-Modells in die Operationen des dimensionalen Modells zu integrieren.

Wesentlich Voraussetzung dafür ist eine klare Unterscheidung zwischen den Operationen für die Grundgesamtheiten oder besser für die Ereignisstruktur in der Grundgesamtheit, die implizit mit der Tabellenproduktion verbunden sind, und den Operationen für die Datensätze selbst. Dazu betrachten wir zunächst nur die Zufallsvariablen, welche die Kreuzklassifikation (γ-Komponente) der Tabelle betreffen. Ist $C : \Omega \to D^k$ diese Zufallsvariable mit einem beliebigen endlichen Wertebereich D^k, so definiert diese Abbildung eine neue Ereignisstruktur in Ω, die von den Urbildern $C^{-1}(\{(d_1, d_2, \ldots, d_k)\})$ der Abbildung C erzeugt wird. In der Praxis wird dabei oft zwischen dem Fall von qualitativen Variablen, welche die Kreuzklassifikation erzeugen (wie im Beispiel von Tabelle 1 und 2) und dem Fall einer Zuweisungsfunktion (zum Beispiel Personen werden Haushalten zugeordnet) unterschieden, doch ist dies für die Struktur nicht wesentlich. Wesentlich ist, dass diese neue Ereignisstruktur eine Äquivalenzrelation für die Einheiten der Grundgesamtheit erzeugt. Diese Äquivalenzklassen können wir zu einer neuen Grundgesamtheit Ω^* zusammenfassen und eine Abbildung $T : \Omega \to \Omega^*$ definieren, die jedem Element von Ω die entsprechende Äquivalenzklasse zuordnet. Sind nun $Y : \Omega \to \mathbb{R}^p$ die ursprünglich gegebenen Zufallsvariablen und $A_C : \mathbb{R}^p \times D^k \to \mathbb{R}^p$ zum Aggregationsoperator $\mathcal{A}_C$ gehörende Abbildung, die jedem Tupel $(d_1, d_2, \ldots, d_k)$ den Wert der Parameterschätzung zugeordnet, so stellt sich die Frage nach der "Projektion" der Zufallsvariablen Y auf die neue Grundgesamtheit Ω^*, die mit der Aggregationsoperation kompatibel ist. Wir suchen also jene Funktion Y^*, die das Diagramm in Abbildung 1 kommutativ macht.

$$\begin{array}{ccc} \Omega & \xrightarrow{(Y,C)} & \mathbb{R}^p \times D^k \\ {\scriptstyle T}\downarrow & & \downarrow{\scriptstyle A_C} \\ \Omega^* & \xrightarrow{Y^*} & \mathbb{R}^p \end{array}$$

Abbildung 1: Funktionsstruktur der Aggregation

Bei gegebener Kreuzklassifikation und dazugehörigem Aggregationsoperator $\mathcal{A}_C$ ist diese Funktion in natürlicher Art und Weise durch die Werte von $\mathcal{A}_C(Y)$ gegeben. Im Falle des Erwartungswertes entspricht dies bekanntlich der Definition des bedingten Erwar-

tungswertes.

Geht man in Abbildung 1 von gegebenem (Y, C) und A_C aus, so entspricht das der traditionellen Betrachtung einer Tabelle und Ω^* und die Transformation T können vernachlässigt werden. Man kann in dem Diagramm aber auch den Output-orientierten Weg wählen und von gegebenem Y und bekannter Transformation $T : \Omega \to \Omega^*$ ausgehen und sich fragen wie Y^*, C und A_C zu definieren sind, damit das Diagramm kommutativ wird. Eine Lösung ergibt sich in natürlicher Weise aus der Anwendung der Überlegungen, die zur Definition von Ω^* führten: Man betrachtet die von T erzeugte Ereignisstruktur in Ω, also die Urbilder $T^{-1}(\{\omega^*\})$ der Elementarereignisse in Ω^*. Diese definieren nun implizit eine Kreuzklassifikations-Variable C auf Ω und wir können bezüglich dieser Variablen C einen Aggregationsoperator $\mathcal{A}_C$ und damit auch die Abbildungen A_C und Y^* definieren.

Die Frage des Wechsels zwischen der α-Komponente und der γ-Komponente ist somit nur eine Frage der Betrachtungsweise des obigen Diagramms. Wenn wir von Y ausgehen und nur die Abbildung A_C betrachten, so ist die α-Komponente durch die Grundgesamtheit Ω gegeben und es liegen Summary-Daten bezüglich der Kreuzklassifikation C vor. Wenn wir hingegen nur die Abbildung Y^* auf Ω^* betrachten, so handelt es sich um Caselevel-Daten für Ω^*. Ebenso ist die Wahl der statistischen Einheit (α-Komponente) bei der Zusammenführung von Tabellen nicht willkürlich, sondern in Abhängigkeit von der vorhandenen Information in Abbildung 1 zu bestimmen. Liegen zwei Tabellen mit gleicher Kreuzklassifikation C und gleichem Aggregationsoperator $\mathcal{A}_C(Y_1)$ und $\mathcal{A}_C(Y_2)$ vor, und sind die Grundgesamtheiten gleich und bekannt, so erzeugt der Aggregationsoperator $\mathcal{A}_C(Y_1, Y_2)$ Summary-Daten für Ω. Sind die Grundgesamtheiten der beiden Tabellen unterschiedlich oder unbekannt, so erzeugt die Kombination Caselevel-Daten $({Y_1}^*, {Y_2}^*)$ für die durch C definierte Grundgesamtheit Ω^*. In der Praxis der Publikation von Tabellen wird auf diesen Sachverhalt meist durch Fußnoten hingewiesen. Natürlich gibt es auch noch die Möglichkeit, die unterschiedlichen Grundgesamtheiten durch ein entsprechendes Modell zu harmonisieren. Im Beispiel von Tabelle 1 und Tabelle 2 könnte man etwa durch ein Modell aus Tabelle 1 die Beschäftigten an den Arbeitsstätten schätzen.

Eine Tabellenberechnung ohne die Betrachtung der zugrundeliegenden Ereignisstruktur kann zu einer Reihe von Anomalien und sinnlosen Aussagen führen. Insbesondere gilt dies, wenn man Tabellen aufgrund von nicht disjunkten Gruppenbildungen erzeugt. Dies ist aufgrund der obigen Ausführungen vom methodischen Standpunkt her selbstverständlich, im Rahmen des rein dimensionalen Kalküls werden aber anstelle dieser statistisch natürlichen Verbindung oft sehr komplexe, inhaltlich motivierte Strukturbedingungen formuliert, oder aber pragmatische Lösungsvorschläge gemacht. Wir wollen die Problematik dieser pragmatischen Vorgehensweise anhand eines Beispiels aus der oben zitierten Arbeit von Willeboordse et al. (2001) etwas genauer analysieren. Die Autoren gehen dabei von folgendem (fiktiven) Beispiel aus:

In einer Umfrage über musikalische Präferenzen wurden 1000 Personen über ihre Vorlieben für die Musikrichtungen ("Genre") "Jazz", "Klassik" und "Pop" befragt, wobei Mehrfachantworten möglich waren. Es handelt sich also um eine Erhebung in einer Grundgesamtheit von Personen mit den drei dichotomen Variablen V_J, V_K und V_P. Um die Ergebnisse tabellarisch darzustellen, geben die Autoren drei Möglichkeiten an. Die erste "kanonisch" bezeichnete Methode entspricht wohl der Betrachtung von (Y, C) und A_C in Abbildung 1: Es werden für die Grundgesamtheit "Personen" die Ergebnisse der

Umfrage gegliedert nach der "Genre" wie in Tabelle 4 dargestellt. Aufgrund der Gesamtsumme 1220 konstatieren die Autoren einen Konflikt zwischen der statistischen Einheit "Person" (α-Komponente) und der Kreuzklassifikation "Genre" (γ-Komponente), die sich aus den Mehrfachantworten ergibt. Im Sinne des statistischen Modells ist diese Tabelle ja auch nicht zulässig, da ja die musikalische Präferenz eine dreidimensionale Variable $V = (V_J, V_K, V_P)$ ist und die Ereignisstruktur, die durch die Urbilder von V erzeugt wird, keine Partition in der Grundgesamtheit "Personen" bildet.

Tabelle 4: Musikalische Präferenzen von Personen

Genre	Anzahl der Personen
Jazz	240
Klassik	360
Pop	620
Gesamt	1220

Zur Lösung dieses Konfliktes werden nun zwei Alternativen vorgeschlagen. Die erste "konservative" Methode entspricht der Betrachtung von Y^* in Abbildung 1: Ausgehend von der statistischen Grundgesamtheit "Stimmen" wird eine neue Grundgesamtheit "Musikalische Präferenz" mit den drei statistischen Einheiten "Jazz", "Klassik" und "Pop" definiert, und für diese Einheiten werden die Antworten dargestellt. Dies liefert als Ergebnis folgende Tabelle:

Tabelle 5: Stimmen gegliedert nach Genre

Genre	Anzahl der Stimmen
Jazz	240
Klassik	360
Pop	620
Gesamt	1220

Diese Tabelle ist vom formalen Standpunkt zwar korrekt, da durch diesen Kunstgriff die dreidimensionale Variable (V_J, V_K, V_P) auf der Menge der Personen zu einer eindimensionalen Variablen "Votum" mit den drei möglichen Werten "Jazz", "Klassik" und "Pop" auf der Grundgesamtheit "Stimmen" umdefiniert wurde und "Genre" eine disjunkte Ereignisstruktur in der Gesamtheit aller Stimmen erzeugt. Allerdings ist der Zusammenhang zu der eigentlichen Erhebungseinheit "Person" unklar, da die Zuordnung von der Grundgesamtheit "Person" auf die Grundgesamtheit "Stimmen" nicht eindeutig, also keine Funktion, ist.

Die zweite als "progressiv" bezeichnete Methode besteht darin, für die Grundgesamtheit "Personen" Kombinationsklassen der Antworten zu definieren und die Ergebnisse der

Erhebung dementsprechend auszuzählen. Dies liefert im Beispiel die in Tabelle 6 dargestellten Ergebnisse für die statistische Einheit "Person". Strukturell entspricht dies der Definition eine neuen Grundgesamtheit

$$\Omega^* = \{\text{Jazz, Nicht-Jazz}\} \times \{\text{Klassik, Nicht-Klassik}\} \times \{\text{Pop, Nicht-Pop}\}$$

und einer Transformation $T = (V_J, V_K, V_P)$ von Ω nach Ω^*. Diese erzeugt in Ω genau die in der Tabelle angegebenen Äquivalenzklassen (offensichtlich kommt die Möglichkeit "Weder Jazz noch Klassik noch Pop" in den Antworten nicht vor). Die Gliederung der Personen in dieser Tabelle entspricht genau der durch die dreidimensionale Variable erzeugten Ereignisstruktur, ist also weder konservativ noch progressiv, sondern einfach statistisch korrekt.

Tabelle 6: Musikalische Präferenzen von Personen

Genre	Anzahl der Personen
Jazz (alleine)	100
Klassik (alleine)	200
Pop (alleine)	500
Jazz und Klassik	80
Jazz und Pop	40
Klassik und Pop	60
Jazz, Klassik und Pop	20
Gesamt	1000

Interessant ist in diesem Zusammenhang auch die Argumentation der Autoren bezüglich der Verwendbarkeit von Tabelle 5. Diese wird als sinnvoll erachtet, wenn jemand am potentiellen Markt für CDs eines bestimmten Genres interessiert ist. Dies ist vielleicht vertretbar, wenn die Grundgesamtheit aus genau den 1000 befragten Personen besteht. Wenn man die Kaufwahrscheinlichkeiten bei Einfachnennungen und Mehrfachnennungen als gleich hoch ansieht, könnte man aus Tabelle 5 als Schätzung für den potentiellen Markt für eine CD des Genres "Pop" 620 Personen angeben. Wenn man aber die Daten als eine Stichprobe deutet, so wird für die Beantwortung dieser Frage sicher nicht die Absolutanzahl der Vorlieben interessant sein, sondern der Anteil der Personen, die eine bestimmte CD kaufen. Unter der Voraussetzung gleicher Kaufwahrscheinlichkeit (unabhängig von Mehrfachinteressen) ergibt sich aus Tabelle 6 ein geschätzter Anteil von 62% von potentiellen Käufern von Pop CDs. Aus Tabelle 5 kann ohne Zusatzinformation über die Struktur der Mehrfachnennungen praktisch keine Schätzung gegeben werden. Bei mangelnder Dokumentation von Tabelle 5 könnte jemand auf die Idee kommen, als Schätzung den Wert $620/1220 = 50,82\%$ zu verwenden, was aber deutlich von der Schätzung aus Tabelle 6 abweicht.

Literatur

J. Bethlehem and F. van de Pol. The future of data editing. In M.P. Couper et al., editors, *Computer Assisted Survey Information Collection*, pages 201-222, Wiley Series in Probability and Statistic New York, 1998.

M.J. Colledge. Statistical integration through metadata management. *International Statistical Review*, 67:79–98, 1998.

G. Kalton, M. Winglee, S. Krawchuk, and D. Levine. *Quality Profile for SASS Rounds 1-3: 1987-1995*. National Center for Education Statistics, Washington, 2000.

W. Koller and G. Zettel. e-Quest: A metadata-based software for electronic raw data collection at Statistics Austria. In P. Nanopoulos, D. Wilkinson, and N. Karavitis, editors, *New Techniques and Technologies for Statistics*, pages 143–154, European Communities (ISBN 92-894-1176-7), 2001.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, 1987.

A. Shoshani. OLAP and statistical databases: Similarities and differences. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 185–196, ACM, 1997.

B. Sundgren. Statistical metainformation systems – pragmatics, semantics, syntactics. *Statistical Journal of the UN ECE*, 10:121–142, 1993.

R. Valliant, A.H. Dorfmann, and R.M. Royall. *Finite Sampling and Inference*. Wiley Series in Probability and Statistics, New York, 2000.

J. van der Hoeven, H. van Hooff, B. Kroese, R. Lok, P. Struijs, and Ad Willeboordse. A Model of the Structure and Activities of Business. Report Statistics Netherlands, Division Research and Development (TMO-102144), 2001.

Ad Willeboordse, C. van Duin, and J.W. Altena. Theme Building by The Art of Cubism. Statistics Netherlands, Division Technology and Facilities (TMO-1819-01), 2001.

Adresse des Autors:

Univ.Prof. Dr. Wilfried Grossmann
Institut für Statistik und Decision Support Systems
Universität Wien
Universitätsstraße 5
A-1010 Wien

Tel. +43 1 4277/ 38610
Fax +43 1 4277 9386
E-Mail: wilfried.grossmann@univie.ac.at