社会調査の管理および分析支援システム SSMS*

横山和典
椿康和

1. はしがき

人文科学や社会科学における諸研究，行政ニーズの把握，企業の市場調査などを目的とするデータ収集の手段として，社会調査は重要な役割を果たしている．

社会調査は一般に，

1．問題点の提起，
2．企画と準備，
3．現地調査，
4．調査結果の整理，
5．集計と作表，
6．分析と報告，

という手順により実行される．このうち4．以下のいわゆるデータの処理の中で，集計や統計計算などある程度定型化され，処理法が確立されている分野の作業は，SPSS や SAS などの統計プログラムパッケージを適用することにより，かなり容易なものになってきている．しかし，その前段階で必要とされるデータの浄化や加工，編集などのいわゆるデータ整理の

---

* 本システムは筆者両名の共同制作によるものである．SAS プログラムは横山が，システム管理と FORTRAN プログラムは椿がそれぞれ主として担当した．このシステムの概要は第53回日本統計学会（昭和60年7月）において報告し有益な助言を受けた．なお，数多くの示唆を受けた本学総合情報処理センターにおける「統計データベース研究会」の諸氏に深甚の謝意を表する．

作業には，なお数多くの問題が残されており，現在その効率化が強く要請されている．

これらの問題の解決を遅らせている主たる原因としては，

1．調査データの形式が多様であること，

2．調査データの処理に不可欠な設計情報と実データとの関連づけが貧弱であること，

などが挙げられる．調査データの形式が多様化するのは，調査票における質問項目の選定や配列，質問に対する回答形式の選定，コーディング形式の選定などのあり方に起因している．従来これら諸事項の決定は，もっぱら調査目的や調査対象との適合性，調査の客観性・妥当性などに主眼がおかれ，他方データ処理の容易さや分析手法との関連にはほとんど配慮がなされず，結局データ処理が必要とされる段階になってはじめてデータに適合する処理法が検討されてきている．このような調査票作成の姿勢からみて，データ処理が極めて個別的で場当たり的な作業とならざるを得なかったのは当然といえよう．

データ形式の変換やデータ処理には，

1．データ形式を決定するに至った調査設計に関する情報，

2．目的とするデータ処理に適したデータ形式に関する情報，

3．その処理手続きに関する情報，

が必要である．これらは一応企画書，調査票，分析マニュアルなどの形で提供されてはいるものの，それらの間に有機的関連性は存在せず，直ちにデータ処理に結びつけることは困難である．結局分析者は実データと対応づけながら，個々ばらばらに処理しているのが現状である．

調査データの処理における中心課題は，上述のような状態を若干でも定型化することである．そのためには，

1．データ処理の立場から見た望ましい調査票やコーディングの形式を明確にし，調査データの管理方法を確立すること，

2．この作業の中に含まれるデータ処理の手続きと，それに必要な調査設計情報を実データに対応づけること，

により，データの処理作業を標準化，一般化して論理的なものにしなければならない．

近年，統計データを管理するための DBMS(データベース管理システム)の研究が盛んに行われるようになってきたが，集計データに比べて調査データは，上述の諸理由に由来するデータ構造の一般化の困難さやその共用度の低さなどから，研究対象とされることが少ない．しかし，調査データであってもデータの収集と分析の主体は必ずしも同一ではなく，共用を前提としたデータ管理の必要性がないわけではない．また分析時の試行錯誤の過程においては同じデータを種々の視点から自由にとり出す操作が求められており，これらに関して DBMS が問題を解決する極めて有力な手段であることには変わりがない．

とはいえ，調査データは管理およびその利用の面において特殊な性質を持っているから，それを一般のデータを対象とした既存の DBMS あるいは統計データを対象にした DBMS (例えば HSDB[2])の管理下にそのまま置くことはできない．これが調査データの構造や処理に適したシステムの開発が必要とされる所以である．

SSMS (Social Survey Management and analysis Support System) はこのような認識と要請の下で調査データの管理と分析を支援することを目的として開発され，現在，広島大学総合情報処理センターに登録されたアプリケーション・パッケージとして公開されている[1]．なお，処理対象は調査データに限定されず，一般の行列形式のデータならば十分処理可能な機能も持っている．

本稿では，SSMS の基本構想を中心にその機能概要を論ずる．その実際の利用にあたっては，同センター発行の利用の手引「調査・統計データ分析支援システム SSMS」を参照されたい．

---

1) 本システムの中心的部分は SAS マクロ言語で，ユーザインターフェース等は FORTRAN で作成されている．

2. 調査データの構造とその特徴

2.1 調査データの要素と集計データセット

調査データは調査票中の質問項目に対する回答の記入によって発生し，そのデータ要素は調査票の構造によって定まる．質問の形式はさまざまな立場から類型化が行われているが，発生するデータ要素に注目すれば，

1．自由回答形式

2．多項選択形式

3．順位回答形式

に大別することができる．

自由回答形式は事前に回答のカテゴリ化が困難な場合に用いられ，回答者に自由に記入させるものである．記入内容によって，数値で回答させる数値記入形式と語句や文章で回答させる語句記入形式とがあり，前者はそのまま集計しうるが，後者は整理に多大の時間と労力を必要とし大量のデータ処理には不適であることから，集計の対象からはずされることが多い．

多項選択形式は，あらかじめ複数個の選択カテゴリを用意しておき，その中から回答を選択させるものである．これには，選択カテゴリの中から1つを選択させる単一選択法と，複数個を選択させる複数選択法とがあり，さらに後者には選択個数を制限するもの（制限複数選択法）と制限しないもの（無制限複数選択法）とがある．単一選択法では質問に対して1つの値が得られ，複数選択法では1組の値が得られる．これらの値は各カテゴリを表すが，あらかじめ順位や得点の意味を持たせることもできる．

順位回答形式は用意された複数個の選択カテゴリに回答者が順位をつけるものであり，カテゴリ全部に順位をつける完全順位法とその一部に順位をつける部分順位法がある．この場合には順位づけられた値の組がデータとして得られる．

このように，調査データの要素は語句記入形式を除けば，

1．自由回答による数値（N），

2．単一選択による1つの値（S），

3．複数選択による値の組（M），

4．順位回答による順位つき値の組（R），

に類別される．また調査では，質問とはやや性格を異にするが，被調査者の属性，例えば，性別，年齢，職業等を調査し，統計表の比較対照の資料にするのが普通である．しかし，これらは通常数値または記号で表現されるから，データの観点からは上の分類N又はSのいずれかに属するものである．

ところで，統計表を作成するデータは，上記の調査データそのものとはやや異なり，データの各項目が，

1．分類に使用されるもの（以下，C変数とよぶ），

2．集計又は解析の対象になるもの（以下，A変数とよぶ），

のいずれかでなければならない．調査データから，このデータを得るためには若干の処理・加工が必要である．また，本来A変数と定義されるものも，階級化によりC変数として使用することもできる．

統計表の中で分類表は，外部表現を別にして，概念的には次のような構造を持っている[2]．まず複数個のC変数の値を組み合せて調査対象を部分集団に分割し，各部分集団に属する件数，A変数の集計量（総和，平均など）がその部分集団の属性として対応づけられる．C変数1個による場合が一元分類，2個，3個のときをそれぞれ二元分類，三元分類，また2個以上をまとめて複合分類という．これを資料行列としてみるとき，C変数の値の組が観測点，すなわち行の識別子であり，対応づけられた統計量がその属性である．この作業を一般に集計といい，またこのデータを集計データセットとよぶことにする．度数分布表，クロス表などの統計表はすべてこの集計データセットからの一種の外部表現である[3]．

2.2 調査データ固有の処理

以下の諸問題は集計前に処理しておかなければならない調査データ固有

2）SSMS では系列表については特別の扱いをしない．

3）この点については，関係データベースにおける種々の同値な変換に関するKent[1]の指摘を統計表形式のデータに適用して横山・椿[3]が既に論じている．

のものである.

(1) 関連質問による集計部分集団の設定

調査票の中には，先行する質問に対する回答結果にしたがって，被調査者の一部のみが回答すべき質問（関連質問）がある．このような質問項目が存在するときには調査票全体を一律に集計することができず，回答に応じて逐次集計集団を設定し，集団ごとの集計に備えなければならない．そうしなければ，本来その質問に答えるべきでない回答者のものまでその集計に含まれてしまうからである．この作業では質問項目と対象集団との関係を明確にしておかなければならず，特に関連質問の多い調査票ではそれらの情報の結合をシステム化することが要求される．

(2) 無効回答および無効票の扱い

調査データの処理が他の統計データと大きく異なる点は，原始データを直接の処理対象としたデータ浄化の必要性があることである．データには回答者の誤記入，コーディング・シートへの転記ミス，パンチ・ミス等による誤った値が数多く含まれている他，データ作成者の設計ミスによって単なる無回答と真の無効回答とが区別し難いものもあり，これらは集計前に十分に処理しておかなければならない．その処理には個々の回答項目ごとにデータの正当性をチェックする比較的単純なものから，矛盾回答の検出や主要な回答項目に占める無効回答の比率による無効票の決定等のように複数の回答項目にまたがるものまである．

(3) 複数選択による回答の処理

このタイプのデータをコーディングする際，無制限複数選択法形式では各選択肢に対応させて回答項目をとり，0(非選択)，1(選択)の値を割り当てて集計を容易にする．他方制限複数選択法形式では選択しうる上限の個数だけ回答項目をとり，選択されたコードだけを記入することが多い．この形式はコーディングの手間とデータ量を減らすのに有効であるが，その反面集計時に各データ項目を単独に処理することができず，それらを一括して処理しなければならないデータとなる．統計プログラムパッケージの中にあるそのような機能（例えば SPSS の MULTI RESPONSE コ

マンド）を利用することもできるが，それよりも前者のような形にデータを変換しておく方が望ましい．それは，各選択肢を単独に集計対象としうるため，分析の幅が拡がるからである．

(4) 選択カテゴリの再分類

設計時に定められた質問中の選択カテゴリは不変なものではない．集計した結果から設定が不適当であったことが明らかになり，その再分類を余儀なくされることもある．例えば，特定のカテゴリに選択が集中し，他のカテゴリの度数が極めて低くなれば，集計値の精度にばらつきが生ずる．このような場合には若干のカテゴリを合併して新しいものを設定した上で再集計しなければならない．また，分析視点の変更から再分類が必要となることもある．

SSMS ではこれらの問題のうち，(1)～(3)は，調査設計情報とデータ処理手続きとを有機的に関連づけることによって解決している．すなわち，(1)については調査設計情報の一部として集計部分集団を明確に定義させ，それによるデータの把握を可能にしている（5.1節を参照）．また(2)，(3)については，データの整合性のチェックや無効票の決定，データ形式の変換に調査設計情報を用いることにより，それぞれに必要な処理手段を容易に提供できるようにしている（6.章を参照）．他方(4)については，分類カテゴリの新規定義とそれによるデータの再分類化の機能が汎用的なデータ処理手続きの１つとして用意されている（7.2節を参照）．

3. SSMS の基本構想

SSMS は図3.1に示すシステム構成をとる．

3.1 SSMS ライブラリ

統計データベース・システムでは DD/D(データ辞書）がデータの管理に極めて重要な役割を果たすことはさまざまな観点から指摘されている．また，そこに記述される情報も，ファイルやフィールドなどのデータ・エレメントに関するメタ・データにとどまらず，データの処理方法をはじめとする DBS をとりまく環境全体に拡張される傾向にある．

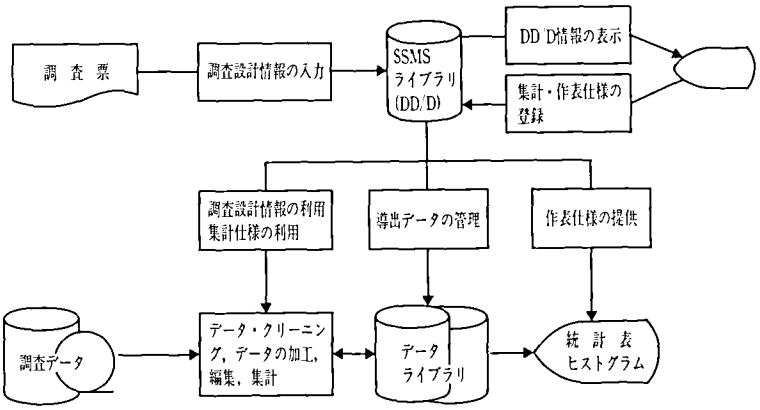

図 3.1 SSMS の構成

調査データを処理するシステムにおいてもまったく同様で，DD/D やそれに類する機能を持つものは，やはり情報管理の有力な手段であり，それは当然システムの中核に置かれるべきものである．このような認識の下で，SSMS では情報管理に用いる SSMS ライブラリを設け，以下の情報を格納している．

(1) ディレクトリ

システムに格納した調査設計情報，実データ，集計・作表に関する各種情報の一覧表．これらはシステム操作の円滑化のために用いられると同時に，利用者にシステムの状況を全般的に把握させるためにも使用される．

(2) 調査設計情報

一般のデータを対象とした場合，システム内に存在するすべてのデータは，意味的な 1 つの実体集合 (entity set) に対応する複数のファイルと，それらの属性 (attributes) であるフィールドという観点から論理的な概念定義がなされ，管理されるのが普通である．統計データの場合，集計データであれば，これらとそれから導出されるデータにはそれぞれ現実世界に存在する各種の統計表を対応させることができるから，同様の方法でデー

タを表現できる．しかし，調査データではやや性格を異にする．

調査データの場合，同様な形で概念定義を明確になしうるのは，調査票のイメージを表現したファイルとそれに含まれる各調査項目に対応するフィールドだけであり，また調査間に共通する概念規定も実際上困難である．それは，各データがサンプリングを経て現実世界からデータの世界へ写像されてくるため，仮に同一名称のものであったとしてもそれらは異なった対象をさすかもしれないからである．このような理由から各調査データは，たとえ対象は同一であっても，それぞれ別個の世界をなすものと理解せざるをえない．またこの世界では，調査票のファイル以外の情報はすべてそこから導出されるものである．SSMS では，各調査ごとに独立したデータ管理を行い，その中心的存在として調査票のファイルを位置づけ，さらに調査データの構造とデータ処理法とは密接に関連していることから，この両者も一体的に管理する．

SSMS に必要な入力情報は次のとおりである．

1．調査データに関する記述的情報，

2．調査票の定義情報，

3．データのコーディング形式に関する情報．

1. はデータ全体の概要を記録するためのものである．2. はデータ構造の管理とデータ処理に必要な情報を提供するためのものであり，関連質問法の採用によって必要となる集計部分集団化のための情報と，個々のデータ項目に関する属性情報とに分けられる．3. はシステムの外部からデータを入力する際のフォーマットに関するものである．これらが入力されると，システムはこれらの情報に基づいて以後のデータ処理に必要な情報や調査データの構造を外部に表示するための情報を自動的に作成し，格納する．

このように管理すべき情報を明示することにより，システムは利用者に対してデータ処理に関してある種のガイダンスを与えることになり，また，これらの情報は同種の調査を設計する際の参考資料としても活用される．

(3) データ作成情報

調査データの分析は，大規模な原始データ全体を対象とするデータの浄

化や全般的分析と，そこから様々な形で出される数多くのサブファイルに対するより詳細な分析という順序で実行されるのが普通である．これらのサブファイルの作成元や作成の経過に関する情報は，システムの状態を知るうえで極めて有用であるが，分析者がいちいち管理することは容易なことではない．このため SSMS ではこの種の情報をシステムで自動的に作成・管理し，随時表示可能な状態にしており，分析者の負担の軽減を図っている．

(4) 集計・作表仕様

集計や作表は対象となるデータを変えつつ同一仕様に基づいて反復的に実行されるのが普通である．また，高度な表現力を持つ統計表を作成するためには与えるべき情報も必然的に多くなるから，情報の入力と集計・作表とを切り離した方がよい．このような観点から SSMS では，これらの情報をシステムで個別に保存し再利用を可能にする方法を採用しており，これにより作業全体の大幅な効率化を図っている．

3.2 データライブラリ

SSMS ではデータの入出力，加工・編集などの操作をすべて TSASDS とよぶ一時的ファイルの上で行う．これは各種のデータ処理を連続的かつ効率的に行うために設定したものである．

一方，処理結果を保存する恒久的ファイルとしては調査データ全体を格納するマスタライブラリと，それから導出された当面の分析に必要な部分のデータや加工された結果を格納する導出ライブラリの２種類が用意されており，データはそれぞれのメンバとして管理される．この区分は，先に述べた調査データの一般的な処理手順を考慮したことと，大容量になりがちなライブラリを長期にわたってすべて保存することの不経済性や利用環境上の制約からなされたもので，利用者は必要に応じて使い分けることができる．

また，システムに対するデータの入力は一般にデータ量が多くなることを考慮して，端末からの直接入力を認めず，外部ファイルからのみの入力に限定している．なお，SSMS ライブラリおよびこれらは，処理速度の

表 3.1 コマンド一覧

機能		コマンド	サブコマンド	内容
分野の選択		PLAN		PLAN セクションへの移行
		DATA		DATA セクションへの移行
		TABLE		TABLE セクションへの移行
ライブラリの管理（自由コマンド）	DD/D 情報の表示	LISTD		ディレクトリ情報
		LISTP		調査設計情報
		LISTS		集計・作表仕様
		LISTH		データセット作成情報
		LISTV		データセットの変数名
	データの出力	PRINT		データの端末出力
		LP		データのラインプリンタ出力
	ライブラリの操作	SAVE		データセットのライブラリへの格納
		NOSAVE		データセットの非格納
		DELM		データセットの削除
		COPY		データセットの複写
	システムの操作	HELP		使用できるコマンドとパラメータの意味の表示
		RETURN		各処理の終了と上位コマンドへの復帰
		END		SSMS の終了
調査設計情報 (PLAN)		PIN		調査設計情報の入力
		PCAT		調査設計情報のライブラリへの登録
		DFORM		データのコーディング形式の登録
		DFORM (NOV)		簡略法による調査設計情報の作成
データの操作 (DATA)	データの入力	DIN		外部ファイルからのデータ入力
	データ・クリーニング	CLEAN	BLUNDER	無効回答の検出
			ZERO	バイナリ回答の0置換
			BLANK	欠損値の確定
			MODIFY	回答の修正と矛盾回答の検出
	データの加工	PROC	MA	MA→SA 展開
			INVALID	無効票の決定
			COMPUTE	回答項目間の演算
			WEIGHT	ウエイト変数の定義
			RENAME	変数名の変更
	データの編集	VIEW	SELECT	条件式による個票の選択
			USELECT	回答集団による個票の選択
			MSELECT	無効回答比率による個票の選択
			PROJECT	回答項目の選択
			UPROJECT	回答集団による回答項目の選択
	データセットの合併	UNION		データセットの縦の連結
	データセットの結合	JOIN		データセットの横の連結
集計・作表 (TABLE)	集計・作表仕様の登録	SPEC	TFORM	値ラベル等の登録
			GSPEC	表の一般的情報の登録
			TSPEC	表の仕様の登録
			SUM	集計情報の登録
	集計・作表仕様の削除	DELS		集計・作表仕様の削除
	集計	AGGREG	RECLASS	変数値の再分類
			SUMMARY	サブグループごとの集計
			NORMAL	集計データセットの標準化
	作表	DISPLAY	FREQ	度数分布表，クロス表の作成
			HIST	ヒストグラムの作成
			TABULATE	登録した仕様による作表

速さとより進んだデータ分析手法の適用を容易にするために，物理的にはすべて SAS ファイルを採用している．

3.3 コマンド（機能）

SSMS はすべて端末から入力するコマンドにより操作される．コマンドの構成は次のような機能体系から成る．

1．自由コマンド群

作業中に随時使用できるコマンドで，DD/D 情報の表示，データの出力，ライブラリの操作，システムの操作などの機能を持つ．

2．PLAN コマンド群

調査設計情報をシステムに入力し，その情報を編成する機能を持つ．

3．DATA コマンド群

データの入力，データ・クリーニング，データの加工，データの編集の他，データセットの合併や結合の機能を持つ．

4．TABLE コマンド群

集計・作表仕様の登録，削除，ならびに仕様を用いる集計および作表の機能を持つ．

これらを表3.1に示し，その概要は以下の各章で述べる．

3.4 誤入力に対する管理

SSMS の基礎プログラムである SAS はシンタックス・エラーやプログラム・エラーが発生すれば，実行を完全に停止する仕組みになっている．したがって，SSMS もこの種のエラーに遭遇すれば，実行を停止し，以後の作業を継続することができなくなる．エラーの発生原因の大部分はコマンドの引数として入力するデータセット又は集計・作表仕様の名前の誤入力，コマンドに引き続いて入力するパラメータの不正確さなどである．

このシステムは，SSMS ライブラリのディレクトリ等を利用して，入力の妥当性を可能な限り診断し，もし誤入力と判定すれば，その原因を表示し，再入力を求める仕組みになっている．その一方で誤入力の原因となる入力量そのものを削減するため，次のような入力の省略形又は簡略法を採用している．

1．引数に当てる値がシステムで限定されている場合は，名前を識別しうる範囲で，その先頭の1文字又は2文字だけ入力させる．

2．調査設計情報の入力において，同一のパラメータを繰り返し入力するような場合，1個だけを入力させ，以後空を返させる．

3．記号的に連続したパラメータの組は省略形による一括入力を認める．

なお，パラメータ入力の際は，入力終了後，利用者にその諾否の確認を求める．

4. 自由コマンド

コマンドには作業中に随時使用できるものと各作業のレベルに拘束されるものとがあり，前者は次の機能を持つ．後者は次章以降で述べる．

4.1 DD/D 情報の表示

DD/D に収められた各種の管理・定義情報を端末に表示するためのコマンドをリスト・コマンドとよび，次の5種がある．

(1) LISTD コマンド

SSMS ライブラリのディレクトリに登録されている調査設計情報，データセット，集計・作表仕様について，名前，OS のファイル名，作成年月日を種類別に表示する．

(2) LISTP コマンド

調査設計情報の内容を表示する．

(3) LISTS コマンド

集計・作表仕様の内容を表示する．

(4) LISTH コマンド

データライブラリに格納されているデータセットの作成経過を表示する．

(5) LISTV コマンド

データセットに含まれている回答項目（又は変数）名を表示する．

4.2 データの出力

データセットの内容を出力させるコマンドをプリント・コマンドとよび，次の2種がある．

(1) PRINT コマンド

データセットの内容を指定した観測点の範囲で端末に表示する.

(2) LP コマンド

データセットの内容をすべてラインプリンタに出力する.

4.3 ライブラリの操作

データライブラリを操作するコマンドをメンバ・コマンドとよび, 次の4種がある.

(1) SAVE コマンド

データ処理で作成された一時的ファイル (TSASDS) をデータセットとして格納する. 同名のものがあれば置換する.

(2) NOSAVE コマンド

作成された TSASDS を放棄して他のデータセットの処理に移行する.

(3) DELM コマンド

データセットをライブラリから削除する.

(4) COPY コマンド

ライブラリ間あるいはライブラリ内でデータセットを複写し, 複写元を削除する.

4.4 システムの操作

システムを操作する基本的なコマンドで, 次の3種がある.

(1) HELPコマンド

実行中の作業レベルで使用しうるコマンド, あるいは上記の各コマンドの引数の意味を表示する.

(2) RETURN コマンド

実行中の作業レベルを終了し, 上位のレベルに復帰する.

(3) END コマンド

すべての作業を終了する.

5. 調査設計情報

5.1 調査設計情報の管理

SSMS で管理する調査設計情報には，利用者が直接入力する始源的情報の他に，これらを基礎にシステムが以後のデータ処理や表示に適合した形に編成したものがある（表5.1）.

利用者が入力すべき情報は次のとおりである.

(1) 調査概要

調査全体の概要を示す記述的情報（表5.2a）.

(2) 調査票定義

a．回答集団

2.2節で述べたように，関連質問法が採用されると，集計のため逐次集計部分集団を設定する必要がある．SSMS ではこれを「回答集団」とよ

表 5.1 SSMS ライブラリ中の調査設計情報

メ ン バ	内 容
SURVEY	調査概要
UNIVERSE	回答集団の定義
ANSWER	回答項目の定義
NOTE	回答集団と回答項目に関する注記
MSTFORM	MA→SA 展開のための情報
UCHAIN	回答集団間の階層的関連性の記述
USELECT	回答集団の選択に必要な情報
UATABLE	回答集団と回答項目の関連表（どの回答集団はどの回答項目に回答すべきかを表すもの）
DFORMAT	入力ファイルのレコードフォーマット
ALENGTH	ライブラリに格納時のデータ長

表 5.2a 調査概要に関する情報

項 目 名	意 味
SCODE	調 査 コ ー ド
SNAME	調 査 名
SDATE	調 査 日
SREGION	調 査 地 域
SSAMPLE	調 査 方 法
SNOTE	注 記

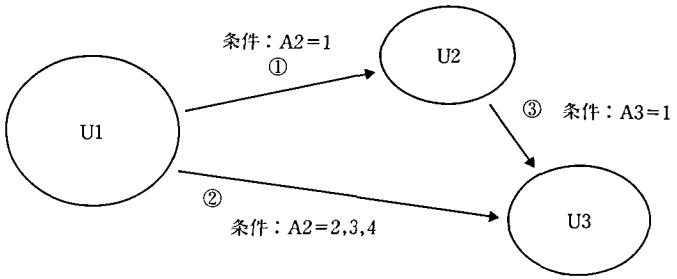

図 5.1

表 5.2b 回答集団の定義情報

項目名	意味	①	②	③
UCODE	回答集団コード	U2	U3	U3
ULABEL	集団ラベル（名称）			
UPARENT	親集団コード	U1	U1	U2
UCOND	条件式	A2=1	2=<A2<=4	A3=1
UNOTE	注記			

び，管理の1要素としている．

1つの回答集団は，それ以前に定義されている回答集団（親集団とよぶ）から，回答項目コードがとる値を条件として導出される．例えば，図5.1は U1 の構成員のうち回答項目 A2 で1を選択した者のみからなる集団を U2 と定義し，U1 のうち A2 で2，3，4を選択した者と，U2 のうち A3 で1を選択した者の集まりを U3 と定義することを示している．このような関係を記述するのが回答集団定義情報であり，表5.2bの①～③はそれぞれ図5.1に対応する記述である．

b．回答項目

SSMS では調査票の定義にあたり，質問ではなくそれに対して発生した回答を単位として採用している．その理由は，2.2節で述べたように1つの質問に対して，複数の値が発生する場合があるからである（制限複数選択法）．回答に対応するデータ項目を管理単位（フィールド）とすること

表 5.2c　回答項目の定義情報

項目名	意味
ACODE	回答項目コード
ALABEL	回答ラベル（名称）
ATYPE	回答タイプ
ANEWV	新回答項目コード
ASCALE	尺度
AMAX	尺度の上限
AUNIT	単位
UCODE	回答集団コード
ANOTE	注記

により，SSMS では多様なデータ形式を比較的簡単に処理することができる．

また，管理対象となるフィールド属性についてもデータ処理のために必要最少限のもののみに限定している．以下，主要な属性についてだけ述べ，全体は表5.2cに示す．

・回答項目コード：各回答項目の識別コードで，システム内のすべてのデータ処理はこのコードで指示する．

・回答ラベル：調査票の質問番号を表現する記号を入れる．これにより回答項目と調査票の対応が明確になる．

・回答タイプ：回答項目の質問形式を示す．単一選択法，制限複数選択法，自由回答による数値の各場合にしたがって，それぞれS，M，Nを入れる．

・MA→SA 展開後の新回答項目コード：回答タイプMの項目はシステムにより MA→SA 展開され，0/1のバイナリ回答に変換されるが，このとき新たに作成される回答項目群のコードを入れる．

・尺度：データの用途を示す．SSMS ではシステム外から入力される資料はすべて数字データとしているから，その数字の意味を規定する必要がある．値が単なる分類コードとしての意味しか持たないものに

はC，値が数値としての意味をも持ち集計操作の対象となるものにはAを入れる．

・尺度の上限：この項目がデータとしてとりうる最大値．尺度がCの場合は選択肢の最大番号，Aの場合は起こりうる最大値を入れる．

・回答集団コード：この回答項目に答えるべき回答集団のコードを入れる．

(3) コーディング形式の定義

調査の実データがコーディングされた形式．これは入力外部ファイルのデータ・フォーマットである．

5.2 調査設計情報の操作

上記の調査設計情報の入力には PIN コマンドを，その SSMS ライブラリへの登録には PCAT コマンドを用いる．また，コーディング形式の入力には DFORM コマンドを用いる．

5.3 処理対象の拡大

SSMS は本来調査データの管理と分析を支援するために開発されたシステムであるが，データの操作，集計・作表の機能は，身近な行列形式を持つ統計データにそのまま適用可能である．この種の利用も想定して，調査設計情報を入力することなく，ただ資料行列の変数の個数だけの入力（DFORM (NOV) コマンド）で，実データの入力を可能にしている．もちろん，このデータには調査設計情報を組み込んだコマンドは適用できない．

上述の方法で複数の統計データを入力し，これらを種々の形で組み合わせ，新しい統計データを作成することができる．

6. データの操作

6.1 調査データの入力

調査データは DIN コマンドによりディスク上の外部ファイルからシステムに入力され，一時的ファイル TSASDS に保存される．

6.2 データ・クリーニング

CLEAN コマンドは調査データに含まれる誤りを調査設計情報などに基づいて修正する．これには次の4種のサブコマンドがある．

(1) BLUNDER サブコマンド

各回答項目について，データの転記やパンチの際に生じた範囲外の値(無効値）を検出し，記号 'B' で置換する．

(2) ZERO サブコマンド

無制限複数選択法による回答のコーディング時に，選択されなかった選択肢に対応するデータ項目を空白にした場合，これは欠損値（無回答）と区別できなくなるため，この種の回答に限り空白を0で置換する．

(3) BLANK サブコマンド

上記以外の欠損値をすべて記号 'M' で置換し，欠損値を確定する．

(4) MODIFY サブコマンド

一定の条件を満たす調査票を対象に，発見されたデータエラーの修正と矛盾回答の検出を行い，調査設計情報で規定された範囲内の値を当てるか矛盾回答を表す記号 'C' で置換する．

6.3 データの加工

PROC コマンドは回答項目間に必要な変換操作を施して新しい回答項目を定義する．これには次の5種のサブコマンドがある．

(1) MA サブコマンド

調査設計情報の回答タイプがMの項目について MA→SA 展開を行い，各選択肢が1つの回答項目に対応する形式に変換し，それぞれに非選択(0)，選択(1)の値を与える．展開して追加された回答項目には調査設計情報の定義時に与えたコードを与える．

(2) INVALID サブコマンド

指定した主要な回答項目について，'B', 'M', 'C' のいずれかの値を持つものの比率が一定値以上の調査票を無効票と判定する．その結果は新しい回答項目 FLAG に，0が有効，1が無効として与えられる．

(3) COMPUTE サブコマンド

回答項目間，あるいは回答項目と定数間の演算により新しい回答項目を

定義する．演算式には通常の算術演算子の他に SAS 関数も使用できる．

(4) WEIGHT サブコマンド

地域により抽出率の相違がある場合など，回答項目の各値に加重して集計する必要があるときのためにウエイト変数を新しく定義する．ウエイト変数の値は指定した回答項目の値に基づいて定められる．

(5) RENAME サブコマンド

回答項目名を変更する．

6.4 データの編集

VIEW コマンドはデータベースでいうビュー機能の一部である．これには次の5種のサブコマンドがある．

(1) SELECTサブコマンド

回答項目とその値からなる条件式を満足する調査票を選択する．

(2) USELECT サブコマンド

指定したコードの回答集団に属する調査票を選択する．

(3) MSELECT サブコマンド

指定した主要な回答項目中で，データ・クリーニングで無効回答と判定されたものの比率に関する条件式によって調査票を選択する．

(4) PROJECT サブコマンド

指定した回答項目を選択する．

(5) UPROJECT サブコマンド

指定した回答集団が回答すべき項目を選択する．

6.5 データセットの合併と結合

ある条件を満たす2つのデータセットの連結操作は次の2種のコマンドで行う．

(1) UNION コマンド

2つのデータセットが同じ回答項目の組を持つとき，両者を縦に連結する．

(2) JOIN コマンド

2つのデータセットに共通する回答項目が存在せず，しかも同一の観測点の組でかつその順序も一致しているとき，両者を横に連結する．

7. 集 計・作 表

7.1 集計・作表情報の登録と削除

以下では「変数」を「回答項目」と同じ意味を持つ用語として用いる．SPEC コマンドは，集計・作表に必要な情報を SSMS ライブラリに登録する．これらは4種に類別され，次のサブコマンドで登録される．

(1) TFORM サブコマンド

作表時にカテゴリを表すために，変数の値に代わる文字列を変数の値ラベルとして登録する．これを TFORM コードとよぶ．これは分類の括りを変更して新しい値を与えたり，数値で表された変数を階級化するためにも使用される．

(2) GSPEC サブコマンド

複数個の表に共通する一般的な情報を登録する．これを GSPEC コードとよぶ．これには次の属性が含まれる．

・GSPEC コード：この情報の識別コード．

・表題：各表に共通のタイトル．

・分類変数：表のマス目を規定する変数．

・解析変数：総計，平均などの統計量の算出対象となる変数．

・加重変数：統計量の算出時にウエイトとして用いられる変数．

・フォーマット：変数に TFORM コードを対応させる．

・変数ラベル：変数名の代わりに表中で用いられるテキスト．

(3) TSPEC サブコマンド

各表の個別的仕様を登録する．これを TSPEC コードとよぶ．表の仕様は1～3次元の表現が可能で，ページ，行，列に対応させて，(2)で指定した分類変数と解析変数の中からコードの組み合わせで表現する．

(4) SUM サブコマンド

データセットの集計法を指定する．これを SUM コードとよぶ．(2)の項目である分類変数，解析変数，加重変数の他に必要な統計量を指定する．

登録した仕様の削除は DELS コマンドで行う．

7.2 集　計

AGGREG コマンドはデータの集計を行う．これには次の3つのサブコマンドがある．

(1) RECLASS サブコマンド

TFORM サブコマンドで登録した TFORM コードを用いて変数値の再分類を行う．

(2) SUMMARY サブコマンド

SUM サブコマンドで登録した SUM コードを用いて集計データセットを作成する．

(3) NORMAL サブコマンド

集計データセットを標準化する[4]．

7.3 作　表

作表処理は DISPLAY コマンドで指示する．これには次の3つのサブコマンドがある．

(1) FREQ サブコマンド

変数名を指定するだけで簡単な度数分布表とクロス表を作成する．

(2) HIST サブコマンド

指定した変数のヒストグラムを作成する．

(3) TABULATE サブコマンド

GSPEC コードおよび TSPEC コードを指定して統計表を作成する．1つの GSPEC コードに複数の TSPEC コードを対応させ，また対象データセットを変えながら，いくつかの表を連続的に作り出すことができる．

8. むすび

われわれは，本システムにより従来顧慮されることが稀であった調査データ管理とそのデータ処理の問題を解決する1つの方法を提案した．そ

4）分類変数の値の組み合わせでできる集計部分集団に属する観測点の度数が0であるときは，通常，それに関するデータ行は集計データセットに現れない．このデータ行を欠損値を含む形で追加することを標準化という．

こに示された必要なデータ管理要素とデータ処理機能は，今後の統計データベース管理システムの研究にもなんらかの示唆を与えるものと信ずる．また，先に述べたように，このシステムは一般の統計データの一次処理および集計・作表にも適用しうる機能も数多く備えており，幅広い利用が期待される（1985.9.5）．

最後に，この研究は昭和59年度文部省科学研究費（特定研究２），「多目的総合統計データバンク」の「統計データベース管理システムの開発」（研究代表者，小林康幸）の助成によるものであることを付記し，謝意を表する．

参 考 文 献

[１] Kent, W., "Choices in Practical Data Design," Proceedings of the 8th International Conference on Very Large Data Bases, 1982, pp. 165–180.

[２] Kobayashi, Y., "Data Dictionary/Directory System of a Statistical Database," Journal of Information Processing, Vol. 7., No. 4., (1984) pp. 233–239.

[３] 横山和典，椿　康和，「地方統計のデータベース化について」，広島大学経済学部紀要　年報経済学，５巻，(1984)pp. 1–25.

[４] SAS USER'S GUIDE: Basics 1982 Edition(日本語版)，1984　アシスト．