数理解析研究所講究録
第448巻 1982年 56-68

統計解析のためのＤＢＭＳの追加機能について

広大　総合情報処理センター　小林康幸
筑波大　社会工学系　池田秀人

1はじめに

統計パッケージは数多く開発されているが，その多くは順編成ファイルを対象としており，データ管理や統計解析以外の目的で構築されたデータベース上での処理には問題がある。

最近データベースシステムを使った統計パッケージが表れデータ管理を容易にしているが，統計解析は通常全データを参照することにより行なわれるため，会話型でのサポートには実時間の応答の面で工夫が必要である。また医学データや社会調査データなどにおいては，原データを公開せず概要のみを公開するといった機密保護の機能も必要になる。

ここではデータベースに対して会話型で統計解析を行なえるようにするために，従来のデータベース管理システムＤＢＭＳの機能に加えてどのような機能が必要かを議論し，それらの機能を満たすための具体的方策を述べる。

2　会話型統計解析のための必要機能

データベースに対して，会話型で統計解析を行なうには，従来のＤＢＭＳの機能に加えて次の機能が要求される。

2.1 統計解析の実時間応答

我々は統計解析の環境を次のように考える。

すでに解析に必要なデータは，データベース上で管理されている。どのような手法で解析すればよいかはあらかじめ決まっていない。会話端末を用い1つの手法で解析を行ないその結果を見て次の解析の手法を検討する。

このような環境のもとでは，解析結果の精度よりはむしろ解析結果を早く出すことが重要である。解析の手法が決まった時点で時間をかけ詳細な結果を出せばよい。

2.2 derived data の拡張

データベース上のデータに何らかの計算を施して得られるデータをderived dataと呼ぶ。利用者が計算式を指示することなく，derived dataをあたかもえのデータとして存在するかのごとく扱えれば便利である。従来のDBMSでは簡単な計算によるderived dataを取り扱っているが，統計解析の為には平均，分散などの基本統計量をderived dataとして扱う必要がある。

2.3 アグリゲートデータの取り扱い

アグリゲーションは統計解析ではよく使われる操作であり，従来のDBMSでも取り扱うことができる。しかしアグリゲートデータをデータベースとして構築するには，あらかじめ

ファイルの定義をしなければならない。このファイル定義は、データベースの管理者の行なう仕事であり，統計解析者にとっては不慣れな操作になる。自動的にファイルの定義が行なわれ，アグリゲートデータが作成される機能が必要である。

2.4 統計データに対する機密保護

統計データの場合，解析結果の公表はされるが，個々のデータについては公表しない場合が多い。例えば医療データで，癌患者数の年度別推移は公表されるが，各人の氏名は機密事項である。従来のDBMSでは，同一人に統計解析をすることは許すが，ケースデータを見ることを許さないように制御することは非常にむずかしい。統計データベースでは，このような制御ができなければならない。

2.5 統計辞書の提供

分布関数や回帰問題で使われる関数の辞書を提供することは，統計についての知識を十分もたない人の解析の支援に必要である。

3 要求を満たすためのアーキテクチュア

統計解析に使われるデータは，大きく分類すると，次の2通りに分かれる。1つは統計解析そのもののために作成されたデータであり，他の1つは何か別の目的のために作成されたが，それが統計解析にも使えるようなデータである。

我々がここで提案するアーキテクチュアは，後者のデータを想定している。つまり従来のＤＢＭＳのもとで管理されているデータに対し，会話型統計解析が可能になるアーキテクチュアである。従来のＤＢＭＳがもっている機能を変更せず新たな機能を準備する方法である。

我々が行なう機能拡張は次の２点である。その１つは data dictionary / directory (DD/D)の拡張である。データ構造定義用の directory に単位，精度，欠測値等や，平均，分散，最大値，最小値等の基本統計量を追加する。また分布関数や回帰問題で使われる関数等の辞書も準備する。我々が行なう機能拡張の他の１つは，summary table (3.2で詳しく述べる)と呼ぶ概要データの追加である。会話端末を用い統計解析を行なう場合，まず解析の対象となる変数群についての summary table を準備する。この summary table をもとに会話型で(試行錯誤で)統計解析を行なう。これらは近似的な解析になるため，解析手順が決まった後にバッチモードで正確な結果を出せる道具を準備する。

3.1 DD/Dの拡張

会話型統計解析のための従来のＤＢＭＳがもっているDD/Dの拡張は次の４つの部分に分かれる

3.1.1 フィールド属性の拡張

従来のDBMSでは，DD/Dでフィールド名とその属性を管理している。フィールド属性としては，最大レコード長，数字か文字かのレコードタイプ等であるが，統計でよく使われる属性として，離散か連続かの区別，欠測値の表現，実レコード数(N)，単位，精度，理論分布，最大値(MAX)，最小値(MIN)，メディアン(MED)，平均値(M_1)，2次モーメント(M_2)，3次モーメント(M_3)，4次モーメント(M_4)，を準備する。これにより，次の統計量を得ることができる。

合計： $SUM = M_1 N$

範囲： $RAN = MAX - MIN$

2次中央モーメント： $C_2 = M_2 - M_1^2$

3次中央モーメント： $C_3 = M_3 - 3M_2M_1 + 2M_1^3$

4次中央モーメント： $C_4 = M_4 - 4M_3M_1 + 6M_2M_1^2 - 3M_1^4$

標準偏差： $SD = \sqrt{C_2}$

変異係数： $CV = SD/M_1$

歪度： $SK = C_3/SD^3$

尖度： $KT = C_4/C_2^2$

この拡張により，利用者はデータ量の多少にかかわらず，これらの統計量を実時間で得ることができる。また次のような条件を満たすレコードの検索も可能になる（WEIGHTはフィールド名）。

$M_1 - 2SD \leqq$ WEIGHT $\leqq M_1 + 2SD$

3.1.2 derived data のディレクトリー

3.1.1で挙げた統計量の他に利用者が，あらかじめ計算式を与えておくことにより，その値をderived dataとして取り扱うことができるよう，derived data のディレクトリーを準備する。

3.1.3 分類のディレクトリー

3.2で述べるsummary tableの作成には，データの分類が重要を役割をはたす。例えば，年令を（～10, 10~20, …60～70, 70～），性別を（MALE, FEMALE）に分類したTable 1はsummary tableの1つの例である。

Table 1

SEX	AGE							
	-10	10-20	20-30	30-40	40-50	50-60	60-70	70-
MALE	4	50	62	66	70	71	65	32
FEMALE	4	41	61	75	82	82	76	49

次に我々が考えている分類の定義の例を挙げる.

〈例1〉 増分値による定義

指定: $W = x_1, x_2, x_3, n$

意味: $n=1$の時 $(-\infty, x_2), [x_2, x_2+x_1), \cdots, [x_2+mx_1, x_3), [x_3, \infty)$

$n=0$の時 $(-\infty, x_2], (x_2, x_2+x_1], \cdots, (x_2+mx_1, x_3], (x_3, \infty)$

〈例2〉 区間による定義

指定: $\{\{\ \}\ x_1, x_2\ \{\ \}\}, \{\{\ \}x_3, x_4\ \{\ \}\}, \cdots, \{\{\ \}x_{n-1}, x_n\{\ \}\}$

意味: 各区間を上限値, 下限値で指定する。

〈例3〉 点による定義

指定: $D=(A_{11},A_{12},\cdots,A_{1n_1}),(A_{21},A_{22},\cdots,A_{2n_2}),\cdots,(A_{m1},A_{m2},\cdots,A_{mn_m})$

意味: 括弧でくくられた値が1分類となる。$n_j=1$ の場合は括弧を省略できる。

利用者がsummary tableの作成時に指定した分類の方法は, あらかじめ用意された分類のディレクトリーに記録される。

3.1.4 統計関数の辞書

分布関数や回帰問題で使われる関数の辞書を準備する。辞書に記録される情報としては, 各関数のサブプログラムの名前, そのパラメーター, 離散型か連続型かの区別, を考えている。

3.2 summary table

summary tableは原データから作成することができるが編集や統計解析の結果として作成することもできる。

3.2.1 summary tableの定義

summary tableは次のシンタックスで定義できる。

テーブル名 [(観測変数名)]:

分類変数名1 (分類名1) [= 形式1]

分類変数名2 (分類名2) [= 形式2]

⋮

summary table のます目に入る値を観測変数名で，分類の対象となるフィールドを分類変数名で指定する。分類変数名がフィールド名でなく，計算式等によって得られる場合は，その形式を指定する。観測変数名と形式では，定数，フィールド名および3.1.1で挙げた統計量による算術式を指定することができる。

例えば Table 1 は次のように定義できる。

TBL1: SEX(D=MALE, FEMALE)
AGE(W=10, 10, 70)=1981−BIRTH

AGEは1981−BIRTHによって得られる。

3.2.2 summary tableの操作

すでに作成されているsummary tableを操作することにより新たなsummary tableを得ることができる。

〈例1〉 TBL1をSEXに関して射影してTBL2を作成する。

TBL2=TBL1|SEX

Table 2

SEX	AGE							
	−10	10−20	20−30	30−40	40−50	50−60	60−70	70−
TOTAL	8	91	123	141	152	153	141	81

〈例2〉 TBL1をAGEに関して再分類してTBL3を作成する。

TBL3 = TBL1. AGE (W = 20, 10, 50)

Table 3

SEX	AGE			
	-10	10-30	30-50	50-
MALE	4	112	136	168
FEMALE	4	102	157	207

上の例の他に，観測変数や分類値の追加などの操作も準備する。

3.3 統計解析と結果の表現

会話型統計解析は，結果の精度よりもむしろ処理速度の速さに重点がおかれ，逆にバッチ型統計解析では，処理に時間がかかっても正確な結果が出ることが重要になる。

また統計処理と，その結果の表現とを分離することにより，多様な解析結果の表示が可能になる。

3.3.1 会話型統計解析

我々は，会話型統計解析はsummary tableを入力として，1つまたは複数のsummary tableを作成する操作だと考える。特別な場合として，1つの値もsummary tableと考える。

summary tableはもとのデータに比べレコード数が少ない為，実時間の応答を可能にする。しかしその解析結果はおおよその値になる。結果の精度と応答速度の重点の置き方によって，summary tableの分類を細かくするか荒くするかは

利用者が選択できる。試行錯誤による統計解析は近似的な結果により行なわれる。解析の方法が決まった時点での詳細な解析はバッチモードで行なわれる。この為，会話型での解析の流れが自動的に記録される機能を準備する。

3.3.2 summary table の表現

次のように定義された3次元以上の summary table は，以下の2通りの方法で2次元または3次元の表示が可能である。

TBL6: GRD (D=Light, Not)
PRT (D=Light, Not)
CHD (D=Light, Not)

〈例1〉 2次元表示

DISPLAY TBL6 (CHD="Child", GRD="Grandfather" *PRT="Parent")

Table 6

Child	Grandfather			
	Light		Not	
	Parent		Parent	
	Light	Not	Light	Not
Light	1928	552	596	508
Not	303	395	225	501

〈例2〉 3次元表示

DISPLAY TBL6 (GRD="Grandfather", CHD="Child", PRT="Parent")

Grandfather=Not

Child	Parent	
	Light	Not
Light	596	508
Not	501	

Grandfather=Light

Child	Parent	
	Light	Not
Light	1928	552
Not	303	395

Figure 1

summary tableを結合した表示や図による表示も準備される。次にその例を示す。

〈例3〉 結合表示: TBL3とTBL3|SEXをAGEで結合

DISPLAY TBL3+(AGE) TBL3|SEX (SEX, AGE)

Table 7

SEX	AGE			
	10-	10-30	30-50	50-
MELE	4	112	136	168
FEMALE	4	102	157	207
TOTAL	8	214	293	375

〈例4〉 図形表示

HISTOGRAM TBL3|SEX

AGE	Frequency
-10	(8) *
10-30	(214) **********
30-50	(293) **********
50-	(375) **********

Figure 2

3.3.3 アグリゲートデータ

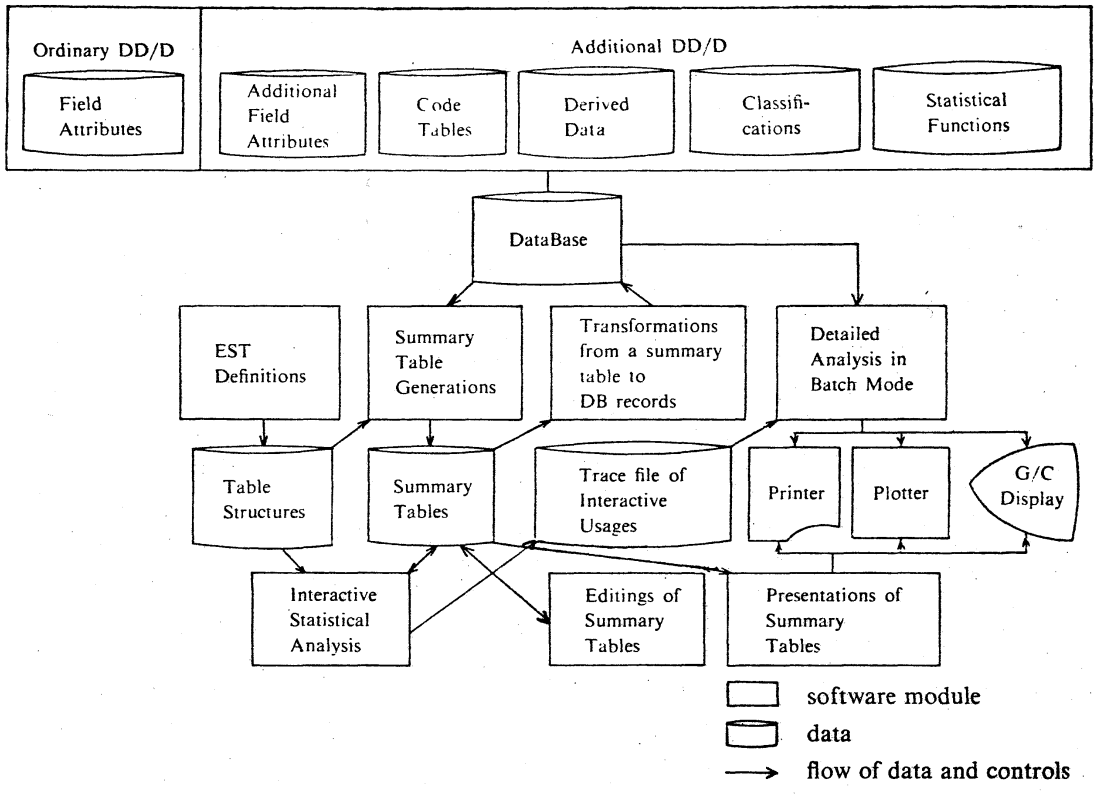
summary table をデータベース上のレコードに変換し，アグリゲートデータを作成するコマンドを準備する。このコマンドは自動的にデータベースのファイル定義を行なう。

3.3.4 機密保護

原データと summary table の機密条件を変えることにより，同一人に原データを参照することはできないが，summary table をもとに統計解析を行なうことができるように制御することが可能になる。

3.4 システムの実現

次の図は広島大学で実現するシステムHSDBの概念図である。



広島大学では現在，データベース管理システムHDMのもとで，管理事務電算化，図書館の機械化，病院の機械化のためのシステムが開発されている。作成されたデータベースに対する統計解析の要求が多く，HDMで管理されているデータベースに対して統計パッケージSPSSを利用できる機能をすでに準備している。しかしこの機能は会話型統計解析の要求を十分に満たさない。我々はHSDBによりこの問題を解決しようと考えている。

参考文献

[1] Ikeda, H. and Kobayashi, Y., Additional Facilities of a Conventional DBMS to Support Interactive Statistical Analysis. Proceedings of the Workshop on Statistical Database Management, Menlo Park, CA Dec. 1981.