社会調査結果の視覚化データベース

Visualized Database of Social Survey Results

吉田　光雄

Mitsuo YOSHIDA

大阪大学　人間科学部

Human Sciences, Osaka University

キーワード: *Mathematica*, 探索的データ解析, 情報処理教育

Keywords: *Mathematica*, EDA(Exploratory Data Analysis), Education for Information Processing

あらまし: 社会調査結果を計算機内に保存し、必要に応じてデータを取り出し、*Mathematica* を用いて、基本統計量を計算すると同時に、ヒストグラム、散布図等のグラフを描き、データの様相を視覚的に検討することのできるデータベースをワークステーション上に構築した。すなわち、保存されたデータをテキストとして取り出せると同時に、それらを視覚的に提示し、探索的にデータの内部構造を探り、データの持つ情報を十分に引き出して分析を深化させるために用いることができる。

本システムは小規模の試作用データベースであるが、メモリー・ハードディスクの容量や高速演算などの十分な計算機資源が得られれば、さらに大規模のデータにも適用することができる。快適に動かして大量のデータを処理するためには、かなりな量のメモリを必要とするが、統計処理のメニュの提示や選択、さらにはより美しいグラフィックスの描画や3次元表示法に対する工夫など、ユーザー・インターフェイスの改良は今後の課題である。

Summary: Social survey results should be preserved as database, so that one can consult it afterwards and extend it to another analysis. If it can be browsed in a visual format, such a database would be an efficient and convenient tool for many kinds of social surveys.

This report summarized how to save data in *Mathematica* as a database, and how to make statistical treatments using graphic techniques, such as histogram, bar-chart, pie-chart, scatter diagram, three dimensional graphic and others in it.

Mathematica seems to be suitable for this purpose, because of its wide coverage of mathematical and statistical usages.

This database seems to be useful, not only to preserve all of the data, but as a visual database that contributes as an EDA(Exploratory Data Analysis) and VDA(Visual Data Analysis) revealing new and hidden structures of the data.

An example of social survey results concerning computer education at Osaka University was demonstrated.

1. はじめに

Mathematica ^{TM}[3] は1986年、Wolfram Research 社から発売され、現在世界中で広く使用されている数式処理ソフトのひとつである。操作の容易さ、機種互換、グラフィックス化等に特徴があり、例えば、式の展開、因数分解、方程式、数値計算、行列演算、微分・積分、微分方程式などの数式処理の他、関数のグラフ表現、統計計算など、数学的処理を容易に行うことができる。強力なパッケージ[2]を多数含み、"A System for Doing Mathematics by Computer" として評判の高いものである。

統計学に関するパッケージも多く含まれているので、さまざまな統計処理にも活用することができる。しかも、グラフィックス機能が豊富であるので、目で見る統計学として有効である。ワークステーション (NeXT) 版およびマッキントシュ版はユーザー・インターフェイスのよい Notebook を介して使用可能であり、コマンド入力、プログラミング、出力、グラフィックスがすべて同一ウインドウ上で実行され、しかも両機種間で高い互換性を有している。

統計パッケージ SAS を用いた社会調査結果の視覚化データベース、ならびにS言語を用いたものも可能であるが[5]、本稿では *Mathematica* を用いて行う方法について報告する[1]。また、まだモジュール化されていない、いくつかの統計的手法については、内蔵の言語を用いてプログラミングを行ったので、それらについても報告する。追加された各種の統計演算はライブラリとして保存されている。

Turkey,J.W.[4]により探索的データ解析 (EDA, Exploratory Data Analysis) の方法が言われて以来、統計データを視覚的に分析する方法 (VDA,

[1]本稿は、文部省科学研究費補助金・重点領域研究、「情報化社会と人間」研究成果報告書、および日本計算機統計学会第8会大会にて報告したものの一部である。

Visual Data Analysis) が注目され、いろいろな方法が提案されている。本データベースはあくまでもデータの統一的保存が目的であるが、保存されたデータをテキストとして取り出せると同時に視覚的に提示し、全体像の把握を容易にすることをも目指す。そして、更に必要な場合には再度統計処理も行うことが可能なものとする。統計処理はすでになされているが、様々な角度からより探索的にデータの内部構造を探り、データの持つ情報を十分に引き出し、分析を深化させるために用いることができる。

大阪大学では、情報処理教育センターを中心にNeXT を端末とする分散処理システムのネットワークが構築され、学生の情報処理教育に活用されている(1)。教材の提供・課題の提示・レポートの収集などが容易に行える教育支援のためのアプリケーションが開発されており、それを活用することにより、登録学生は容易に本データベースにアクセスすることができる。

2. データの作成と *Mathematica* の基本

2.1 保存および参照

Mathematica を起動する前に、カレント・ディレクトリにデータをテキスト (Ascii) 形式で保存しておく。データの形式は、各行にサンプル (ケースまたはオブザーベーション)、各列に変数 (調査項目) をあて、行列データの入力形式に従って、各行ごとに { } を使用する。

データは連続量でもよいし、数字化された属性データでもよい。ただし、処理や結果の解釈に際して、データの属性に注意する必要があり、現在のところデータベースのシステム内に、可能な統計処理をチェックする機能は組み込まれていない。

調査票の項目一覧も必要であり、*Mathematica* 内の変数番号と、調査票の項目番号の対照も必要である。本計算機システムが NeXT ワークステーションであるため、多重ウインドウを開くことが可能であり、項目対照を明示するウインドウを常時開いておくと、操作が容易となる。項目の選択は変数名やラベルを使用するのではなく、通し番号で行う。

調査票は多くの場合、回答者のバイアスを避けるためランダムに並べられているが、計算機に保存する際には、予め質問項目番号、回答カテゴリの方向性などを整理し、整合性を保って保存しておいた方が便利であろう。調査項目一覧もデータに合わせて並べ換えておく。

メモリなど、計算機資源が豊富な場合には、全データを一括して読み込ませておくと迅速な処理が可能であるが、十分でない時は、必要な項目を処理の都度読み込めばよい。

一旦保存されたデータは修正ができないよう、ライトプロテクトをかけておく。データを加工しても、保存はせず、必要な修正はその都度行うようにしておかないと、データベースの機能を果たさなくなる。従って、以後の修正が無用となるよう、最初に計画的に保存しておく必要があろう。

現在のところ NeXT 版 *Mathematica* (Ver. 2.1) は日本語対応とはなっておらず、出力に日本語が使用できないのは難点である。ただし、処理結果をワープロや作図ソフトに取り出して編集すれば、日本語は使用可能である。

2.2 *Mathematica* の基本

Mathematica を起動し、ハードディスク上のカレント・ディレクトリ (~) に保存されているデータ・ファイルを *Mathematica* に読み込む。必要に応じて、特定の項目に関する全サンプルの回答を抽出し、編集しなければならないが、そうした作業を便利に行うことのできるコマンドも多く用意されている (表1)。また、統計処理を行うためには、統計パッケージを予め読み込んでおかねばならない(2)が、これらの一連の初期作業をルーチン化してプログラミングしておいてもよいし、*Mathematica* との対話の内容は *Notebook* に保存されているので、それらを利用して copy & paste しつつ、再度実行することも可能である。

3. 統計処理

3.1 基本統計量の算出

統計パッケージ (記述統計学) を用いることにより、平均・メディアン・モード・分位数などの代表値、分散・不偏分散・標準偏差・範囲・四分偏差などの散布度のほか、歪度・尖度なども直ちに計算することができる (表3)。

3.2 区間推定・検定

パッケージ (信頼区間、検定) を用いて、母分散や母平均に関する区間推定・検定を行うことができる (表4)。データは項目番号を指定して粗データをベクトルとして取り出す。その際、オプションを用いて、有意水準、両側・片側などを設定する。

表 1: データ読み込み・編集のためのコマンド

パッケージ　*Statistics'Master'* *Statistics'DataManipulation'*	
`ReadList["~/file1.dat",Number]`	
`ReadList["~/file1.dat",Number, RecordList->True]`	
`TableForm[data]`	`Column[data, n]`
`Column[d,{n1,···}]`	`ColumnTake[d,n]`
`ColumnDrop[d,{n1,···}]`	`ColumnJoin[d1,d2,···]`
`Row[data, n]`	`RowTake[d,n]`
`Row[d,{n1,···}]`	`RowDrop[d,{n1,···}]`
`RowJoin[d1,d2,···]`	
`BooleanSelect[d,sel]`	`TakeWhile[d,pred]`
`LengthWhile[d,pred]`	

表 2: データ集計のためのコマンド

パッケージ　*Statistics'DataManipulation'*	
`Frequencies[data]`	`QuantileForm[d]`
`CumulativeSums[d]`	`BinCount[d,{min,max,dx}]`
`RangeCount[d,{c1,c2,dx}]`	
`CategoryCounts[d,{cat1,cat2,···}]`	
`BinLists[d,{min,max,dx}]`	
`RangeLists[d,{c1,c2,dx}]`	
`CategoryLists[d,{cat1,cat2,···}]`	

表 3: 基礎統計量算出のためのコマンド

パッケージ　*Statistics'DescriptiveStatistics'*	
`Mean[data]`	`Median[d]`
`Mode[d]`	`GeometricMean[d]`
`HarmonicMean[d]`	`Quantile[d,q]`
`Quantiles[d]`	`LocationReport[d]`
`Variance[d]`	`VarianceMLE[d]`
`StandardDeviation[d]`	`MeanDeviation[d]`
`StandardDeviationMLE[d]`	`SampleRange[d]`
`QuartileDeviation[d]`	`Skewness[d]`
`DispersionReport[d]`	`Kurtosis[d]`
`CentralMoment[d]`	`ShapeReport[d]`
`covariance[`$x1, x2$`]`	`correlation[`$x1, x2$`]`
`multipleCorr[`x, y`]`	`partialCorr[`x, i, j`]`
`canonicalCorr[`$x1, x2$`]`	`contTable[`$x1, x2$`]`
`association[`$x1, x2$`]`	

表 4: 推定・検定のためのコマンド

パッケージ　*Statistics'Confidenceintervals'*
`MeanCI[data]`
`MeanCI[d, KnownVariance->v]`
`MeanDifferenceCI[d1,d2]`
`MeanDifferenceCI[d1,d2, KnownVariance -> {v1,v2}]`
`VarianceCI[d]`
`VarianceRatioCI[d1,d2]`
パッケージ　*Statistics'HypothesisTests'*
`MeanTest[d,mu]`
`MeanTest[d,mu, KnownVariance->var]`
`MeanDifferenceTest[d1,d2,diff]`
`VarianceTest[d1,var]`
`VarianceRatioTest[d1,d2]`

パッケージ (連続型分布) には、予め正規分布、t−分布、F−分布、χ^2−分布 などの標本分布が組み込まれているので、上側・下側・両側確率、上側・下側・両側パーセント点を求めることができる (表 5)。従って、任意の値の自由度についての統計数値が計算可能であるし、プログラミングすることにより、粗データからではなく、標本平均・標本分散などから、直接区間推定や検定を行うこともできる。

4. データの視覚化

グラフ用パッケージを読み込み、グラフを描くことができる (表 6)。BarChart で描くのは棒グラフであるが、ヒストグラムの代用とすることもできる。

標本統計量から分布のパラメータを求め、確率分布をあてはめて、密度関数・分布関数などのグラフを描くこともできるし、データの分布と重ね

表 5: 標本分布

パッケージ　*Statistics'ContinuousDistributions'*
`NormalDistribution[0, 1]`
`ChiSquareDistribution[df]`
`StudentTDistribution[df]`
`FRatioDistribution[df1,df2]`
NonCentralChiSquareDistribution[df,lambda]
NonCentralStudentTDistribution[df,lambda]
NonCentralFRatioDistribution[df1,df2,lambda]
`CDF[dist,x]`
`Quantile[dist,q]`

表 6: グラッフィクス

パッケージ *Graphics'Graphics'*	
`BarChart[data]`	StackedBarChart[d1,d2,···]
`PieChart[d]`	`ListPlot[d]`
`TextListPlot[d]`	`LabeledListPlot[d]`
`DisplayTogether[plot1,plot2,···]`	
`DisplayTogetherArray[{{p1,p2,···},` {···}, ···, {···}}]	
パッケージ *Graphics'Graphics3D'*	
`BarChart3D[data]`	`ScatterPlot3D[d]`
パッケージ *Graphics'MultipleListPlot'*	
`MultipleListPlot[d1,d2,···]`	
MultipleListPlot[d1,d2,···,PlotJoined->True]	

表 7: ユーザー・グラフィックス関数

1 変数データの統計図	
`11.msd[`*dname*, *item*`]`	平均・標準偏差
`12.mvec[`*vec*`]`	平均値ベクトル
`13.psing[`*vec*`]`	比率
`14.distsing[`*vec*`]`	度数分布
2 変数データの統計図	
`21scatter[`*vec*`]`	散布図
`22.ctab[`*vec*`]`	相関表
`23.cortab[`*vec*`]`	相関表 (多数)
多数データの統計図比較	
`31.hist6[`*vec*`]`	ヒストグラム
`32.mvprof6[`*vec*`]`	平均値プロフィール
`33.pcomp6[`*vec*`]`	比率の比較
`34.distcomp6[`*vec*`]`	度数分布の比較
`35.mpc6[`*vec*`]`	平均 (比率) の比較

て描き (表 6)、確率分布のあてはめの程度を検討することもできる。

調査項目 (変数) をいろいろ抜き出し、グラフを描くことによって、データ全体を見渡すことが出来るし、かつ outlier などの発見も容易となる。

グラフィックス・コマンドはオプションが豊富に用意されており、凡例・カラー・線の太さと種類・ラベル・テキスト挿入などを自由に設定することにより、様々な美しいグラフを作成し、プレゼンテーションの効果を上げると同時に、EDA としても活用することができる。データから探索的にさらに多くの情報を引き出すことが、本データベースの目的でもある。

5. ユーザー・グラフィックス関数

データの種類、例えば、連続量データかカテゴリデータかにより、描くグラフの種類も異なってくる。どのような種類のデータに対して、どのようなグラフが描けるかを、先の標準パッケージをもとに、描画し易いようにユーザー関数としてプログラミングを行った。

社会調査の場合、属性データの分類に際しては、0,1,2,··· の離散型データが用いられ、名義尺度として数量化されるケースが多い。さらに順序データ、連続量データであっても離散型で処理されるケースも多く、属性データの処理は基本であろう。

連続量データの場合でも度数分布表に集計して、順序尺度として使用されるケースが多い。従って、多用されるのも、カテゴリカルデータのグラフ化である。

統計処理としては、度数の集計、度数分布、分割 (関連) 表などであり、そららの図示として、円グラフ、帯グラフ、分割表の 3D 表示などが用いられる。

連続量データの全体像を見るためには、粗データをそのまま用いて、散布図を描くこともできるが、プレゼンテーションを工夫して、3 次元 (立体) 散布図を描くことも可能である。

さまざまなグラフ化に対応するため、標準グラフを活用したユーザ関数を作成した。パッケージを使用しているので、それらの読み込みが必要である。

データセットから項目 (変数) を指定して、平均・標準偏差を算出する関数が msd である。データ名を *dname* で与え、項目を *item* で選出する。*item* が複数に及ぶときは、mvec を使用する。このとき関数の引数は *vec* のみであるので、関数を呼ぶ前にデータセット dname[1]=*datasetname* と項目番号 *vec=itemnumbers* を予め設定しておかねばならない。*vec* はリストである。mvec はグラフを描く。

データが (0,1) の場合は、回答 (yes=1) の比率を求め、ヒストグラム (棒グラフ) を描く。

下位反応カテゴリの度数を集計し、比率を求め、分布状況についてのヒストグラムを描くのが psing である。これらは複数項目の処理が可能であり、事前にデータセット名と処理項目番号を選定しておく。

2 変数の連続量データについて、平均・標準偏差・共分散・相関係数を算出し、散布図を描くのが scatter であり、項目を *vec* で選択する。選択

された *vec* の項目から 2 項目づつを取りだして処理する。

離散量データの場合、セルの度数のカウントとなり、散布図では度数が 1 点としてしか図示されないので不合理である。図示方法を変更し、3 次元立体図と 3 次元表面図で図示するのが ctab である。1 項目の単純集計 (marginal distribution)、2 変数のクロス集計 (joint distribution) の度数も出力する。単純集計のグラフでは標準関数の制約から、2 項目のみの描画ができず、従って 1 項目を反復して、並べて描くこととする。

ctab は 2 項目間の処理のみ、cortab は多数項目を *vec* で与え、その中から順次 2 項目の組み合わせを取り出して処理するものである。

以上は単独のデータセットについての処理であるが、それが複数となったとき、データ間の比較を行うためのユーザー関数が、番号 31 以下である。

まず、hist6 は度数分布を集計しヒストグラムを描く。mvprof6 はで指示する項目の平均値を求め、それのプロフィールを描く。pcomp6 は比率の比較、distcomp6 は度数分布の比較である。mpc6 は平均値 (または比率) の比較を行う。

予めデータセット名 (dname[i]=$filename_i$) とその数 ($ng \leqq 6$) を与え、処理する項目を ($vec=i_1, i_2, \cdots$) で与える。データセットの数の上限は 6 である。あまり多くのグラフを重ねると相互の比較が困難となるが、上限の 6 をプログラム内で修正することは容易である。

グラフはいずれも折れ線グラフを描き、相互比較の便を期す。

6. 多変量解析

現在のところ、多くの多変量解析の手法は *Mathematica* に組み込まれておらず、他の統計ソフトを使用しないのであれば、自らプログラミングせざるをえない。行列やベクトルの積、逆行列などの行列演算や、固有値・固有ベクトルの計算などは *Mathematica* 内に定義されているので、それらを用いれば多変量解析のプログラムを書くことは、さほど困難な作業ではない。むしろ、*Mathematica* に内蔵のプログラム言語では、ベクトル演算がサポートされており、一般の言語より短く、簡単に書くことができる。

これらを発展させ、多変量解析の例として、重回帰分析、判別分析、主成分分析のプログラミングを行った。必要に応じてデータベースから簡単にアクセスすることができる。さらに、FORTRAN, C, Pascal などの一般の言語で書かれた実行形式のプログラムをリンクさせて、用いることもできる。

7. 具体例

社会調査の一例として、大阪大学生に対して実施された情報処理教育に関する調査結果[1] を使用する。調査項目は以下の通りである。

- 学部・専攻などのフェースシート項目
- 大学入学以前のコンピュータ経験
- 入学後の計算機使用の実態
- 計算機に対する意識
- 情報処理教育に対する意見

全学部 1 年生のランダム・サンプルとして 999 名 (在籍学生のほぼ 35%(文科系 489 名、理科系 510 名)) のデータを入手し、計算機内に保存し、データベースとした。

調査項目は、ID 番号をも含めて No.1 ～ No.106 である。5 段階表示の順序データ、(0-1) 表示のカテゴリデータなどが混在している。

サンプルは、例えば (1) 学部別、(2) 文科系・理科系別、(3) 全調査対象を合併したもの、などの構成とすることが可能であり、3 群毎にファイルを作成しておけば、必要に応じて目的のデータを読み込むことができる。

基礎統計量の計算、単純・クロス集計、グラフ化が可能であるが、一例として、図 1 ～図 6 に各種のグラフを示す。

文　献

[1] 松浦敏雄他 (1993) 教育用計算機システムの運用と学生の意識, 行動計量学, **40**, 17-31.

[2] Phillip Boyland (1992) Technical Report, Guide to Standard Mathematica Packages, Wolfram Research.

[3] Stephan Wolfram (1988) Mathematica, Addison-Wesley, 訳書 (1992) アジソン・ウエスレイ.

[4] Tukey, J.W. (1977) Exploratory Data Analysis, Addison-Wesley.

[5] 吉田光雄 (1993,95) 映像情報データベースの開発, 文部省科学研究費補助金・重点領域研究「情報化社会と人間」, 研究成果報告書.

565 吹田市山田丘 1-2, 大阪大学人間科学部
Tel: 06-879-8051, Fax: 06-879-8054
E-mail: yoshida@hus.osaka-u.ac.jp

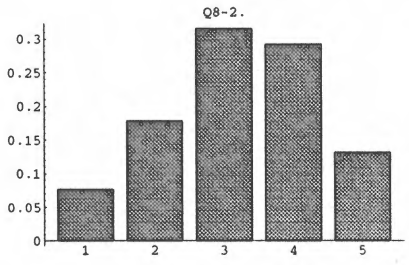

図 1: Q8-2. コンピュータが好きだ (文科系、賛成の程度)

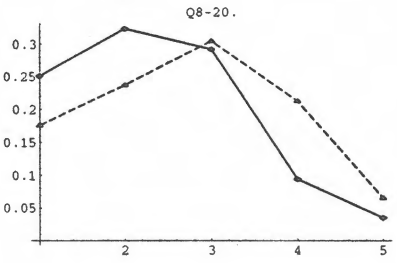

図 2: Q8-20. 卒業後はコンピュータを使用する仕事をしたい (文 (実線)・理 (点線) の比較)

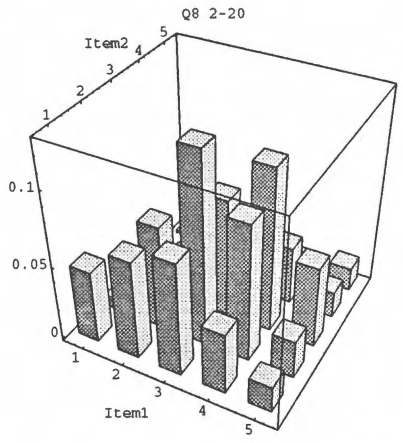

図 3: Q8 2(Item1)-20(Item2) のクロス集計 (文科系)

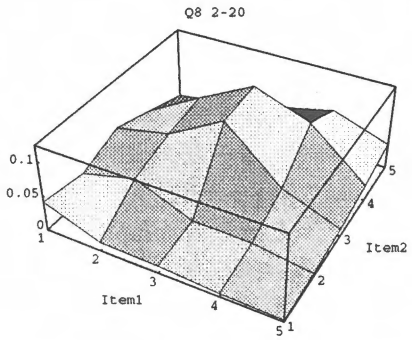

図 4: Q8 2-20 のクロス集計 (理科系)

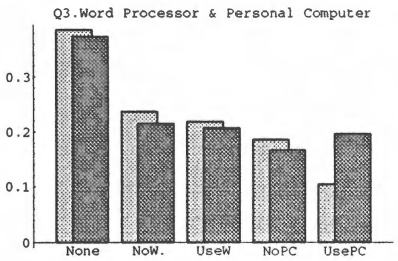

図 5: Q3. 自宅におけるワープロ (W)・パソコン (PC) の有無と使用 (文 (淡)・理 (濃))

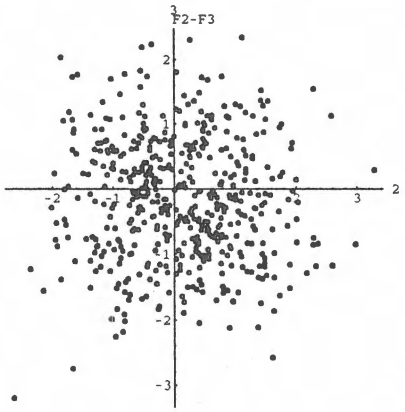

図 6: 主成分得点散布図 (文科系・F2-F3)