# 普通ではないデータの普通の解析法

北海道大学　情報基盤センター
先端データ科学研究室　水田 正弘

## １．はじめに

最近、「仮説」をキーワードとしたビジネス書を数多く見かけます。「99.9%は仮説」(竹内, 2006)では科学的な考え方におけるいわゆる定説について議論し、仮説こそ科学の基本であると主張しています。「仮説思考と分析力」(生方, 2010)では、仮説・分析さらには仮説の検証に基づく行動について扱っています。「仮説力を鍛える」(八幡, 2007)では、実例を通して、適切な仮説を立てて行動することを説明しています。

これらの「仮説」は、データ解析における「モデル」に対応すると考えられます。数値などの集まりであるデータからモデルを構築したり、逆にあるモデルとの適合性を議論したりすることがあります。このとき、基本的なことはデータをどのように記述するかという問題です。モデルもデータの記述方法に依存します。

初等的な統計学の教科書では、$n$ 個の数値が得られたと設定し、その平均や分散を計算させます。さらには、ヒストグラムなどで分布の状況を視覚化します。多次元データ解析では、$p$ 個からなる数値の組が $n$ 個あるデータ、すなわち $p$ 変量データを想定し、変量間の関係をはじめ多くの事項を探索・検証します。このような $n$ 個の $p$ 変量で記述されるデータを「普通のデータ」と呼ぶことにしましょう。

シンボリックデータ解析 (Symbolic Data Analysis) の提唱者である Diday 教授は Complex Data (複雑データ)を「通常の $x$ 変数からなるデータと扱うことのできないデータ(Any data which cannot be considered as a standard “observations x variables” data table.)」と定義しています。すなわち、「普通ではないデータ」です。「普通ではないデータ」を解析するためには、それぞれの記述方法に従って解析手法を考えなくてはいけなくなります。これは、多くのデータ解析のユーザにとっては大変な作業です。しかし、幸いなことに、「普通ではないデータ」に対する多くの解析方法が提案され、さらには統計解析システムで扱うことができるようになっています。すなわち、「普通の解析法」により解析できます。

本報告では、基本的な「普通ではないデータ」とその解析法の紹介から始めます。また、より高度な「普通ではないデータ」についても扱いたいと思います。

## ２．もの同士の関係を表したデータ

各解析対象(個体)を個別に測定することは出来ないが、2 つ個体の関係を測定できる場合がよくあります。例えば、2 国間の貿易量、2 人の顔の似ている度合いなどです。そのようなデータを直接、解析するのは容易ではありません。でも、安心して下さい。「普通の解析法」としては、クラスター分析と多次元尺度構成法があります。

クラスター分析は、似ているものをグループにまとめ、個体を分類する手法です。特に、階層的手法では、個体間の類似度（または非類似度、距離など）が得られた場合、個体をいくつかのグループに分けることができます。階層的手法としては、最短距離法、最長距離法、ウォード法などがありますが、どれも S-PLUS を含む大部分の統計解析システムで実行できます。クラスター分析の結果はデンドログラム(樹状図)で表示することができます(図 1)。ただし、クラスター数は利用者が決めなくてはいけません。

多次元尺度構成法は、個体間の類似度から、その類似度を表現する空間配置を求める手法です。別の言い方をしてみます。地図に描かれた 2 つの都市の距離を求めるのは非常に簡単です。定規や巻尺を使うこともできますし、緯度経度や(数学的な)座標が分かっていたら、ピタゴラスの定理などを使い電卓でも計算できます。ところが、全ての都市間の距離がわかっていても、地図を再現するのは大変な作業です。この大変な作業を実施するのが多次元尺度構成法です(図 2)。入力された類似度と点間の距離との近さの基準により、値自体の近さに着目した計量的多次元尺度構成法と、値の大小関係に着目した非計量的多次元尺度構成法があります。

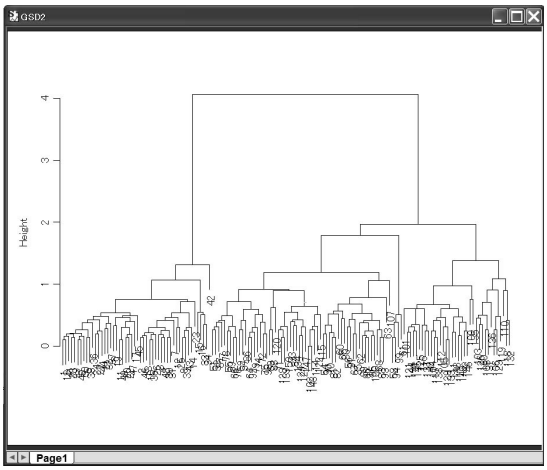


図１．樹状図の例

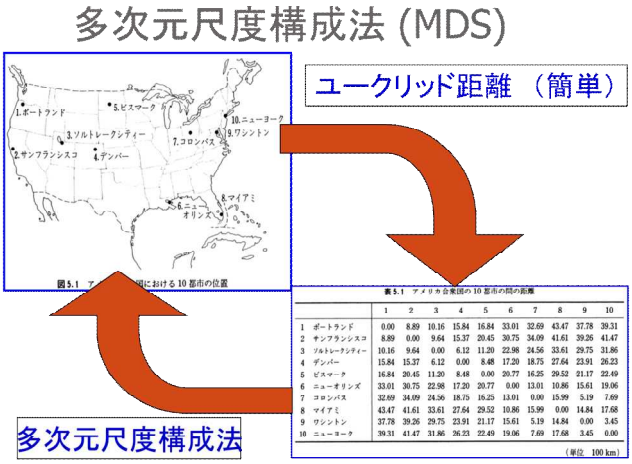


図２．多次元尺度構成法とは？

## ３．時間的に変化するデータ

観測対象が時間や周辺の環境に従って変化する場合があります。経済指標や気温などが典型的な例です。これらのデータを時系列データと呼び、多くの解析方法が提案されています。さらに、時間に限定せずに、ある量に従い変化するデータを「関数」として解析する関数データ解析(Functional Data Analysis)が Ramsay 教授と Silverman 教授により提唱されています(Ramsay and Silverman, 2005)。関数データ解析について簡単に紹介いたします。

通常、データが関数として得られることはありません。しかし、ある間隔で関数の値がサンプリングされていると解釈できる場合があります。そこで、関数データ解析における初めのステップは、サンプリングされた値から関数を生成させることです。従来から研究

されてきた関数当てはめの手法、例えば、B スプラインやフーリエ関数による補間を使うことが出来ます。得られた関数を関数データと呼びます。関数データに対する解析方法としては関数主成分分析法、関数回帰分析法、関数クラスター分析法など多数、開発されています。これらの手法の基本的なアプローチは、関数を基底関数により展開し、その係数を通常のデータ解析法に適用することです。関数データ解析を実行するためのプログラムも公開されていますので、「普通の解析法」となっています。詳しくは、Ramsay and Silverman(2005)による成書および http:// www.psych.mcgill.ca/misc/fda/が参考になります。

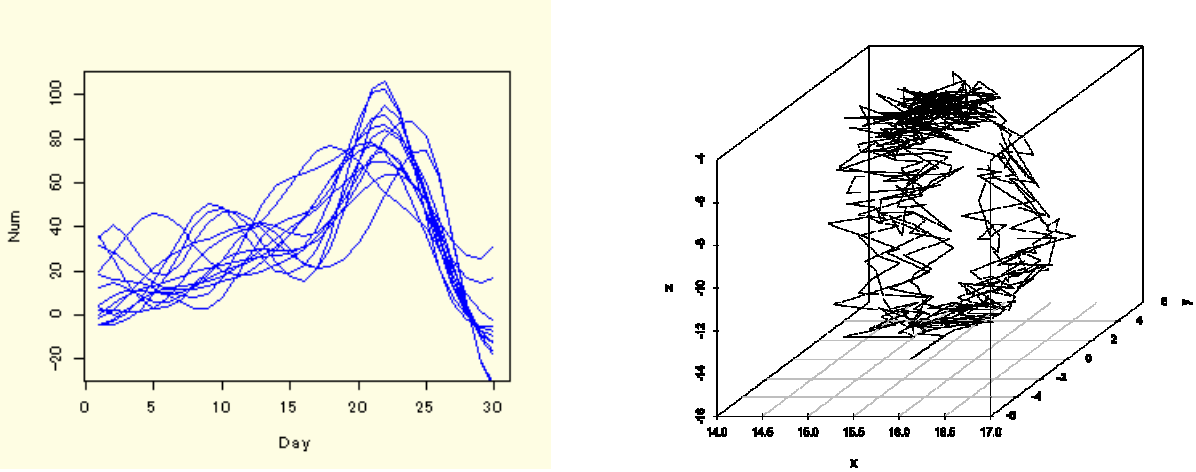


図３．関数データの例

## ４．集まりを表したデータ

インターネットやデータベースから大量のデータを容易に入手できる時代になってきました。それに伴い、2つの問題が発生しました。1つは、あまりにも大量なデータであるため、解析の目的によっては通常のコンピュータでは解析ができないという問題です。第2の問題は、いろいろな型のデータが混在することが多くなったことです。これらの問題に対する方法として、Diday教授は、シンボリックデータ解析法(Symbolic Data Analysis)を提唱しました。この名称は少し混乱する可能性があります。特に「シンボルのデータの解析法」ではないことに注意してください。

「シンボリックデータ」とは、通常の観測対象のクラス(集まり)の内部における変動を考慮したデータ(Any data taking care of the variation inside classes of standard observation.)です。個体を*p*変量の値で記述すると、各個体の内部の変動が考慮されないことになります。個体を区間値やヒストグラム、分布値などで記述することにより内部の状況を表すデータ、すなわちシンボリックデータとなります。さらに、シンボリックデータでは、各個体の記述として、連続的な数値や離散的な量、さらには区間値などを同時に利用することもあります。シンボリックデータの概念により、集合やクラスを解析対象とすることができます。より一般的には、コンセプト(Concept; 概念)を解析対象としています(図4)。

ただ、残念ながらシンボリックデータ解析の具体的な手法の開発は限定的であります。

研究の中心は、区間値で記述されたシンボリックデータの解析法が大部分です。シンボリックデータを解析するためのソフトウェアとしては、SODAS (Symbolic Official Data Analysis System)が無料で公開されています(Diday and Noirhomme-Fraiture; 2007)。

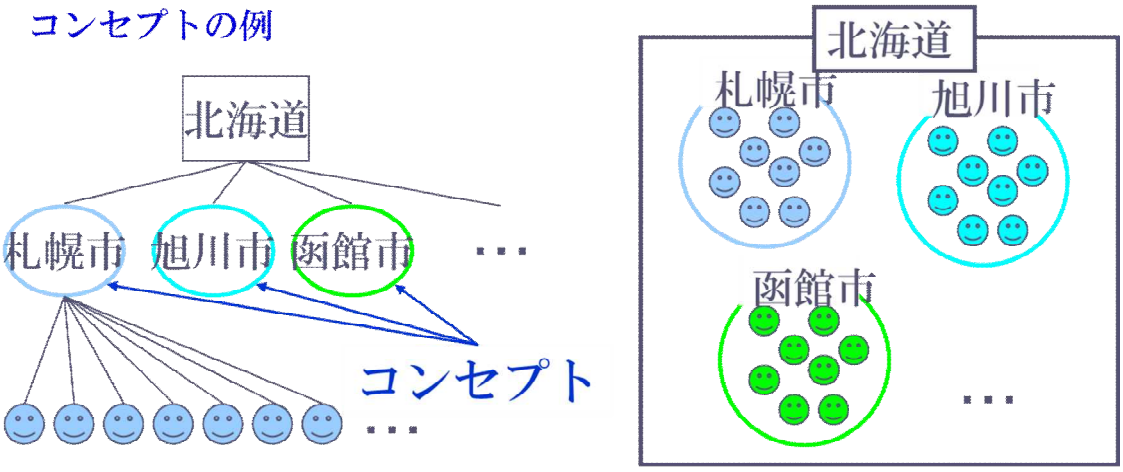


図４．シンボリックデータ解析におけるコンセプトのイメージ

## ５．おわりに

今日、私たちが直面しているデータは複雑になっています。それに伴い、データ解析の研究者および統計解析システムの開発者は新しい手法を開発し、コンピュータ上に実装する努力を続けています。「普通でないデータ」と思われるものも「普通に解析」できる場合があります。データを解析する場合、何であれば既存の手法で対応できるかを把握することは重要です。本報告で紹介した以外にも多くの解析方法が開発されています。それらを全て理解するのは不可能ですが、一度、代表的な統計解析システムに実装されている手法の一覧を眺めてみるのはいかがでしょうか？

## 参考文献


Ramsay, J. O. and Silverman, B.W. (2005) Functional Data Analysis -2nd Edition, New York: Springer.

Diday, Edwin and Noirhomme-Fraiture, Monique (2007), Symbolic Data Analysis and the SODAS Software, Wiley.

水田正弘・山本義郎・南 弘征・田澤 司(2005), S-PLUSによるデータマイニング入門, 森北出版.

竹内 薫(2006), 99.9%は仮説 思いこみで判断しないための考え方 (光文社新書).

八幡紕芦史(2007), 仮説力を鍛える (ソフトバンク新書 51).

生方正也(2010), ビジネススクールで身につける仮説思考と分析力―ポケットＭＢＡ〈5〉(日経ビジネス人文庫).