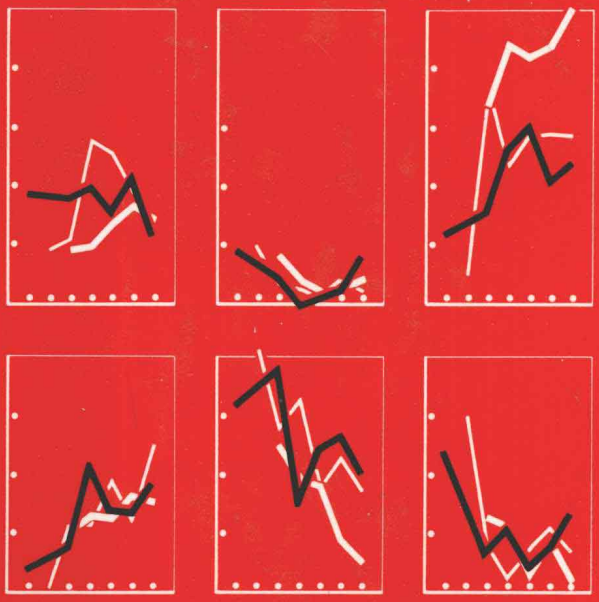
データ解析の考え方

林 知己夫 著



東洋経済新報社

データ解析の考え方

林 知己夫 著

東洋経済新報社

著者紹介

1919年　東京都に生まれる．
1942年　東京大学理学部数学科卒業．
現　在　文部省統計数理研究所長．
著　書　『数量化の方法』東洋経済新報社，『多次元尺度解析法』（共著）サイエンス社，『日本人研究』（共著）至誠堂，『比較日本人論』（共著）中央公論社など．
現住所　三鷹市井の頭2-23-11

データ解析の考え方

昭和52年6月20日発行

著　者　林知己夫（はやしちきお）
発行者　宇梶洋司
発行所　東京都中央区日本橋本石町1の4　東洋経済新報社
郵便番号 103　電話 03 (270) 4111(大代表)　振替口座東京3-6518

3033-3952-5214
Printed in Japan

序　　文

本書は，前書『数量化の方法』1974年に続くものであるが，数量化をも含めてデータによる現象解析(データ解析と略称する)の考え方，方法論，方法についての基本，を述べたものである．主として前書以後に発表した小論をまとめたものであるが，「数量化」以外のものについては，古いものでも前書にもれた部分をここに収録したものもある．

データ解析は，理論と実際とが一体となってはじめて有効なものであり，これには体験や統計的方法以外の知識や洞察力，知恵，勘のよさ，いわゆるセンスをも必要とするものであって，きわめて人間的なところがあるといわざるをえない．ここにおもしろさもあるのであるが，こうしたデータ解析を定型化して述べることはいまのところ私には困難なのである．それほど体験や知識の集積があるわけではなく，データの性格もさまざまなので，すべてを知って整理するということも，現実的にはできうべくもないので，「実例を通してその核心を感得してもらう」といういき方よりほかに手がないように思う．

私自身，新しいデータに接することによって衝撃を受け，データ解析の方法を工夫していくという道をたどりつつある．たとえば，日本人に対する社会調査のデータをみていただけでは突っ込めなかったことが，ハワイの日系人調査

データをみることによって衝撃を与えられ，目がさめたような気持になって新しい視点を得た．この視点から日本人に対する社会調査データを見返すことによって，これまでわからなかったことが浮かび上がり，目前の鱗の落ちる思いをしたものである(第9章)．

ここでは，私のいろいろの体験を通して感得したことをまとめてみることにした．それらはその都度，機会を得て発表したものである．全体を整理してまとめ直すことも考えてみたが，それよりも過去にそれを書いたときの熱気というものが現われているほうが望ましいと考え，そのまま収録し，いくらか手を加えるという形をとった．

第Ⅰ部は，実際的な研究を進めるについて，データ解析の基本的な基盤になる考え方をまとめた．すなわち第1章の「データとは何か」は，データによる現象解析の基礎となるデータについての考え方をまとめておいたもので，全体への序曲ともいうべきものである．次の二つの章は予測に関するものである．前者の「予測について考えること」は，前著の予測に関する考え方に関係が深いのであるが，それに続くもので，過程制御という立場からの見方を加えた．これは，データに基づく予測と行為に関係することであるが，ある現象においてははっきり予測し，それに基づいて行為するということができにくい．したがって，その場合には試行錯誤の過程の高能率化をねらわざるをえない情況にいたるということになるが，それにおける予測についての考え方を示したものである．こうしたものも複雑靱強な相手に立ち向かうとき必要な態度であろう．「予測におけるベイズ推論」では，予測(推論)における事前確率とそれを用いる考え方を示したもので，予測をまじめに取り上げるとき避けて通れないところである．実際のデータ解析においては，いわゆる理論からの「背のび」「はみだし」をしなくてはならないので，これを現実の場でどう処理するかというところが問題点となる．逐次近似の形で一歩一歩進むより仕方のないところである．

次の第4章が集団と個の問題であり，集団の統計的分析とその結論の個への還元のあり方というところに帰着する．これを裏づける確率にもう一度基礎的

な考察を加え，永遠の課題ともいうべき集団と個との関係，そのデータ分析を通してのかかずらいに思いをいたしたものである．

第Ⅱ部は数量化に関するもので，前書の直接の続編でもある．第5章の「心理学にとって数量化とは何か」は，前書の「人文・社会科学における方法論的諸問題——統計数理の立場から」に続くものであるが，ここでは心理学が数量化によって得たもの，失ったものについて考えをめぐらし，データ解析の立場から心理学において数量化はいかに考えるべきかを論じた．ここに数量化というのは，狭い意味の数量化ではなく，広義のものを意味する．

第6章の多次元尺度解析——MDA-OR と MDA-UO は Multidimensional Scaling (MDS) の一環としてつくられた Minimum Dimension Analysis (MDA) の二つの種類 -OR と -UO の特色を示し，その使い方を述べたものである．前書の「質の数量化は進む」において示された MDA は，ここでいう MDA-OR のことである．二つのものの間の関係が名前だけの分類である場合を取り扱うのが MDA-UO である．この二つは，はなはだ異なった特色をもつものであるので，その区別を理論と実際において知ることが大事であると考え，実例について特色と使い方を示した．また MDA-OR の方法の使い方の注意——MDA-OR の方法のもつ性格からみて，好ましくない使い方となるようなデータのあること，そのような場合の処置など——を示し，MDA-UO の具体的な例をも示しておいた．これらは MDA にかぎらず多次元尺度解析を用いるときの注意ともなるもので，データ解析とデータの性格に関する議論の一端をなすものである．

第7章は第6章に続くもので，多次元分析法のうまい適用とそれについての注意を述べたものであるが，併せてデータをいかに正しく取るかの重要性を示したものである．正しく取られていないデータをいくら形式的に解析しても無意味であると知りながらも，解析の技術のみに酔う楽しみを忘れえない向きもあり，またそれを利用する人々もあるので，あらためて「データを取ることの重要性」を強調しておいた．

第8章は昔の研究であるが，データ解析のときの考え方としてたいせつな意

味をもつところがあるので収録した．これは尺度点のつくり方を統計的にはどう考えるかという問題であり，こうしたことの考え方の応用は広いと考えたからである．

第9章は，社会調査データの解析から質的な知見を得た例であって，いわゆる数量化の方法によって“思想”を描くことができる，という一つの物語というべきものである．えてして統計的分析は平板できめが粗いといわれるが，こうした方法は，それを乗り越えようとする一つの努力であるとみなされよう．

第Ⅲ部は，データ解析が新しい分野に広がっている現状から，適宜何らかのヒントになるものを選んでおいた．まず，第10章から第12章までは医学関係のデータ解析についてのものである．疫学的研究，フィールド調査，診断とか治療の科学という問題を取り上げた．これは，私が医学方面の方々と共同研究をした結果，感じとったもの，あるいは患者として感じたものに基づいている．共同研究の範囲は，公衆衛生方面，循環器に関する健康管理，循環器疾患方面，眼科，別の見方をすればME関係などである．医学は人間というかけがえのないものを取り扱うむずかしさがあり，これと統計的解析の冷徹さをどう調和させていくかという重大問題があるので，データによる現象解析としていかなる立場に立つべきかを深く考えなくてはならなくなる．

第13章の前半は社会調査に関するもので，比較文化(社会)研究における問題点を検討したものである．容易なようであって比較はむずかしいものである．何のための比較，比較とは何か，という根本問題に立ち入らねば研究は始められるものではない．こうした点に触れたものである．後半部分では，標本調査をうまく運用すれば容易なのであるが，まずく行なうとたいへんな手間を要し，しかも正確な数字を得ることがむずかしいという問題について述べた．また「知りたい」ことの表現をいかに客観的にとらえるかという問題について考察し，具体的目的意識，問題意識と調査との関係について論じたものである．データをいかに取るかは，データ解析の重要な点であるので調査に注意を喚起する意味で本章と次の章を加えておいた．

第14章は型変りの標本調査の例である．この型変りの一つは，前書で「確

率によってノウサギの数を知る――生態モデルとランダムネス」において示したが，この研究はその後，個体数動態の問題に発展している．このため野性動物の野山における年齢分布を知ることがきわめて重要で，これが臼歯を支える骨の断面に現われる縞によって知りえられることが，共同研究者の柴田義春氏(林業試験場北海道支場)と北大歯学部の大泰司紀之氏によって見出された．これに従って目下研究が進められているが，ノウサギの年齢分布はきわめて若いほうに偏り，平均年齢は1.5年前後という短いものであることが見出されている．このほか足跡調査，被害調査，捕獲調査が同一地区(新潟県名立地区，村上地区)で共同研究者の豊島重造氏，高田和彦氏(新潟大学農学部)によって繰り返され，データが蓄積されつつあり，併せて population dynamics 解明の道が開けつつある．これができてはじめて，野性動物管理の基礎ができあがることになる．さて，14.2節では，ソシオメトリック指標の標本調査に関するものであって，関連標識に基づく一つのサンプリング調査のモデルである．また14.3節は境界を推定するための一つのモデルであって，こうしたことの要望も，計量的生態研究が進むにつれてしだいに高まってくると思われる．

第15章以下では，データ解析と関連の深い諸学問との関係について考える．まず第15章は情報科学に関係する問題である．ここでは情報科学において統計的考え方や方法がいかなる役割を演ずるか，統計的方法を用いなければ，情報科学のソフト的な部面は有効妥当なものにならないことを論じ，情報科学によって，統計はいかに発展していくべきものかを考察したものである．第16章では，データ解析の立場(利用者側)よりするコンピュータ観であって，コンピュータの専門家とは異なった見解があろうかと思う．

最後の第17章は学問の方法についてのものである．私が現在，専攻分野として強く関係している統計学，行動計量学(Behaviormetrics)，科学基礎論の三つについて，データを通してものをみる．詳しくいえば，データによる現象解析ということを通してものをみる私の立場から，それらの学問の目ざすものを考えてみたものである．学問がそうしたことの方向へ向かうとき，データ解析にとって非常な幸いと考えられるのである．

以上が，本書の内容概観である．ここでも前書と同じく，個々の具体的問題の解決に心身を使って格闘することによって，一般的なもの，体系的なものが感得され，それが外に向かって表現されてくるという姿勢で貫かれている，といってよい．

昭和52年4月

林　知己夫

目　　次

第 I 部

データ解析の基盤

1

データとは何か

1.1 データとデータ解析

ある要素 i ——これは即物的なものであっても論理的なものであってもかまわない——を測定し，標識づけを行ない，X_i と表現したもの($i=1, 2, \cdots, N$; N は要素の数)，あるいはN個のXを分析して出された結果をデータという．このXは——添字 i を落として一般的表現とする——数量的表現でも質的(属性的)表現であってもよい．数理統計学では，ユークリッド空間内のベクトルXで表現するというところから論述が始まるわけであるから，何がデータであるか，の意識が稀薄になるおそれがある．D. J. Finney(エディンバラ大学)は data, number の区別を強調しており*)私もその論旨には同感である．

表現されたものは数であるとしても，データはそれをそこまで表現するにいたった過程をもち，したがって表現されたものの性格があるわけであって，これに基づいて情報が紡ぎ出されねばならないわけである．数理統計学では，X

*) D. J. Finney, "Problems Data and Inference," *Journal of the Royal Statistical Society*, Series A, 1974, および "Numbers and Data", *Biometrics*, Vol. 31, 1975. なお氏が来日の折の講演の翻訳は，「数，データ，推論」(日本マーケティング・リサーチ協議会，昭和51年7月)として印刷されている．