多変量解析*

—その起源と発展に関する回想—

C. Radhakrishna Rao

University of Pittsburgh
and Indian Statistical Institute

訳者 柳井 晴夫 竹内 啓
千葉大学文学部 東京大学経済学部

*1981年11月20日にコネチカット大学 (University of Connecticut) で行なわれたテレビ講演の内容をアメリカ統計学会の保存のために記録した Pittsburgh 大学 Center for Multivariate Analysis のテクニカルレポートNo. 82-23 を著者の許可を得て翻訳したもの。なお、翻訳に際して、講演の最初と最後の謝辞に関する部分は割愛した。

図1. ゴールトンによる親と子供の身長に関する散布図

多変量正規分布は19世紀におけるガウス(Gauss), ブラヴェー (Bravais), ショルス (Shols), ゴールドン (Galton) およびエッジワース (Edgeworth) の仕事にその起源をたどることができる。しかし、これらのうちもっとも重要な人物は遺伝法則に関する知識の探究によって、多変量正規分布の理論と応用の発展のきっかけをつくったフランシス・ゴールトンであろう。ゴールトンは1885年9月王立人類学研究所で行なわれた講演¹⁾で、親と成人の子供の身長からなる変数のクロス表(図1参照)を用いて次のような観察を行なった。

- (i) 親の身長が与えられた場合の子供の身長の条件付き平均値は、ほぼ直線となる。

1983年8月10日受付

(ii) 親の身長が与えられた場合の子供の身長の条件付き分散は、ほぼ一定となる。

(iii) 2変量分布によって規定される等確率の点の軌跡は、ほぼ楕円形になる。

上記の3つの特徴、すなわち、回帰の直線性、等分散性および等確率点の楕円形、の性質から、図1の分布が2変量正規分布になることは自明であるが、不思議なことに、ゴールトンは、自分自身でこの点に気づくことができなかった。しかし、ゴールトンが上記の3点についてケンブリッジ大学の数学者ハミルトン-ディクソン(J. D. Hamilton-Dickson)に話したところ、彼は即座に図1の分布が2変量正規分布になることを示した。

ゴールトンは、図1の身長データの対する観察から、現在の回帰理論の発展に貢献したいいくつかの興味深い知見を見だしている。まず彼は、親の身長が極端に高くなった場合、子供の身長の条件付平均値は、親の身長から離れて全体の平均値に近づく、先祖がえり(reversion)の現象を発見した。このreversionが後になって回帰(regression)と呼ばれたもので、ゴールトンはその強さを表わすための数量的指標として r を導入した。 r は今日、相関係数を表わす記号として広く採用されているが、カール・ピアソン(Karl Pearson)によれば、相関係数の略称 r は、まさにreversionの最初の1字に他ならないものであった。

ゴールトンの提唱した r とその2変量正規分布への関連の仕事は、ウェルドン(Weldon)、エッジワース、カール・ピアソンに受け継がれ、引きつづき研究された。ゴールトンは、 r をco-relationの指標と呼んだが、エッジワースはそれを、言葉としてより重みのある相関係数(correlation coefficient)と呼びかえた。ウェルドンは、2変数の多くのサンプルを用いて、 r の値を実際にいく通りも計算したが、 r についてはゴールトン関数(Galton function)と命名した。このようなウェルドンとエッジワースの研究が下積みとなって、カール・ピアソンとシェパード(Sheppard; シェパードの補正で著名)は、今日広く使われている相関、偏相関および重相関の表現を与えた。また、大標本のもとでの r の標準誤差も与えている。

相関係数 r の分布はゆがんでいたために、小標本、またはある程度の大きい標本の下での推論を行うには、相関係数の厳密な分布を求める必要のあることが認識されていた。このための多くの試みは徒労に終わったが、1915年になってフィッシャー(R. A. Fisher)が、初めて、 r の厳密な標本分布を与えた。フィッシャーが相関係数の標本分布を与えた著名な論文は、多変量データに関する統計的推論を行うための手法を開発したという意味で、多変量解析の研究の嚆矢というべきものであろう。 r の正確な分布を与えた式はかなり複雑なものであったから、フィッシャーはまもなくフィッシャーの z 変換として知られている変換 $z = \tanh^{-1}(r)$ を導入し、その結果、標本相関係数 r にもとづく種々の推論の方式がかなり簡素化されることになった。なお、フィッシャーは、20世紀の始めに F 分布を導入し、回帰係数、偏相関係数および重相関係数に関する帰無仮説の検定方式を開発している。さらにこの期間にフィッシャーは最尤推定法を導入しただけでなく、実験の計画と分析に関する様々な手法を開発した。1925年に出版されたStatistical Methods for Research Workersは、これらの手法をまとめたハンドブックというべきものであろう。上記の書物は、まもなくポピュラーになり、大量のデータを収集していた生物学者の間でとくにひろまった。そして、データに適切な統計的解析を施すことによって、新しい知識を獲得することができる道が開かれた。現在の視点から上記の書物の欠点を挙げるとすれば、帰無仮説の検定、および、1%と5%の有意水準の使用を著しく強調している点であろう。しかし、付言するならば、フィッシャー自身は、自分の研究において、これらの有意水準を厳格に用いたわけではない。賢明にも、彼は、応用場面において、研究者が観察された検定基準を解釈するための一つの安全な処方せんとして、有意水準の効用を考えていたのであろう。

上記のフィッシャーの書は、広く世界中の研究者の注目をあびた。そして、1925年から1935年におたる10年間の間に、主として彼の影響下で多変量解析にいくつかの重要な進展がみられた。1928年にウィシャート(Wishart)

は、2変量分布に関して得られたフィッシャーの仕事を、一般の p 変量正規母集団から得られた標本分散共分散行列の同時分布に拡張した。ホテリング (Hotelling) は1931年に学生 t 分布の多変量への拡張である T^2 統計量を導入している。また、彼は1933年に主成分、1936年に正準相関の概念を導入した。ウィルクス (Wilks) は1932年に多変量正規分布の仮定のもとで、平均ベクトルと共分散行列に関するいろいろな仮説を検定するために、尤度比基準の研究を開始した。フィッシャーの助言によって、マルチン (Martin) は、1936年にある墓地で発見された顎骨 (がくこつ) が男性のものであるか女性のものであるかを判定するために線形判別関数を構成した。スミス (Fairfield Smith) は同年に、植物選択 (plant selection) のための判別関数を発表している。なお、この研究は、現在、経験的ベイズ推定²⁾と呼ばれているものの先駆的研究でもあった。

アメリカ人であるホテリングとウィルクスが有意性の検定に興味をもっている間に、インド人であるマハラノビス (Mahalanobis) は、いくつかの異なる母集団の相互関連の研究に関心をもっていた。マハラノビスは、この目的のために、2つの母集団の距離を示す指標であるマハラノビス D^2 を考案した。なお、このマハラノビス D^2 は、異なる種族間の類似性の尺度であるピアソンの C^2 の1つの改良でもあった。1938年にボース (Bose) とロイ (Roy) はマハラノビス D^2 の対立仮説 (二つの母集団の平均ベクトルが等しくない場合) のもとでの分布を導いたが、それは、10年前にフィッシャーによって発見された重相関係数の分布と同じ形をもつものであった。なお、マハラノビス D^2 と重相関係数の分布の類似性は、1960年になってバウカー (Bowker) が D^2 を2つの独立なカイ二乗統計量 (一方は非心カイ二乗分布に従う) の比で表わすというエレガントな証明によって、初めて明らかにしたものである (訳者注1)。

フィッシャーは、重相関係数の分布を導くために、幾何学的な議論を展開しており、その証明は、フィッシャーの多くの著作と同様に、途中のステップが省略されていたため、理解するのに大変骨の折れるものであった。マハラノビスは、後にフィッシャーの伝記を概略した際、「フィッシャーは、彼の心の中で問題が解けるまでは決して解析の結果を書きおろそうとはしなかったので、ときおり、問題解決に至る重要な鍵を忘れてしまうことすらあったと告白している」と述懐している。

フィッシャーは、2群の判別分析の理論を3つ以上の母集団が存在する場合に一般化して、行列方程式の根に関する仮説のもとでの標本分布を導き、それを1939年に発表した。上記の結果導かれる根はウィジャート分布に従う2つの独立な確率変数行列の一方に対する他方の行列の固有値に他ならないものであった。上記の分布は、同年に、ロイによって導かれた2つの分散-共分散行列の同等性に関する仮説のもとでの検定基準と同一のものであったが、これがフィッシャーが導いた標本分布に関する最後の主要な仕事になった。

上記の研究の流れが、1940年に至るまでの多変量解析の主要な発展を特徴づけるものであったが、多数の測定値を分析するもうひとつのタイプの研究に因子分析があった。因子分析は20世紀の初めにスピアマン (Spearman) によって創始され、1935年に“Vectors of Mind”という奇想外なタイトルのつけられた書物が出版された (訳者注2) もの、統計学者からはあまり注意が払われなかった。おそらく、因子分析モデルにおいては、多数の未知パラメーターを推定しなければならないこと、主観的な因子の解釈、さらには、因子分析によって得られた結果の有用性についての疑わしさなどがその理由であろう。フィッシャー自身の著作には、既存の因子分析の手法について言及したものはないが、1918年に書かれた親族の相関に関する古典的論文の中で、人間の測定値の変動を説明するために、因子分析型のモデルを仮定している。しかし、通常の因子分析家と異って、フィッシャーは実際の因子数と効果的因子数の区別に慎重で、抽象的ではあるが後者が推定可能でしかも概念的意味をもつものであると述べている。

1940年に、私は統計学を特定のテーマとして学び終え、数学科を卒業した。戦時中であつたので、数学科の学生の就職には乏しく、数学の研究を続けていくための奨学金も十分ではなかつた。どこにも雇用されないままに約6ヵ月が過ぎたとき、イギリスの軍隊に入隊するか、さもなければ、統計学とは異つたもっと将来性のある研究分野に進むかの二者選択を迫られることになった。そのとき、私は数年前にマハラノビスによって設立されたインド統計研究所のことを耳にした。その当時マハラノビスは、魅力的な職務に就けることを保証して、新しい学問である統計学の勉強を志す多くの優秀な学生を勧誘しているところであつた。私は、これを聞いてただちに統計研究所に入所できるよう志願したところ、マハラノビスは、統計学者を育成する研究所の養成コースに私を採用してくれた。

研究所に入所してまもなく、私はボース、ロイとマハラノビスに会うことができた。その当時ボースは実験計画における組み合わせ論的問題を研究していた。ロイは、以前から行っていた分散共分散行列の同等性に関する仮説検定のための行列方程式の根の研究を続けていた。一方、マハラノビスは、大規模な標本調査に関する方法論とその理論の開発に取りこんでいた。マハラノビスは、日頃から統計学の教育と訓練の相違を強調していたが、ある時私に、統計学の訓練の一環として、人体測定に関するひとつのプロジェクトを与えてくれた。その結果として、私は21のカーストのグループと種族に分類された多数の個体に関する12の形態学上の測定値の分析を担当することになった。この研究において、私はマハラノビス D^2 を主な分析の道具として使用するよう指示されたが、その為には、大規模な共分散行列とその逆行列、および多数の2次形式を計算しなければならなかつた。しかし、その当時研究所に備えられていた卓上計算機では、上記の計算は到底不可能であつた。そこで、私が最初に成すべき仕事は、与えられた計測値(変数)の個数を減少させていかに計算を簡略化するか、を工夫することであつた。その為に変数選択の基準³⁾が必要となつた。

変数選択のための最初のステップとして、私はいくつかの異つた母集団における平均値ベクトルの同等性を検定するための統計量 V_r を開発した。 V_r は異なる母集団の全てのペアについて、選択された r 個の測定値にもとづく D^2 の重みつき平均値である。分散共分散行列が既知、または大標本下での推定値が与えられた場合の V_r の分布は、私の1943年の修士論文⁴⁾で報告され、1945年に1つのノートとして発表された。私が perimeter test と呼んだ V_r 統計量は、今日トレース基準と呼ばれているものに相当し、フィッシャーの導いた多群の判別分析に現われる行列方程式の根の総和に一致する(訳者注3)。私は、この基準を用いて p 個の変数から選ばれる r 個 ($r < p$) の変数のすべての組み合わせについて V_r の値を計算し、 V_r を最大にする変数の組を見出すことを試みた。しかし、この計算も卓上計算機の手にも負えるものではなかつた。そこで、 V_r をすべての変数の組み合わせについて計算せずにすむ何らかの簡便法が必要となつた。

次に私は、測定された変数の個数を減らしても、あまり多くの情報が喪失されないということを確認する必要に迫られることになった。これが契機となつて、私は、判別関数に含められたある変数の集合に、付加的な変数を加えることによって増加する情報の統計的有意性を検定する基準を導き、それを1946年に Sankhya⁵⁾ に発表した。ここで導いた基準は、ある指定された判別関数に対する Fisher-Bartlett 検定のひとつの一般化を与えるもので、この基準を用いることによって、真の判別関数の係数ベクトルが、ある特定の部分空間に含まれているか否かを検定することができる。

大規模な共分散行列の逆行列を求めるかわりに、私は相関のある変数と無相関な変数の集合に変換するために段階的手順を用いた。この変換は、各変数を判別における有用性に基いて序列化し、それを三角型に並べ換えるものであつた。後になって、上記の変換は、グラムシュミットの直交化に他ならないことに気づいたが、この変換に用いられた計算の方法は現在のコンピュータープログラムに採用されているものとはほぼ同等で、数値的にも

安定したものであった。

多群の判別分析の固有方程式の根に関するフィッシャーとロイの研究を参照しながら、多数の測定値の線形結合として得られる一般化された級内相関 (intraclass correlation) として定義される家族相関 (familial correlation) の概念を1945年の論文⁶⁾で発表した。いくつかの世帯における兄弟 (または姉妹) の多数の計測値によって計算される家族相関の値は、フィッシャーの多群の判別問題で与えられたように、級内共分散行列に対する級間共分散行列の固有値の関数となる。

カルカッタの研究所で、人体測定のデータ解析に没頭していた2年間の間に、私は、多変量解析の専門家としての地位を確立することになった。そのために、マハラノビスは、イギリスのケンブリッジ大学の人類学者である J.C. Trevor の要請で私をケンブリッジに送り、イギリスの探検隊がアフリカの Jebel Moya で収集した骨格データにマハラノビス D^2 を適用した解析を行うことを命じた。

私は、1946年の秋にケンブリッジに到着し、人類学の博物館のあるダックワース (Duckworth) 実験室に入った。しかし、その実験室には、骨と石が散在しており、どうみても統計学の研究には縁遠いところのように思われたが、自分に与えられた新しい使命に挑戦する気持で、その実験室に腰を落ち着けることにした。なおこのとき、私のコンパニオンとして、電動式卓上計算機が与えられた。

私のケンブリッジ訪問は大変実りの多いものであった。フィッシャーは、かつて、種族の類似性の指標であるピアソンの C^2 を批判する論文を書き、骨格データに関するすべての研究を不毛なものと非難していたにも拘わらず、骨格データの分析によって Ph.D. 論文を作成する私の指導者になることを承諾してくれた。その間に、私が2つ以上の母集団が関与する場合の判別関数に関するフィッシャーの研究の拡張を行ったことをフィッシャーは大変喜こんでくれた。しかし、フィッシャーは、私をもうしばらく彼のマウス実験室において、染色体地図の作成に従事させることが望ましいと考えた。その結果、私は2つの場所で仕事を行うことになり、ダックワース実験室では、死人の骨を取り扱い、Whitting Lodge にあるフィッシャーの実験室では生きたマウスを用いて、異ったタイプの交配やその子孫について研究した⁷⁾。

イギリスの人類学者によって測定されたアフリカ人の骨格データの場合、すべての測定は発掘された場所で行なわれた。そして、よく保存された骨の一部分のみがケンブリッジの博物館に送られていた。私がケンブリッジに着いた時、計測値はすでに IBM の機械によって処理されており、発掘された場所の深さに応じて骨が3つの水準に分類され、それぞれの特性についての平均値、標準偏差値が計算されていた。

カルカッタで人体測定のデータ解析に従事していた間に、私は、フィールドで収集された大量データの処理にあたって、データの記録の誤り、はずれ値の存在、調査者の個人差などに留意すべきであることを学習していた。そこで、私は IBM の機械から出力された見事な印刷結果を脇において、まず、オリジナルな調査票に誤りがないかを詳しく調べることにした。データの精査、フィッシャーの言葉を借りれば、データの訊問 (cross examination of data) はルーチン化された仕事ではなかった。しかし、統計解析を行なう人は、さまざまな種類の誤りによって生ずるデータの不規則性、明らかなデータ作為にもとづく規則性には細心の注意を払わなければならない。したがって、データの誤りを検出するための手法としての irregularity analysis、データの捏造 (ねつぞう) を検出するための regularity analysis が重要となってくる。後者の1例としては、フィッシャーが行ったメンデルのスイートピーのデータ解析がある。この解析の結果、メンデルのデータはおそらく期待値にきわめてよく合致するように補正されていたであろうと推測された。

多変量データを精査する為には、まず、それぞれの変数の分布を吟味すること、すなわちまずヒストグラムを描く、つづいて、最初の4つのモーメントと、歪度、尖度を求め、極端に大きい (または小さい) 値があった場

合には、それらを取り除いて、ふたたび歪度、尖度を計算し、その変化を調べるとよい。

主成分⁸⁾、または2変数のクロス表を図示することは、データの誤りやはずれ値を発見するうえで、大変役に立つことがある(訳者注4)。アフリカで収集された骨格データに関するオリジナルなデータを詳しく分析したところ、その多くが解析に耐えられないことが判明したので、それらのデータはすべて廃棄した。そのかわりに、発掘場所から送付され、ダックワース実験室で計測された小標本のデータのみを解析の対象にすることにした⁹⁾。

直面したもうひとつの問題点は、データの不完全さであった。頭蓋骨のほとんどは破壊されており、どの頭蓋骨についても予定したすべての計測値を得ることはできなかった。たとえば、頭蓋容量は、頭蓋骨がまったく損われていない場合のみ測定できる。データが不完全な場合、どのようにして平均値や分散共分散を推定できるであろうか。最近、最尤推定と尤度比基準を用いた欠測値推定の方法がある雑誌に発表されたが、この研究に、欠測値がランダムに現われるという仮定が採用されていることを知って、私は大変びっくりした。欠測値がランダムに生ずるということは、現実のデータではまず起こり得ない。よく保存された頭蓋骨は、一部が破壊されているものに比べると、すべての計測値が全般に小さな値をもつ。したがって、よく保存された頭蓋骨は、頭蓋骨の真の母集団を代表する標本とは到底いいがたくなる。この意味で、すべての既存データを用いて尤度を計算することは、誤った結果を導くもとになる。したがって、頭蓋骨のオリジナルな母集団を特徴づけるパラメーターの推定値を得るには、適切な補正が必要となってくるわけである。

墓地の中で、頭蓋骨の保存の仕方が異っているという現象は私にとって大変興味深いものであった。そこで、私は、この現象について、統計学のある会合で講演することにした。その講演の中で、私は小さい頭蓋骨は、大きいものに比べるとよりよく保存される可能性が高くなることを、骨格データの一連の解析から立証しようとした¹⁰⁾。講演の終わりになって、座長がコメントを申し出た。彼が、パリの博物館で、旅行者と一緒にガイドの説明を聞いていた際、ガイドはあるケースの前にとまって「そこにあるよく保存されている頭蓋骨がシーザーのものです」と説明した。旅行者の中ですでに「ラオの発見」に精通していたある人が「シーザーの頭蓋骨はこのように小さいのだから、よく保存されているのは当然だ」と叫んだ。これを聞いたガイドはすぐにいいかえした。「本当をいえば、これはシーザーが16歳の時の頭蓋骨です……。」

次の問題は、比較すべきグループ間について計算された D^2 の値の利用法とその解釈である。その当時、クラ

図2. 二つの正準変数によるカーストのグループの布置

スター分析のアルゴリズムは開発されていなかったが、計算された D^2 の値を用いて、2次元の図に多数のグループのおおまかな相対的位置を示すことは可能であった。このような方法をより客観的にするために、私自身、正準変数¹¹⁾ (canonical variables) と命名した測定値の線形結合 $l_1, l_2, \dots$ を構成し、そのうちの最初の k 個 ($l_1, l_2, \dots, l_k$) がグループ間の差をもっともよく反映するような k 次元の部分空間を構成するようにした。このような線型結合は、級内分散行列に対する級間分散行列の固有ベクトルを用いて計算されるものである¹²⁾。図2は、最初の2つの正準変数を用いて構成された2次元的な位置づけで、図に示されている各グループの名称は、インドのある1つの州に属するカーストの名称である。最初の3つの正準変数、さらには、いくつかの正準変数から選択された3つの変数を用いて3次元的表示を行うことは可能である。しかし、高次の次元によって特徴づけられるデータを低次元の空間に布置する場合、何らかのゆがみが生じてくる。したがって、それぞれのグループ間の距離に関して生ずるゆがみの量を計算しておけば、その知識を低次の空間に表現された各グループの布置の解釈に役立てることができる¹³⁾。多次元尺度構成に関するトーガソン (Torgerson) の研究にならって、私は、1964年の論文¹⁴⁾で、 D^2 の値の表、またはそれぞれのグループ間の非類似度を用いて直接的に正準変数を計算する方法を発表した。今日、正準変数を用いたグラフィカルな表現がポピュラーになっているのは喜ばしいことであるが、それによって生ずる距離のゆがみについてあまり考慮が払われていないのは遺憾である。私が命名した CANCORD という正準変数の略称は、ニャナデシカン (Gnanadesikan) によって提唱された CRIMCORD という略称に比べると、より好ましいもので、罪を感じさせないであろう。

1950年以降、クラスター分析の各種の手法が発表され、そのアルゴリズムを記載した数冊の書物が出版された。しかし、私が初期の著作で言及したように、クラスターを形成するためのきまったルールというものとは存在しない。異った場面においては異った手法が用いられなければならない。1つのグループが他の2つのグループと連絡しながら、全体のグループが連結している場合を考えてみよう。このような場合に、すべてのグループをいくつかの異ったクラスターに強制的に分類することは、誤った解釈を導くものになる。私は、かつていくつかのグループがオーバーラップしているかもしれないすべてのクラスターをまずリストアップし、それぞれのクラスター内の各グループが与えられた数値よりも小さな非類似性係数をもつようにする方法を提唱したことがある¹⁵⁾。その一例を図3に示した。図3中の頂点が各グループの中心位置を示すもので、ある閾値より小さな値

図3. 最大部分グループのグラフ

をもつグループのみが連結されている。こうして作られた図は、クラスター分析の樹状図 (デンドログラム) より多くの情報を有しているといえよう。図3の Mmy のような点は、たとえある非類似性係数の水準によって他

のクラスターに分類されるとしても、通常のデンドログラムには反映されにくいものである。しかし、図3で見出された Mmy と他のグループの関連は、歴史的事実にも符合する。

先に触れたフィッシャーの書「統計的方法…」は私の学生時代、唯一の高度な統計学の書物であったが、学生の教科書として指定されるか一般向けの書として推薦されるようなものではなかった。何故ならば、この書は、途中の証明がスキップされていたり、ある結果を導くための仮定が明確に記述されていないため、統計学の基礎を学ぼうと志す人々にとって、決して読みやすいものではなかったからである。「もしその本を以前に読んだことがないならば、その本を読むべきではない」というたぐいの話を私自身しばしば耳にしたことがある。しかし、私にとって、フィッシャーの本は大変刺激的なもので、全力投球して読むだけの価値があるものと強く確信していた。そこで、フィッシャーの本のすべてのページを読破し、その理論と実際の応用について討論する勉強サークルを結成した。そうしているうちに、ある日、フィッシャーの本に欠けている理論とその数学、さらにはひとまとまりのデータに関する解析法とその解釈等についての注釈 (commentary) を書きまとめることは、大変意義ある仕事になるという考えが頭にひらめいた。カルカッタとケンブリッジで従事した人体測定のデータ解析の仕事は、フィッシャーの方法を論評し、フィッシャーの本に含まれていない多変量解析の方法論を構築するための素晴らしい材料を提供してくれる筈である。さらにその当時、私は、多変量解析に関するインドの統計学者の仕事を広く普及させる機会を求めていたので、ケンブリッジにいる間に1冊の本を書き始めることにしたが、その本の標題を長い間決めることができなかった。その本は、数理統計学と呼ぶにふさわしいほどの統計的理論とその数学を具備していたので、数理統計学というタイトルであれば、より広い関心を集めることができた筈である。しかし、その頃、若気のいたりで、使命感に燃えていたので、結局のところ、統計学の実際の側面を強調して *Advanced Statistical Methods in Biometric Research* というタイトルをつけた。

ケンブリッジで Ph.D. を取得し、上記の書物の最初の草稿を携えてカルカッタに戻った私は、統計学的方法論に関しての価値ある研究を行うためには、生きたデータに接触することが不可欠であるという確信をより強く抱くようになった。

多変量解析の研究は、過去25年の間に2つの異った方向に進んでいる。1つの方向は、アンダーソン (Anderson)、ジェイムス (James)、フィライ (Pilai)、および他の研究者によるもので、行列方程式の根の対立仮説のもとでの厳密な分布を求め、その結果に基づいてさまざまな検定基準を導いたり、有意性検定のためのパーセント点の数表を計算している。多変量解析においては、いくつかの仮説を同時に検定したり、いくつかの母数を同時に推定する場合が少くないが、このような場合における推測の方法が、正規性の仮定のもとでクリシュナイ (Krishnaiah) やその他の人々によって研究されている。

もうひとつの方向は、多変量正規性や確率的考察の枠を離れたもので、ケンドール (Kendall)、クラスカルと他の研究者によって、考古学的素材の時間的順序づけ、歴史的またはその他の文脈のもとでの進化論的樹木構造 (evolutionary tree structure) の構築、および多次元尺度構成等が行なわれている。Cavalli-Sforza と Nei は、ある種の遺伝学的モデルのもとで、生物進化に関する樹木構造を推定している。

多変量解析の研究は、今日においても発展の途上にある。多くの問題に対して、いまだにその解答が見出されていない。実際的な問題に多変量解析を適用するにあたって、十分な解答が与えられていない問題も山積みしている。私はいま、かつてフィネイ (Finney) がいった「多変量解析の研究は、その論文をかいた研究者の数学的才能のはけ口としてしか役立つにすぎないもので、実際の効用は伴わない」という言葉と、テューキー (Tukey) による「状況を理解しえない場合に多変量解析が利用されている」といった言葉を想い出す。このような批判の言葉に象徴される状況は、これからの研究者が実際の問題から生ずる問題の解決にもっと眼を向けていくように

すれば、きっと変っていくであろう。

コンピュータープログラムに対する無定見な使用についても言及しておこう。不幸なことに、現在の研究者には、自分が当面する研究課題の解決を既成のコンピュータープログラムに委せきってしまう傾向がみられる。ここで、一例として、アフリカで発見された歯の化石から、アウストラロピテクス・アフリカヌスが人類と類人猿のどちらに近いかを定める問題¹⁶⁾を考察しよう。この問題に適したプログラムは判別分析のプログラムであるから、データを判別分析のプログラムにかけたところ類人猿よりは人類に近いという解答が与えられた。しかし、実際には、観察された化石は、人類にも類人猿にも適合していない。何故ならば、上記のプログラムには与えられた標本が、特定化されたグループのいずれにも属さないという可能性が全く考慮されていなかったわけである。化石の実際の位置でなく、判別関数の方向に正射影された位置がマハノビス D^2 の計算に用いられたために、人類への値が類人猿への D^2 値に比べて小さくなり、その結果、誤った結論が導かれたわけである。

いかなる研究テーマについても、きっちりとした仕立服のようなコンピュータープログラムというものは存在しない。しかし、研究テーマに必ずしもびったりしない既成のプログラムを用いて解析した結果といえども、研究の最終的な方向の決定に関しては有用な情報を与えてくれるであろう。

最近発展しつつあるグラフィカルな方法についても一言述べておこう。グラフィカルな方法の適用によって、多変量データからはずれ値を見つけ出し、データの背後に潜む構造や、統計的解析の結果についてより深い洞察を行うことができるようになる。その詳細については、最近出版されたエナデシカンによる書物（訳者注5）、さらには、チャーノフ（Chernoff）によって提案された多変量データに関する絵画的表現等を参照されたい。

— 完 —

付記：ページの関係で割愛したが、この他にウィルクスのラムダ基準の分解にもとづくパートレットの検定方式¹⁷⁾に関する批判が述べられている。

著者による注

- 1) この日は奇しくも、C. R. Rao の誕生日にあたる9月10日であった。
- 2) Rao, C. R. (1953): Discriminant function for genetic differentiation and selection, *Sankhya*, 12, 229-246.
- 3) Anthropometric survey of the United Provinces, A statistical study (with P. C. Mahalanobis and D. N. Majumdar), *Sankhya*, 9, 90-324 (1949).
- 4) *Sankhya*, 6, (1943), p. 348, (ここでの基準は、フィッシャーの固有方程式の根の和になる)。
- 5) Tests with discriminant functions in multivariate analysis. *Sankhya*, 7, 407-414 (1946).
- 6) Familial correlations or the multivariate generalization of the intraclass correlation, *Current Science*, 14, 66 (1945).
- 7) フィッシャーは4つの劣性遺伝子（波形の尾、虚弱、多毛、皮膚色）をもったマウスを産ませる課題を与え、私はそれを最小数の交配でなしとげることができた。
- 8) The use and interpretation of principal component analysis in applied research, *Sankhya*, 26A, 329-358 (1964).
- 9) *The Ancient Inhabitants of Jebel Moya* (jointly with R. K. Mukherjee and J. C. Trevor) Cambridge University Press, 1955.
- 10) On a formula for the prediction of cranial capacity, (with D. C. Shaw), *Biometrics*, 4, 247-253

(1948).

- 11) Anthropometric Survey of United Provinces, A statistical study, (with P.C. Mahalanobis and D. N. Majumdar), *Sankhya*, 9, 90-324 (1949).
- 12) The utilization of multiple measurements in problems of biological classification, *J. Roy. Statist. Soc.*, B10, 159-204 (1948).
- 13) Taxonomy in anthropology, *Mathematics in Archaeological and Historical Science*, 19-29, Edin. Univ. Press (1971).
- 14) The use and interpretation of principal components in applied research, *Sankhya*, A26, 329-358 (1954).
- 15) Bengal Anthropometric Survey, 1945. A statistical study (with D.N. Majumdar), *Sankhya*, 201-408 (1958).
- 16) *Linear Statistical Inference and its Applications*, Wiley, 1973. (奥野忠一他訳 統計的推測とその応用, 東京図書, 1977).
- 17) Discussion on the paper by M.S. Bartlett, Multivariate Analysis, *J. Roy. Statist. Soc.*, B9, 176-197 (1947).

訳者注

- (1) Bowker A. H. (1960): A representation of Hotelling's T^2 and Anderson's classification statistic, *Contribution to Probability and Statistics*(Essays in honour of Harold Hotelling), 142-149, Stanford University Press を参照せよ.
- (2) この著は, 知能の多因子説を発表した Thurstone L.L. によるものである.
- (3) 全ての群の共分散行列を等しいと仮定しそれを Σ , 群 j の母平均ベクトルとデータ数を κ_j , n_j とおくと,

$$V_r = \frac{1}{n} \sum_i \sum_j n_i n_j (\kappa_i - \kappa_j)' \Sigma^{-1} (\kappa_i - \kappa_j) = \text{tr}(\Sigma_A \Sigma^{-1}) = \lambda_1 + \lambda_2 + \cdots + \lambda_r$$

ただし $\phi(\lambda) = \det|\Sigma_A - \lambda \Sigma| = 0$ で, Σ_A は級間共分散行列

となる. なお, トレース基準とは, 下記の文献を参照のこと.

Hotelling, H. (1951): A generalized T -test and measure of multivariate dispersion, *In Proceedings of the Second Berkley Symposium on Mathematical Statistics and Probability*, University of California Press,

- (4) (5)に示すニャナデシカンの書の第6章に詳述されている.
- (5) *Methods for Statistical Data Analysis of Multivariate Observations*, John Wiley (1977) (丘本正, 磯貝蒙史訳 統計的多変量データ解析, 日科技連, 1979) である.