多変量解析を理解するために(1)

多変量解析の基本的な考え方

高木 廣文 | Takagi Hirofumi

東邦大学医学部看護学科国際保健看護学研究室教授

■1950年生まれ。74年東京大学医学部保健学科卒。79年同大大学院博士課程修了。同年米国国立衛生院奨励研究員として米国国立環境保健学研究所勤務。81年聖路加看護大学講師、82年同大助教授。89年文部省統計数理研究所助教授。99年新潟大学医療技術短期大学部看護学科教授、同年同大医学部保健学科教授。2006年4月より現職。著書に『ナースのための統計学 — データのとり方生かし方』(医学書院)、『多変量解析ハンドブック』(現代数学社)、『健康科学とコンピュータ』(共立出版)など。

統計学は、人間を対象として調査研究を行う科学分野では必須アイテムの1つである、と言って反対する人はいないと思う。しかし、統計学の本は沢山あるのだが、中身を見ると見た目が数学に似ているので、とにかく苦手な人が多いのではないだろうか。特に、多変量解析となると使いたいのだが、もしくは既に使っていても、間違って理解している場合も多く見られるようである。

パソコン用の統計解析ソフトの普及は、当初から指摘されていたように、方法を理解しないでも使えるために、誤用や結果の解釈の誤りなどをそこら中で引き起こしているようである。このような現状の改善に少しでも役立つように、本シリーズでは主要な多変量解析の手法について、本当にやさしい解説を試みたい。

1. 多変量解析とは何をする方法なのか?

何か調査を行った場合、我々は1人の対象者から様々な情報を得られるように、調査計画を立てるのが普通だろう。例えば、対象者の基本情報である性と年齢は常に調べるし、職業や学歴などもできれば収集したい情報の1つである。このように各対象から得られる多数の特性に関する情報は、「**多変量データ** multivariate data」と呼ばれている。

まずデータ解析の初めに行う基本的手順として、年齢や血圧などの平均値や標準偏差を求めたりする手法は「**単変量解析** uni-variate analysis」と呼ばれることがある。また、多くの変数間には大なり小なり関係があるため、我々は、相関係数を求めたりクロス表を作成したりと、2変数の関係を解析することもある。しかし、我々人間の持つ特性は極めて多様であり、それぞれの関係は極めて複雑である。従って、より多くの変数を一度に解析する必要があるのではないかと考えられるだろう。

このように、ある現象を解明するために、3変数以上の変数から成る多変量データを、

それぞれの変数相互の関係を考慮して、目的に応じて分析する統計学的な解析方法は「多変量解析 multivariate analysis」と呼ばれている。

一概に多変量解析と言っても、分析方法によって解析目的は、それぞれかなり異なっている。とは言え、多変量解析の各手法の主な目的は、以下のように4つに分けて考えることができる。すなわち、

①ある現象をより簡潔に記述したり情報を圧縮したりする。
②ある現象の背後にあると仮定される潜在因子を探索する。
③ある現象に対する他の変数の影響を検討する。
④ケース（個体）の所属するグループを判別したり予測したりする。

と言える。

以上のように、多変量解析とは、それぞれの解析の目的に応じて考案された多くの方法の総称である。以下に、それぞれの目的による多変量解析の考え方と主要な方法について、その概略をまず説明することにしよう。

2. 情報の圧縮とは何か

大学入試では、幾つかの科目の試験が受験生に課される。ある大学の入学試験では、英語、数学、世界史、物理学、現代国語の5科目であるとしよう。このような場合、大学の試験担当者は、入学試験の結果をどのように評価するのだろうか。最も簡単な方法は、

合計点＝英語＋数学＋世界史
＋物理学＋現代国語

のように、「合計点」を計算するのではないだろうか。

合計点を求めるという操作は、あまりにも普通に行っているので、合計点とはどのような意味を持っているのかを考えたことはあるだろうか。元々5科目それぞれの得点を眺めて、どの受験者を入学させるかを決めるのは、やってみれば分かるが大変なことである。教授会のように大勢の人間が決める場合には、英語を重視したり、数学を重視したりとなかなか合格者を選べないのではないだろうか。

単純に、多数の科目の得点を1つにまとめたものが「合計点」であるが、これは5科目の情報を1つにまとめたものである。このように、複数の科目の得点から1つの合計点にまとめることは、多くの情報をまとめて（すなわち圧縮して）、より理解しやすくしていることになる。このように、合計点とは多くの科目の得点では判断しにくい場合に、より意志決定をしやすいように、情報をまとめるという意味があることが分かるだろう。

ところで、合計点の計算をするのに、各科目の得点を単純に足しただけでよいのだろうか。例えば、

合計点＝2×英語＋数学＋0.5×世界史
＋物理学＋2×現代国語

ではいけないのだろうか。もしくは、

合計点＝英語＋2×数学＋世界史
＋3×物理学＋現代国語

ではどうだろうか。

このように、それぞれの科目の得点に、ある数値を掛けて合計点を計算することもあるだろう。各科目に掛ける数値（重み）をどのようにすればよいかは、どのような受験生を

大学に入学させたいかによるだろう。理学部ならば数学や物理学の能力の高い人を選びたいと考えるかも知れない。一方、人間科学部ならば、英語や現代国語の能力の高い人を選びたいと考えるかも知れない。結局のところ、どのような目的でその試験が行われているのかということで、それぞれの科目に掛ける重みも変わるのが当たり前である。

3. 分析目的と合計点の求め方の違い

大学入試は、入学後に優秀な成績を修めて無事に卒業してくれるような受験生を選抜したいと考えて実施するものではないだろうか。そうすると、入学後の成績の平均点と入学試験の関係が大きくなるように、各科目に掛ける重みを決めればよいのではないかと考えることができるだろう。

英語や数学などの試験科目が5科目ある場合、それらを科目1、科目2、...、科目5のように書くと、重み付けをした合計点の計算は、

合成得点＝重み1×科目1＋重み2×科目2
＋･･･＋重み5×科目5

のように書くことができるだろう。ただし、上の式の途中の「･･･」は、科目3と科目4でも同様に重みを掛けることを省略して示したものである。また、合計点という用語ではなく、各得点を単純に足し算しただけではないことを示すために、「合成得点」という言葉を使うことにした。

実際の計算では、重み1や重み2のところには、1.5や0.5などの数値が入り、科目1や科目2には、60や80などの得点が入り、それぞれ掛け合わされて、合成得点が求められることになる。

実際に各科目に掛ける重みを決めるには、

①合成得点と入学後の成績の平均点の差を最小にするようにする。

②合成得点と入学後の成績の平均点との相関を最大にするようにする。

のような基準で重みを決めればよいのではないかと考えられるだろう。実際にやってみると、この2つの基準で解析すると同じ結果を得ることができる。この2つの基準で解析を行う方法は、「**重回帰分析** multiple regression analysis」と呼ばれている。

それでは、入学試験の目的が、大学生活に対する適応や不適応を事前に予測する目的を持っている場合には、どのように科目に掛ける重みを決めればよいだろうか。

このような目的を達成するための考え方として、大学生活や入学した学部への適応グループでは重み付けした合成得点の値はプラスになるようにし、一方、不適応グループでは合成得点がマイナスになるようにできればよいのではないだろうか。そのように計算された合成得点を使用することにより、入学試験の成績から適応者と不適応者を簡単に分けることができるようになるだろう。

とにかく、2つのグループについて、どちらのグループに所属するかで、その得点に大きな差が付くようにすればよいのである。このような考え方は結局、「**2つのグループでの平均得点の距離を最大にする**」ように各科目の重みを決めればよいということである。そのような基準の下で解析を行う方法は、「**判別分析** discriminant analysis」と呼ばれる多変量解析の方法である。

4. 多変量解析での変数の分類

重回帰分析や判別分析の計算式で使用する変数は、大きく２つに分類できる。入学後の成績や大学への適応・不適応のように、合成得点との関係で計算式の重み付けの基準となる変数は、

・**基準変数** criterion variable
・**従属変数** dependent variable
・**内生変数** endogenous variable
・**外的基準** external criterion

などと呼ばれている。

変数の呼び方が数種類もあるのは、多変量解析の各手法が多くの分野で発展してきたためである。それぞれの学問分野でその呼び方に違いがあるので注意が必要である。

入学試験科目の英語、数学、物理学、世界史、現代国語のように、計算式の右側に来て、重み付けした合成得点を求めるために使われる変数は、

・**説明変数** explanatory variable
・**独立変数** independent variable
・**外生変数** exogenous variable
・**内的基準** internal criterion

などと呼ばれている。

今後の多変量解析の説明では、「基準変数」と「説明変数」という言葉を主に用いることにする。従って、他の多変量解析の解説書で異なる用語を用いている場合でも、同一の意味で用いていることが多いので、この点を思い出していただきたい。

5. 基準変数がない場合の多変量解析

調査や研究によっては、重回帰分析や判別分析とは異なり、特別な基準変数がない場合もある。例えば、もっと科目を増やして、英語、数学、現代国語、漢文、古文、物理学、化学、世界史、日本史の試験成績で何が分かるのかを考えてみよう。

各科目の学力試験は、その科目に関係する能力や知識を測ることを目的として行われるのである。よく考えてみると、本当に試験によりその科目の能力が測れるのだろうかという疑問がないわけではないだろう。このように、学力試験によりその能力を測定するということは、暗黙に学力に関する因子が潜在的に存在し、それを試験成績によって測定できるということを仮定しているのである。これは、よく使われている知能検査なども同様である。

このように多くの科目の試験成績を基にして、それぞれの学力に関する潜在的な能力因子を探索することができれば、試験の得点を分析することで、「理科的能力」や「文科的能力」を測定することが可能になるかも知れない。このような解析では、基準となる変数はなく、用いた変数間の関係から潜在的な因子などが調べられることになる。

そのような目的のための分析方法には、「**因子分析** factor analysis」や「**主成分分析** pricipal component analysis」と呼ばれる方法がある。

今回は、多変量解析の概略を説明したが、次回以降は各分析方法の解説とともに、必要とされる基礎的な統計学の知識についても解説していく予定である。