

関数データ解析によって高まる ネットワーク社会の効率と安全性

ユーザー訪問 / 北海道大学 情報基盤センター 先端データ科学研究室

さまざまな問題解決に役立つ 関数データ解析

統計解析というと、理化学実験や臨床試験の結果判定、疫学など、主に専門的な分野で用いられるとのイメージが強いかもしれない。しかし最近では、心理学・行動科学などの社会科学から、経営、製造、マーケティングといったビジネス領域まで、その応用範囲が急速に拡大しつつある。「それに対応するには、関数データ解析という新しい概念を統計学のなかに取り入れ、統計解析が扱うデータの範囲を広くすることが重要です」と解説するのは、北海道大学情報基盤センターの水田正弘教授だ。

関数データ解析は、1990年にカナダのマギル大学のジェームズ・O・ラムゼイ博士が提唱した新しい統計学で、従来の統計解析が数値をデータとして扱うのに対して、関数をデータとする手法だ。これを大学受験の模擬試験を例に説明すると、従来の統計解析では合格率の判定を、何回か行っ

た模擬試験の点数で行ってきた。それに対して、試験の点数というのは受験者の体調や学習計画などによって変動する「関数」であると考えるのが関数データ解析だ。この関数をデータとすることによって、統計解析を合否の確率だけでなく「どのように勉強して試験に臨んだら合格できるか」といった問題解決の手段とすることができるのだ。

水田教授らは、国内で最初に関数データ解析を導入し、さまざまな応用研究に取り組んできた。例えば最近の研究テーマには、企業などがインターネットで発信するウェブページ(ホームページ)の有用性を左右する要因の分析がある。

これは研究室のウェブページのなかに、学生による札幌近郊の飲食店の紹介など、不特定多数の人が閲覧するようなページをいくつか設定し、そのアクセス記録から、閲覧行動を分析するというもの。従来のように、情報ごとのアクセス頻度が判定されるだけでなく、関数データ解析を応用することによって、ドメインの種類ごとの閲覧行動などさまざまな分析も可能だ。

これまでに得られた結果から、閲覧者は「新しい情報に敏感な群」「食べ放題や大盛りなど食事の量を重視する群」など、クラスター分析によっていくつかのタイプに分けられることが判明した。また、対応分析法を用いることで、飲食店情報の地域によって、ヤフーなど一般的な検索エンジンを経由してくるユーザー層と、検索エンジンによらずに直接アクセスしてくるユーザー層の存在がうかがわれることや、各検索エンジンと地域



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 正弘教授(左) 南 弘征助教授(右)

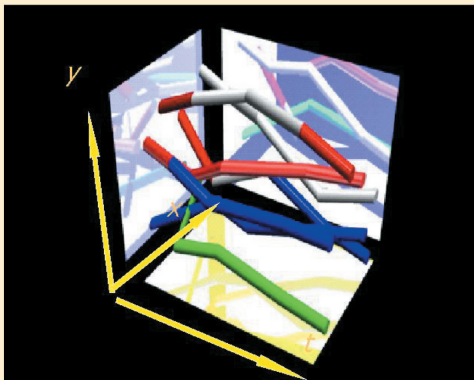
との関係などが分かってきたという。「ウェブページには、膨大なアクセス記録が残されますが、インターネットは匿名性が高く、これらにはあまり意味がないと考えられてきました。しかし関数データ解析をはじめとする統計解析法を使うことで、ウェブページの情報発信メディアとしての効率性を高める戦略を立てることも可能なのです」と実際に解析を行った南弘征助教授は語る。

豊富な解析メニューと 高速の情報処理が必要

まさに、これまで捨てられてきた情報の山のなかから、宝を見つけ出す技術の究極が、関数データ解析といえるだろう。その利用のための統計パッケージには、どのような能力が必要なのだろうか。「関数データ解析は、従来の統計解析と比べて入力するデータ量が膨大なものになります。豊富な解析メニューと高速の情報処理機能の両者を兼ね併せたものが必要ですね。関数データ解析を提唱するラムゼイ博士は、そうしたパッケージのひとつであるS-PLUSで実行可能なプログラムを発表しています」(水田教授)

水田教授らは、いま関数データ解析を用いることで、ネットワークへの不正アクセスの検知やスパムメール(迷惑メール)を効率的に除去する手法などを開発中である。新しい統計分析は、インターネット社会の情報流通の効率と安全性を高めることにも大きく貢献しそうだ。

関数クラスター解析概念図



関数データの分類は、刻々と変化する。分類結果も、時間など(z軸)によって変動するx, yの数値や色など多次元関数だ。