内容解説資料

じっきょう
数学資料 No. 75

データサイエンスと数学

滋賀大学データサイエンス学部長　竹村彰通

本年４月に日本初のデータサイエンス学部が滋賀大学に開設された。定員は１学年 100 名である。私はこの学部の開設まで２年以上準備に関わり，学部のカリキュラムや育成する人材像について検討を重ねて来た。この４月に無事新入生を迎えることができ，実際の教育が開始されたが，全く新しい学部であり，これらの新入生が狙い通りの人材に育ってくれるか，そして社会に巣立ってくれるか，不安な点も多い。データサイエンスはすぐれて文理融合的な分野である。そのような文理融合の観点から，データサイエンスに必要とされる数学について述べてみたい。また私は統計に関する知識や活用力を評価する全国統一試験である「統計検定」にも 2011 年の開始以来関わり，その中で統計と数学の関係について考えてきた。本稿では統計検定の過去問も題材として取り上げる。

まず，データサイエンスとは何かということについて，滋賀大データサイエンス学部の考え方を述べよう。よく言われることであるが，最近はビッグデータの時代である。そしてビッグデータから有用な価値を引きだすための学問がデータサイエンスである。その技術的基礎はデータを処理するための情報学及びデータを分析するための統計学であり，この部分は理系的である。しかし以下で説明するように，その応用先は文系的な分野が多く，この意味でデータサイエンスは文理融合的である。ビッグデータ時代の到来のきっかけとなった一つの象徴的な現象は，スマートフォンの普及であり，人々が情報をやりとりする手段が根本的に変化してしまった。スマートフォン普及のきっかけとなった iPhone がアメリカで発売されたのはわずか 10 年前の 2007 年である。日本では１年遅れて 2008 年に発売された。その後，今から数年前からは，地下鉄の中でも電波が届くようになった。電車の中では 10 年前までは多くの人が新聞や文庫本を読んでいたが，今ではスマートフォンを使って情報のやりとりをしている。新聞や本には，重要な情報が整理されて効率的に情報を得ることができるという利点があるが，読者は情報を一方的に受け取る受動的な立場である。これに対して，スマートフォンを使った情報交換では，ユーザはメッセージを送ったり検索をおこなったりといった能動的な行動もおこなっている。そのような双方向性がスマートフォンの便利さであり，もう以前の新聞や本の時代に戻ることはないであろう。

もくじ

このように人々は常時ネットワークに接続して双方向的な情報のやりとりをしている。そして人々の行動の履歴はネットワーク上に記録されていく。これらの詳細な情報の記録が可能となったのは，通信速度の向上とデータ保存のコスト低下によるところが大きい。以前では，保存することができずに捨てていたデータが長期間保存できるようになったことにより，行動履歴がそのままの形で記録できるようになった。例えば，個人の健康情報なども，生まれてからのデータが途切れることなく記録される時代になると思われる。ビッグデータの時代における顕著な変化は，以上のように人々の行動履歴が直接にデータとして得られるようになったことである。そしてネット通販における個人向けの商品推薦のように，このようなデータを活かした新たなサービスが次々と提供されている。人々が常時ネットワークに接続して情報のやりとりをしているのは，このようなサービスの便利さが理由の一つである。つまりデータサイエンスの応用としては，人や社会といった文系の応用が大きい。

もちろん人々の行動履歴以外にも，ビッグデータは文理を問わず気象，医療，遺伝，物流などさまざまな分野で得られるようになっている。これは計測技術やセンサーの発展のおかげである。最近では，例えば遺伝子検査の価格は個人でも手が届くようになってきている。そして，これらのさまざまなデータを組み合わせることも重要である。人々が遺伝情報を含め自分自身の健康に関する情報を豊富に得られるようになれば，健康のために消費行動を変えるかもしれない。個人向けの情報提供は，単に物やサービスをその人に売るためというビジネスだけでなく，さまざまな面で人々の生活を豊かにする可能性を持っており，データサイエンスは行政サービスの改善などにも貢献するものと考えられる。

データサイエンスは以上のように文理融合的であり，滋賀大学のデータサイエンス学部も文理融合をうたって学生を集めている。私のクラスで新入生にアンケートをしたところ，理系と文系の比は6:4であり，新入生については文理融合という目的は達成できたように思われる。ただし，日本の高校の受験では文理を区別する傾向が強く，滋賀大学データサイエンス学部の入試説明会でも，文系と理系のどちらが有利なのかというような質問が何回も出された。そのような質問は，高校の進路指導の先生から出されることも多かった。入学試験は競争試験であり受験の有利不利が関心事であることはもっともであるが，私としてはデータサイエンスという分野自体が文理融合的な人材を必要としていることをまず理解してほしいと思ったものである。

以上でデータサイエンスの性格とその応用分野について述べてきたが，ここからはやや具体的に数学との関わりについて述べる。すでに述べたように，データサイエンスの技術的な基礎は統計学と情報学であるが，さらにそれらの基礎は数学である。実際，滋賀大学のデータサイエンス学部では1年次で解析と線形代数を演習を含めて必修としている。また解析，線形代数，確率については高校との接続を念頭においた準備的な講義も提供している。それでは，具体的にどのような数学が必要とされるかについて主に私の専門である統計学の観点から論じる。

まず線形代数であるが，具体的な行列の記法や演算が統計学に必要なことは，表計算ソフトを考えれば明らかである。一度にメモリーに読み込めないほどのビッグデータは別であるが，普通のサイズのデータであれば通常は表計算ソフトに入力する。特に数値的なデータからなる表であれば，それ自体を行列と考えることができる。そして，データの平均や分散・共分散などの基本的な統計量の計算は，行列の演算に対応している。このように行列の記法で表せば，重回帰分析の最小二乗法の解なども簡潔に表示することができる。

データを行列で表す時には，行列のサイズは 2×2 のように固定されているわけではないことに注意しよう。例えば $n=50$ 人の生徒の5科目の試験の点数を表計算に入力する際には，各行を生徒に，列を科目に割り当てれば，行列は50行5

列となる。行列を簡潔な記法と考えれば，サイズの大きい行列こそ 1 文字（例えば X）で表すことの有用性がある。このようにデータサイエンスでは，まずベクトルや行列を，データを簡潔に表す記法として使い慣れることが重要である。

また，線形代数には 2 次元や 3 次元からの類推で，高次元の空間も幾何学的に考えられるというメリットがある。例えば n 組の x と y のペアのデータ $(x_1,\ y_1),\ \cdots,\ (x_n,\ y_n)$ の相関係数は，$\bar{x}$，$\bar{y}$ をそれぞれの平均値として，2 つの n 次元ベクトル $(x_1-\bar{x},\ \cdots,\ x_n-\bar{x})$ と $(y_1-\bar{y},\ \cdots,\ y_n-\bar{y})$ の間の角度のコサイン $(\cos)$ である。このような幾何的な見方をすると，相関係数の性質が理解しやすくなる。2 つのベクトルの角度が 90 度より小さい時は相関係数は正となり，90 度より大きい時は相関係数が負となる。ここで一つの例として，x，y，z の 3 変量を考え，x と y の相関係数 r_{xz} が正，y と z の相関係数 r_{yz} も正となる場合を考えよう。この場合，x が大きいと y も大きい傾向にあり，y が大きい時に z も大きい傾向がある。このように言葉で説明すると，3 段論法的に考えて，x が大きい時に z も大きい傾向にある，つまり x と z の間の相関係数 r_{xz} が正になる，と言えそうな気がする。しかしながら，x と z の相関係数 r_{xz} は負であることもあり得る。その理由は，x と y の間の角度が 90 度よりは小さいが 45 度よりは大きく，また y と z の間の角度も 90 度よりは小さいが 45 度よりは大きい場合，x と z の間の角度が 90 度より大きくなることは十分あり得るからである。このように幾何的に考えれば $r_{xy}>0$，$r_{yz}>0$ であっても $r_{xz}>0$ とは限らないことが納得できる。

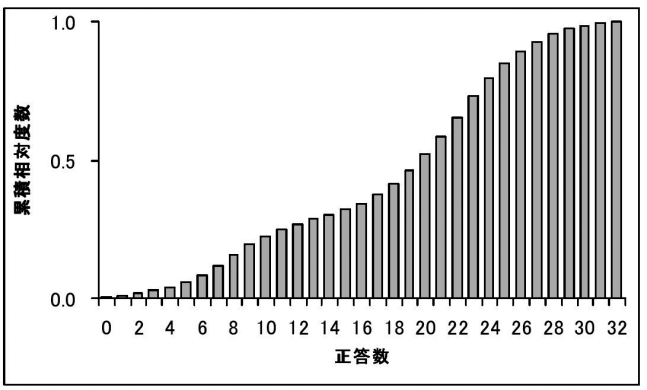
幾何的に考えると，高次元のデータを低次元の空間に射影するというような見方も非常に有用である。これはしばしば次元の圧縮とよばれる。原データから比較的少数の統計量を求めることは原データをそれらの統計量の次元の空間に射影していると考えることができるし，原データへモデルをあてはめることはデータの情報をモデルのパラメータに縮約していると見ることができる。特に線形の重回帰分析は低次元の線形部分空間への直交射影と見れば明確に理解することができる。

次に微分積分について考えてみる。微分や積分は多項式などの滑らかな関数について学ぶことが多いが，統計学では実は離散的な場合の差分や和分の形から慣れることが必要である。例えば，ある店の毎月の売上高のように，一定の時間ごとに観測される時系列データについては，原系列とともに前月との差をとって考えることが多い。このように差分をとることは時系列データを扱う際に基本的な処理であり，まずはそれらの操作に慣れることが必要である。

さらに，通常のヒストグラムにおいても，累積度数分布を扱うと差分と和分の関係が現れる。ここでは例として統計検定 3 級の 2012 年の問 10 をとりあげてみる（次ページ参照）。問題に与えられているのは累積相対度数分布であり，求められているのはこれにあうヒストグラムを選ぶことである。この問題を解くとき，微分と差分が意味的には同じ操作であること，また関数の傾きが微分であることの 2 つの事実を知っていれば，解答が非常に容易になる。いま問に与えられている累積相対度数分布を曲線と見て，横軸（正答数）にそって曲線の傾きを見ていくと正答数 8 あたりまで傾きが増加し，その後正答数 14 あたりまで傾きが減少し，その後また正答数 24 くらいまで増加していることが見てとれる。したがって答は④である。この問題はもちろん微分積分を習っていなくても解くことができる。このような具体的な問題で関数の傾き（増加率）を読み解くことができる能力は，微分積分の概念を理解する以前の能力である。教育上も微分積分を習う前にこのような例題を何回も解いていることが望ましいと思われる。

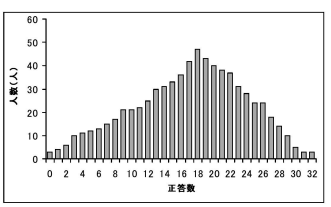
更に付け加えると，この問題では傾きの変化を見ているから，実は 2 階の微分を背後で考察しているのである。累積分布の正答数 8 は累積相対度数の曲線が「下に凸」から「上に凸」に変化する変曲点に対応している。つまり正答数 8 までは累積相対度数の増え方が増えており，8 以降 14 までは増え方が減っている。この問題のよう

問 10　ある学校で 32 問の数学のテストを行った。次の図は横軸に正答数を少ない順から 0 ～ 32 を 1 つ刻みでとり，縦軸にその数以下正答した生徒の割合を棒グラフで用いて表している。

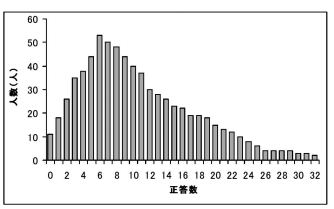


このデータを用いて，正答数のヒストグラムを描いたときに得られるグラフとして適切なものを，次の① ～ ⑤ のうちから一つ選べ。 10

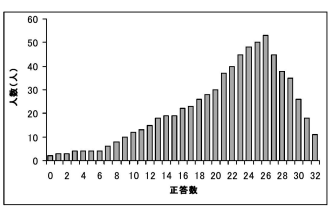
①



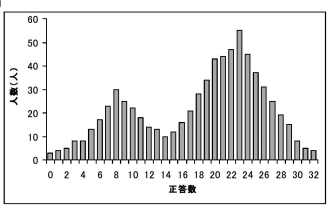
②



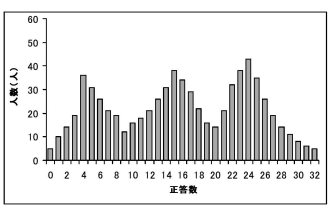
③



④



⑤



（統計検定 3 級　2012 年 11 月試験）

な簡単な問題でも，関数の凹凸の概念をも具体的に説明することができるのである。

さらに偏微分は多変数関数の最大化をおこなう際に基本的である。インターネット上のさまざまなサービスは，効果が最大となるように調整されている。具体的には，複数のパラメータを含む適切な評価関数を設定して，パラメータを動かして評価関数を最大化している。多変数関数を最大化するには，関数の勾配方向に進んでいけばよい。

以上，相関係数や累積相対度数の例を用いて，線形代数や微積分が統計学さらにはデータサイエンスに必要とされることを述べた。最後に，データサイエンスは文理融合的な分野であり，これらは必ずしも理系にのみ求められる数学ではないことを強調しておきたい。

（統計検定®は統計検定センターの登録商標です）