統計学は何を可能にするのか

上越教育大学大学院学校教育研究科・学校教育学系
准教授　奥村太一

1. はじめに

相関係数を考案したカール・ピアソンは，統計学を「科学の文法」と表現した。ここ最近になって,「ビッグデータ」や「データサイエンス」といった言葉を目にすることが多くなった。ICT の急速な普及により，SNS でのツイートや「いいね！」はもちろん，どのような言葉がインターネット上で検索されており，どこのコンビニでどのような商品がどのような客層に人気があるのか，膨大な量のデータがリアルタイムに蓄積されるようになっている。こうしたデータは規模こそ膨大であるものの，あらかじめ計画を立てて集めたものではないため，形式も不揃いで測定の質も粗いものである。そのため，これらは利用価値の低い粗大ごみのようなものとして放置されてきた。しかしながら，様々な分析を駆使することで，こうしたデータからマーケティングはもちろん，感染症の拡大予測から自動翻訳に至るまで様々な用途に利用できる情報を取り出せることがわかってきた。粗大ごみはビッグデータとして一躍宝の山となり，専門的なスキルを駆使してビッグデータを操るデータサイエンティストは時代の寵児としてもてはやされている。2017 年 4 月にはデータサイエンティストを専門的に養成する初の学部も設置され，一般の読者を対象とした統計学関連の啓発本も数多く出版されている。統計学が社会に変革を起こすための「手段」として脚光を浴びるようになったのは時代の必然とも言えるが，その本質を 100 年以上も前に見抜いていたピアソンの先見の明には敬服するばかりである。

2. 統計学の成り立ち

こうしたブームとは対照的に，統計学が実社会でどのように生かされているのか，具体的なイメージをつかめる人はそう多くないだろう。ここで，統計学の成り立ちを大雑把に把握するために，図 1 を用意した。統計学の理論的基盤はもちろん確率論である。ここに，平均や標準偏差を始めとする様々な統計量のふるまいについて議論をする数理統計学という分野があり，ここまでが統計学の根幹を成している。一方，統計学を道具として見た場合，得られるデータの特徴や扱おうとする現象によって様々な生かし方や制約が発生する。そのため，自然科学に限らず社会科学や人文科学においても，統計学は各分野に特化した発展をとげてきた。図 1 でいえば樹木の枝葉に相当する部分であり，しばしば〇〇統計学や計量〇〇学のように領域名を冠して呼ばれる。

一般に統計学者と呼ばれる人たちの多くは，もともと経済学や生物学，心理学などを専攻する学部や大学院において数理統計学のトレーニングを受けたのち，こうした諸領域での応用，すなわちデータの集め方と分析の方法論に関する研究に従事していることが多い。

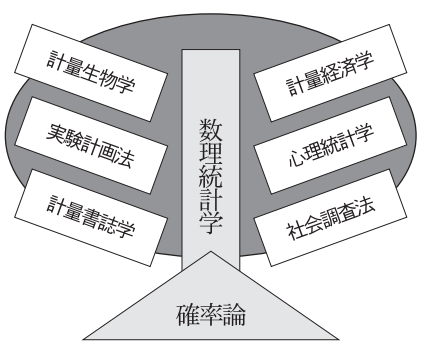

図 1　統計学のイメージ

分野に応じて得意とする手法が細分化されているとはいえ，もちろん共通する部分も少なくはない。また，統計学を学ぶなら誰しも避けては通れないルーツとも呼べる分析方法もある。ここでは，その代表として t 検定と回帰分析を取り上げる。教育的，実用的観点のいずれにおいても，これらの重要性はビッグデータ時代の今も色褪せる

ことはない。

3. 因果関係を検証する

何らかの働きかけ（処置）とそれに伴う反応について検証したいという場面は，様々な領域で生じる。古くは，品種改良や施肥によって穀物の一株あたりの収穫量が増えるかということが農業の近代化において重要な問いであった。現在では，どのような付加価値が顧客に商品を魅力的に感じさせるか分析することはマーケティング戦略において必須であるし，新たな医薬品に既存薬より高い治療効果が見込めるかといったことや，啓発活動によって慢性疾患のリスクを低減できるかといったことを検証することは，少子高齢化社会においては切実な問題である。

科学は，こうした因果関係に関する問いに答えるために実験と呼ばれる手法を洗練させてきた。今，ある処置 T が変数 X に及ぼす影響について検証したいとする。このとき，最も基本的な方法は，処置 T を受ける群（処置群）と受けない群（対照群）にそれぞれ n 個体を無作為に割り当て，X の平均を比較するというものである。

今，変数 X が独立に平均 μ，分散 σ^2 の正規分布に従うことを $X \sim N\left(\mu\,,\,\sigma^2\right)$ と表すとすると，標本平均 $\overline{X}$ の分布は

$$\overline{X} \sim N\left(\mu,\ \sigma^2/n\right)$$

となることが知られている。処置群と対象群の標本平均をそれぞれ $\overline{X}_1$，$\overline{X}_0$ とすると，もし処置 T が変数 X に何の効果も及ぼさなければ，差 $\overline{X}_1 - \overline{X}_0$ の分布は $\overline{X}_1 - \overline{X}_0 \sim N\left(0\,,\,2\sigma^2/n\right)$ となる。これを少し変形すると，

$$z = \frac{\overline{X}_1 - \overline{X}_0}{\sigma\sqrt{2/n}} \sim N\left(0\,,\,1\right)$$

と平均 0，分散 1 の正規分布（標準正規分布）に従う変数 z を得る。このとき，$|z| > 1.96$ となる確率は 5% にすぎないことが知られている。

ここまでの話を整理すると次のようになる。仮に，処置 T が変数 X に何の影響も及ぼさないと仮定しよう。これを帰無仮説と呼ぶ。すると，処置群と対照群に無作為に割り当てられた各 n 個体から X の平均を算出し，その差を既知の σ^2 を用いて z に変換した場合，その絶対値が 1.96 を超えることはめったにない。逆に言うと，もしこの手続きによって $|z| > 1.96$ なるデータが得られたとすると，処置 T が変数 X に何の影響も及ぼさないと考える事自体に無理があるということになる。この場合，帰無仮説は誤りであると考えて棄却し，処置 T は変数 X に影響を及ぼすと考えたほうが賢明である。このような手続きを統計的仮説検定という。帰無仮説が棄却された場合，処置群と対照群では変数 X の平均に統計的に有意な差があった，もしくは処置 T の変数 X に対する効果は統計的に有意であった，などと表現する。

さて，データから z を算出するには分散 σ^2 の値が必要であるが，当然のことながらこれは未知である。そこで，これをデータから計算された分散で代用する必要が出てくる。変数 X について処置群と対照群の標本分散をそれぞれ S_1^2，S_0^2 とし，

$$S_{pooled}^2 = \frac{n\left(S_1^2 + S_0^2\right)}{2n-2}$$

とすると，z の式で σ^2 の代わりにこれを用いたもの

$$t = \frac{\overline{X}_1 - \overline{X}_0}{S_{pooled}\sqrt{2/n}}$$

は，もし処置 T が変数 X に何の影響も及ぼさないのであれば，自由度 $2n-2$ の t 分布に従うことが知られている。t 分布は標準正規分布によく似ているが，それよりもいくぶん裾が広がった形をしている。自由度とはこの裾の広がり具合を決める働きをするもので，例えば $n = 20$ であれば $|t| > 2.02$ となる確率が 5% となる。すなわち，この条件でデータを集めた場合，もし $|t| > 2.02$ なる結果が得られれば，処置 T は変数 X に何の影響も及ぼさないという帰無仮説は棄却され，処置 T の変数 X に対する効果は統計的に有意であると報告することになる。

このように，処置群と対照群に無作為に個体を割り当てて結果を比較する手続きのことを，無作為化比較試験と呼ぶ。因果関係を検証する様々な方法の中でも，無作為化比較試験は最も単純でありながら最も強いエビデンス（科学的根拠）を与

える方法のひとつであるとされている。これは，割り当てが無作為に行われることで，2 群の個体が処置の有無を除いて等価であることが期待されるためである。例えば，サプリメントによるダイエット効果を検証するために，処置群（サプリメントを摂取する）に 20 代の女性を，対照群（サプリメントを摂取しない）に 30 代の男性を割り当てたとしよう。この場合，2 群は明らかに等価でない。おそらく 20 代の女性は 30 代の男性に比べてやせ願望が強い傾向にあるだろうし，そもそも男女では体格からして違いがある。そのため，一定のサプリメント摂取期間を経て処置群と対照群の平均体重に差が見出されたとしても，それをサプリメントの摂取によるものだと結論づけることはできない。処置の効果を同定するには，年齢や性別といった他の要因がそれと連動しないよう統制しておくことが必須である。

無作為化比較試験は，群への割り当てを無作為にすることで要因の統制を行うものである。もちろん，無作為割り当て以外の方法でも要因の統制を図ることはできる。例えば，処置群にも対照群にも 20 代の女性を割り当てることで，上記で指摘した問題は解決できる。しかし，同じ 20 代の女性といっても，都心で働くビジネスパーソンと地方都市に居住する専業主婦とでは痩せ願望の強さは違うかもしれない。もし，これに気づかないまま処置群には前者を，対照群には後者を割り当ててしまえば先ほどと同じ問題が生じることになる。こうした要因を事前にすべてリストアップして，それらがどれも等価になるように意図的な割り当てを行うことは現実的に不可能である。一方，割り当てが無作為であれば，年齢や性別に限らずあらゆる点において異なりうる個体がいずれの群にも等しく割り当てられる見込みがあるわけであるから，潜在的に存在するどのような要因についても群間で等価であることが期待できる。もちろん，期待できるだけであって必ず等価になるという保証があるわけではない。しかし，処置の効果がなかったとしたら変数 X の値がどれくらい不等価になり得るか確率的に議論することはできる。すなわち，どのような個体がどちらの群にたまたま割り当てられたのかという偶然生じうるレベルを超えて処置群の平均が対照群を上回っていれば，その処置には変数 X を高める効果があるといえる。これは，上で述べた統計的仮説検定のロジックそのものである。

ここまで取り上げたように，処置の有無においてのみ異なる 2 群に個体を無作為に割り当て，平均の差に関する仮説を t 分布にもとづき検定することを，「独立な 2 群の平均値差に関する t 検定」などと呼ぶ。t 分布を発見し，t 検定の理論的整備に貢献したのは，イギリスの統計学者ウィリアム・ゴセットである。彼はカール・ピアソンの教え子で，ギネスビールの醸造技師でもあった。彼の功績は，同じイギリスの統計学者ロナルド・フィッシャーによって実験計画法という理論体系として結実することになる。無作為化比較試験は，現在でもヒトを対象とした新薬の臨床試験（治験）など因果関係の検証に広く利用されている。

4. 相関関係を予測に生かす

話を因果から相関に移そう。変数 X と Y に相関関係があるということは，一方の値から他方の値を予測できるということである。今，変数 X から Y に関する予測を行うことを考えよう。このとき Y を目的変数，X を説明変数などと呼ぶ。目的変数 Y の予測値を $\hat{Y}$ と表すと，

$$\hat{Y} = \alpha + \beta X$$

という 1 次関数を考えるのが最も簡単である。これを Y から X への線形回帰といい，この式で表される直線を回帰直線という。回帰直線の切片 α は $X = 0$ のときの $\hat{Y}$ であり，傾き（回帰係数）β は X の増分 1 に対応する $\hat{Y}$ の増分に相当する。

予測値 $\hat{Y}$ と実測値 Y とのずれ $Y - \hat{Y}$ を残差という。これを e で表すと，Y は

$$Y = \alpha + \beta X + e$$

となる。残差 e は平均 0，分散 σ^2 の正規分布に従うと仮定されることが多い。

回帰直線の切片および傾きは，予測の精度が最大となるように決めるのが合理的である。そのた

め，残差の 2 乗和$\sum e^2 = \sum (Y - \alpha - \beta X)^2$を最小にするような$\alpha$と$\beta$を求めることを考える。これはこの式を$\alpha$と$\beta$について偏微分して 0 とおいたものを 2 元連立方程式として解くことに帰着する。この解をそれぞれ$\hat{\alpha}$，$\hat{\beta}$と書くことにすると，

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} \qquad \hat{\beta} = r\frac{S_Y}{S_X}$$

が得られる。ただし，$\overline{X}$と$\overline{Y}$は標本における平均，S_XとS_Yは標準偏差，rはXとYの相関係数を表す。このような切片と傾きの求め方を最小 2 乗法という。

回帰係数は，結局のところ相関係数に標準偏差の比という正の値をかけたものでしかない。しかし，相関係数と異なりここにはXとYのもとの単位が保存されている。そのため，単なる関係性の記述から，予測という具体的な行動を起こすための情報を与えてくれる。イギリスの統計学者フランシス・ゴルトンは，Yを子の身長，Xを親の身長（いずれも成人時，単位はインチ）としたとき，およそ$\hat{Y} = 20.5 + 0.67X$なる関係が見られたことを報告している。これにもとづくと，例えば親の身長が 72 インチ（約 183cm）と平均を大きく上回る場合，その子の身長は 68.7 インチ（約 175cm）とより平均に近い値となると予測される。ゴルトンは，この現象を「平均への回帰」と名づけた。これが回帰分析の始まりである。

実際の回帰分析では，例えば生徒の学力テストの得点を学習意欲や自己評価，家庭の社会経済的地位の高さから予測するといったように，複数の説明変数を用いることが一般的である。こうすることで，目的変数の個体差がどのような説明変数によってどう定まっているのか，複雑な現象の成り立ちを多様な観点から切り分けて理解することが可能となる。このとき，$\hat{Y} = \alpha + \sum \beta_k X_k$のように目的変数と説明変数に線形関係が成り立っていると仮定するものを重回帰分析と呼ぶ。PISA や TIMSS といった国際学力調査では，学力と合わせて生徒の学習に対する意識や家庭の経済状況といった情報もアンケートを通じて収集しており，こうした分析を適用することで政策提言に資する様々な知見が提供されている。

一方で，目的変数によってはこのような線形関係を仮定することが適切でない場合もある。心理学では，刺激に対してヒトが感じる主観的な強度（感覚量）は，実際の刺激の強度と比例関係にないことが知られている。例えば，Xを音の大きさ，Yをそれに対するヒトの感覚量とすると，$Y = \beta \log X$なる非線形関係が見られる。つまり，音の大きさを 1 デシベルずつ上げていっても，ヒトが感じる音の強さの変化は段々緩やかになってゆくということである。これをウェーバー・フェヒナーの法則と呼ぶ。これ以外にも，何らかの経験の有無や事象の生起確率を予測したい場合など，得点を足し合わせただけでは予測できない現象は少なからず存在する。このようなケースでは，得られたデータの特徴をグラフや散布図に表わすことでどのような関係性が成り立っているか吟味し，目的変数の種類に応じて最適な関数を選択する必要が生じてくる。これらはいずれも回帰分析を拡張したものであり，送られてきたメールのタイトルや本文の特徴からそれがスパムメールであるか判断したり，病気や事故の発症率やかかった医療費に関するデータから適切な保険料を算定したりと，社会のあらゆる場面で私たちの生活に欠かせない役割を果たしている。

5. おわりに

統計学がこれほどまでに活用されるようになったのは，計算機の性能が飛躍的に向上したためである。数十年前であれば大型計算機を用いて何日もかかった分析が，今では個人のパソコンで瞬時に実行できる。無償で利用できるソフトウェアやデータベースの整備も進み，統計学の門戸は万人に開かれている。是非，自分たちでデータを集めたり，身近にあるデータを使ったりして実際に分析してみてほしい。きっと新たな発見に心を躍らせるだろうし，同時にデータを集めさえすればエビデンスになるというわけではないということにも気づかされるだろう。