Internet world
ネット時代に生きる
櫻井 哲朗

第17回
「統計学」は世界を暴く

「本質」を浮かび上がらせる ビジネス世界でも不可欠に

前回の冒頭で、寒さも一段落してきた…と書き始めたのですが、まったくもって、そんなことはなかったです。東京では、2月8日、14日と降った雪は1994年に記録した23cmと同等またはそれ以上の積雪が観測されました。ちょうど20年ぶりの大雪となった、今年の雪。

それらのニュースを見ながら突然、記憶の扉が開き20年前の大雪の日のことがよみがえってきました。20年前の大雪となった1994年2月12日、その日はとある新作ゲームの体験イベントの日で友達何人かといっしょに出かけました。

吹雪の中、半ズボンで

その新作ゲームは超ビッグタイトルで、小学生だった私にとっては発売日が待ち遠しい、できることならば今すぐやりたいゲームでした。そんなゲームを実際にプレイすることができるこのイベントに行かないわけにはいきませんでした。横殴りに吹雪き白一色に染まった静かな道を歩く小学生の一団、目指すはまだ見ぬエルドラドこと新作ゲーム。

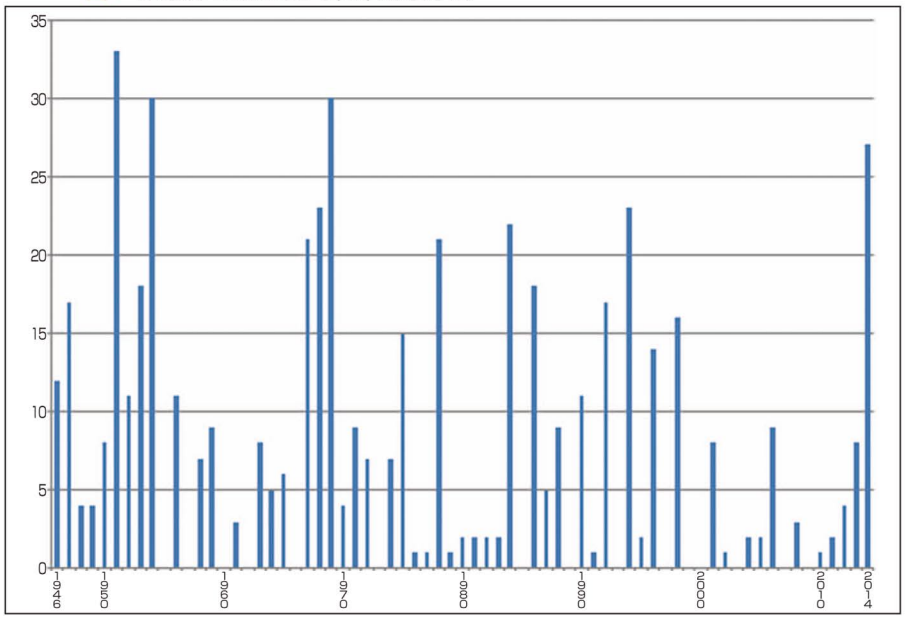
会場までの途中、半袖半ズボンの昭和の健康優良児スタイルだった私は道行く人から応援をもらいながら進んでいきました。そんなこんなで艱難辛苦を乗り越えて到着した会場。そこには我らと志を同じくした同士であり、難関で困難な壁を乗り越えてきた猛者たちが好敵手として集っていました。

いま思い出せば、新作ゲームの当時の人気から見ればかなり参加者が雪によって減っていたのだと思います。そのおかげで何度かゲームを体験することができました。そういえば、会場で先行販売をうたっていた攻略本が、帰り道に寄った地元の本屋さんで既に販売されていたのは衝撃的でした。子供心に商売ってすごいなと思いました。

データは印象づける

そんな20年ぶりとなった今年の大雪。実際には、どれぐらいの積雪だったか調べてみますと図1のようになりました。これを見ていただくとハッキリとわかりますが、20cm以上の積雪は20年ぶり、25cm以上の積雪になると1969年に記録した30cm以来の実に45年ぶりの大雪だったことがわかります。また、この1946年以降のデータの中でも4番目に大きいデータであることがわかります。

図1 東京都の最深積雪（cm）気象庁調べ



このようにデータで見ると、この前の雪がどれほどすごかったかが、よりわかりやすく伝えやすくなるかと思います。実際、筆者も過去のデータと比較することによって読者の皆さんに印象づけようと思い、このデータを持ち出しました。そこで今回はデータを扱う上で欠かすことができない統計学について考えていきたいと思います。

統計学の重要性

大学でもおざなり

今回使ったデータはとても簡単なデータで単純な使い方から素朴な結果を出しましたが、前回紹介したデータサイエンティストのように膨大なデータから高度な理論を駆使して有用な結果を出すことを求められる職業も最近では登場してきました。

いま現在は一部の人たちに限られていますが、このようにデータを活用する機会が増えてきました。今後、一部の人たちだけではなくより多くの人たちがデータを分析する機会が多くなってくることが予測されます。そうなってくるとデータを扱う上で有用な方法である統計学を扱う機会も同時に増えてきます。

統計学という学問を学ぶ機会は、そう多くはありません。統計学を学んだことがある人の多くは大学だったと思われます。統計をよく使う専門学科でない限り、短くて半期だけ、長くても前期と後期をあわせた1年ぐらいで終わってしまっているのがほとんどではないでしょうか。

そのため、多くの方は学んだには学んだがなんだか消化不良のまま受講を終えてしまっているかもしれません。そんな、みなさんにとっては大学で学んだ多くの科目の1つで印象も思い出もあまりなかった統計学、その統計学がいま必要となってきています。

MBAや高校生まで

実際、統計学がビジネスで使われるシーンは多く、筆者が前に在籍していたビジネススクールでも統計学の講座は開講されていましたし、そこに在籍していた学生さんの多くはMBA（Master of Business Administration：経営学修士）取得のためや実務のために統計学を駆使していました。また国家試験の1つである公認会計士試験においても2006年度から統計学が試験科目に加わったりしています。

さらに高校での必修科目である数学Ⅰに2014年度から導入された新学習指導要領から「データの分析」という内容が加わりました。このような統計的な内容が加わった背景には、これからのビッグデータ時代の到来が影響していると思われます。そのため、ビッグデータ時代に対応するために統計的な能力が必要であると考えられ、これからの子供たちは私たちが高校生のときには学ぶことがなかった統計的な知識を持って社会に出てくることになるでしょう。

使える能力が一般的に

IT革命と言われていた時代、高校に情報の科目が加わりコンピュータ・リテラシーと言われるコンピュータを操作して目的を達成する能力を持った高校生たちが社会に出てきたときと同じ状況が来ています。いまでは大人も子供も多くの人たちがコンピュータを使ってメールを送ったり調べ物をしたりしています。このように、これからは多くの人がコンピュータ・リテラシーに対して統計リテラシーとも呼ばれるような統計学を使える能力が一般的になるのかもしれません。

実際、SF小説の大家として知られているH・G・ウェルズは1

「統計学」は世界を暴く

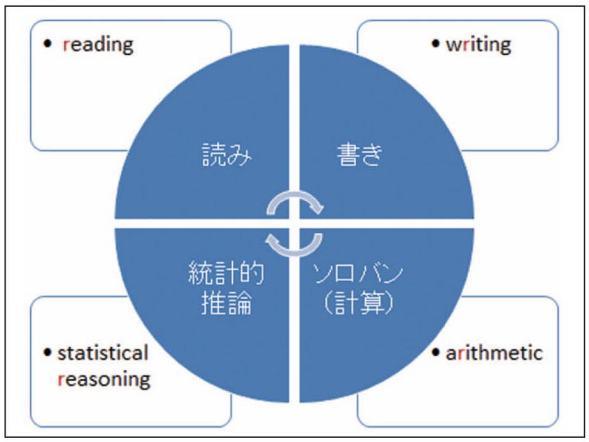
００年以上前の作品で統計的思考の重要性を訴えていました。ちなみにＳＦが詳しくない人のために説明すると、Ｈ・Ｇ・ウェルズは２００２年に映画化された「タイムマシン」や２００５年にスティーヴン・スピルバーグ監督によって再び映画化された「宇宙戦争」の原作者です。このようなことから図2のように、「やさしい統計入門」[1]では、これからは基礎学力につながる３Ｒである読み・書き・ソロバン（計算）に統計的推論を加えた４Ｒが重要になってきていると解説しています。

最強本との出会い

そんな時代の流れをよんだのか、２０１３年１月にある本が出版されました。ご存じの読者の方もいるかもしれませんが、その本は「統計学が最強の学問である」[2]です。ときどき略称で「最強本」ともいわれています。筆者が初めてこの本を見たときの衝撃は今でも忘れません。東京駅構内にある本屋さんで偶然この本を見つけました。

この本の著者の西内啓先生には失礼なのですが、「なんて中二病なタイトルの本なんだ。超カッコイイ」と思いました。誤解がないように書かせて頂きますが、内容は真面目な統計の本でとても興味深く統計学の本質をわかりやく解説しています。この本は統計関連の本では異例の25万部以上のベストセラーとなっているようです。そんな最強本にあこがれて筆者も今回の原稿タイトルに「目を引くタイムリー」なタイトルを考えてみました。ちなみに、今回のタイトルはロボットアニメならお任せのサンライズ制作の大人気アニメ「革命機ヴァルヴレイヴ」で使用されていたキャッチコピー「世界を暴く」をオマージュしてみました。

図2

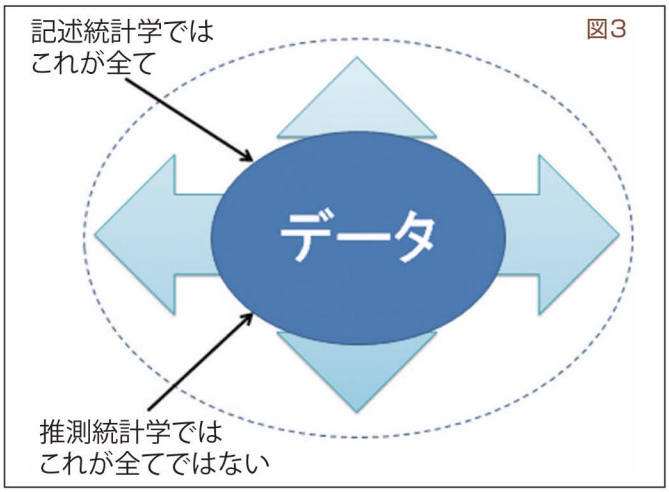


なぜ、このようなキャッチコピーを作ったかといいますと、十分な数のきちんとしたデータがあれば統計学を使って現象の解明を行うことができます。そこで現象＝解明＝暴露することで、コピーを作ってみました。そこで、ここからは統計学の考え方だったり使い方だったり言葉遣いだったりについてトピックスごとになるべく簡単に説明してみたいと思います。

統計学の二つの側面

そもそも統計学を使うとなにができるようになるのかについて説明したいと思います。データさえあれば統計学を使って全ての事柄に対して妥当な判断を下すことができます。だからこそ、参考文献[2]の最強本では「統計学が最強の学問である」といっています。統計学には二つの側面があります。それは記述統計学と推測統計学です。

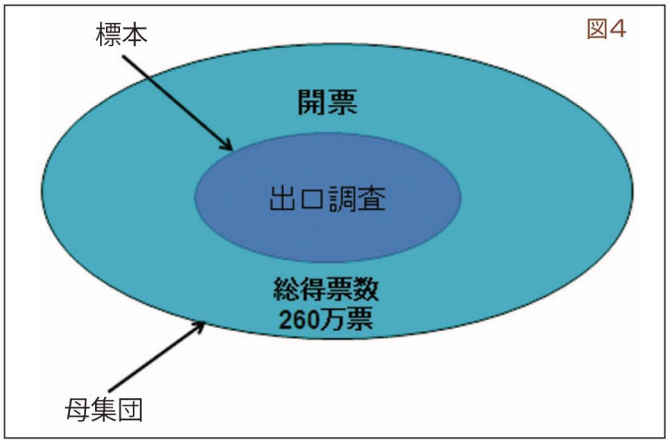
図3



記述統計学
全数調査で判断する

二つの違いは、データの違いにあります。図3のように記述統計学では得られたデータが全てであ

図4



り、これ以外にデータは存在しないとします。推測統計学では、逆に得られたデータは全てではなく、これ以外にもデータが存在し、得られたデータはごく一部のデータであると考えます。

これらの違いを選挙で説明することができます。図4のように総得票数260万票だった選挙があったときに当選者を決めるためには全てを開票、集計して確定します。ですが、テレビなどでは選挙速報として出口調査の結果から当選を予測しています。

記述統計学は全てを調べる全数調査のため判断に間違いがありません。その反面、全てを調べるための時間と費用、つまりコストがかかります。

推測統計学 効率いいが間違いも

推測統計学は一部分を調査し全体を予測するため判断に誤りを含みます。しかし、全調査に比べ調査にかかるコストを減らすことができます。このとき、推測統計学では得られたデータを標本といい、その背後になるデータをとってきた集団を母集団と呼びます。

さきほどの選挙の例で考えてみたいと思います。全てを開票するには時間と人件費などの費用がかかりますが、それに対して出口調査では全ての投票者に聞き取り調査をするわけではないので時間と費用のコストをおさえることができます。

とくに当選者と落選者の得票数の差が大きいときに、出口調査ではしばしば開票と同時に当選予測の発表を出すことができます。このように書くと出口調査の良さばかりクローズアップされますが実はそうとは限りません。出口調査の欠点として、どうしても間違いを取り除くことはできません。従って、どうしても間違いが許されない場合には全数調査をする必要があります。

さくらいてつろう

中央大学大学院理工学研究科を卒業し、専攻は統計学。コンピュータなどによって計測される大量のデータをまとめる多変量解析の研究。現在は、諏訪東京理科大学共通教育センター講師。東京都出身、30歳。

推測統計学の時代

このように書きますと、どんな場合でも間違いをしない全数調査をしたほうがいいのではないかと思います。ですが、現実的な問題として全てのデータを調べることが不可能な場合もあります。例えば、明日の天気を調べる場合に天気に関係のある無数の項目全てを調べ上げる必要があります。実際、天候はある場所での蝶々の羽ばたきが影響するともいわれるほど予測が困難なものの一つです。

このように微少な変化も後々の結果に大きな影響をおよぼすような状況下では、すべての項目を観測するというのは現実的には不可能です。そんなような状況でも使えるのが推測統計学になります。歴史的にも、記述統計学が19世紀から20世紀にかけて発展し、その後、より幅広い分野を扱えるようになった推測統計学が20世紀から発達していきました。また、これより統計学という学問は他の学問と比較すると若い学問であることもわかります。ここから先の統計学の説明は、推測統計学に関する説明を指します。

推測統計学の限界

では次に、どうして間違いを取り除くことができないのかについて説明させていただきます。まず思い出してほしいポイントとして、統計学つまり推測統計学は全部を調べないで一部分だけを調べるということです。全部と一部、このギャップを埋めるのが確率という理論です。そのため、統計的に求められたものは全て確率的に求められたものであり、そのため100%正しいという結論を得ることはできません。

このことを皮肉って、「統計に絶対は絶対ない」だとか「統計的に100%正しいことが一つある、それは統計学は間違える」といった言葉があるぐらいです。また統計学はデータがあってはじめて計算や判断を下すことができます。そのため、データが存在しないところでは統計学を使うことができません。

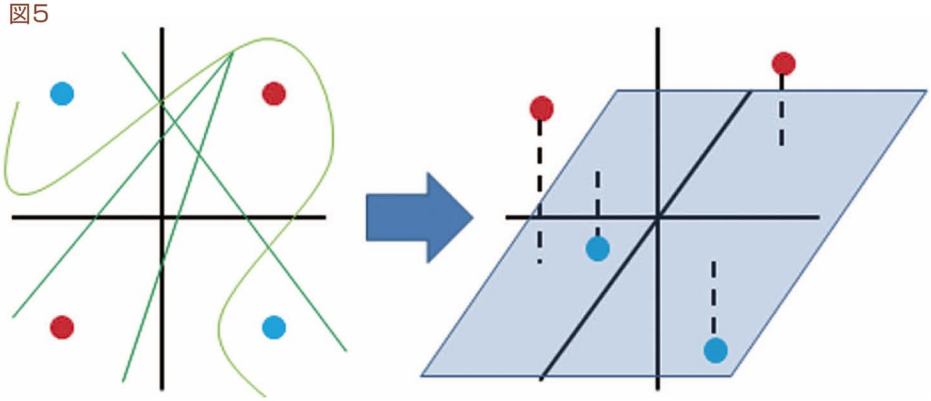
確率使う無作為抽出

また、統計学が妥当な判断を下すためのお約束または前提条件について説明させて頂きます。統計

学では、全体は調べることが難しかったりコストがかかってしまうため一部分のみを調べて全体の状況を知ろうとします。このとき、この選ばれた一部分が全体の状況をうまく反映していないと全体を知ることはできません。

先ほどの選挙の例で説明します。

図5



例えば、出口調査をする時に若い人だけを調査対象にするなど偏ったデータをとってしまうと、若い人以外の意見を知ることはできなくなってしまいます。では、この偏りなく全体をうまく反映するようにデータをとるための方法について考えてみましょう。

全体をうまく反映するためには全体を知っていないとそのようなデータを取り出すことはできません。しかし、いま全体を知るために調査をしているのですから全体の状況を知ることはできず、そのためそのデータを作り出すことは不可能です。つまり、全体を知るためのデータを作るためには全体を知る必要があるといった一種のパラドックスに陥ってしまいます。

そこで登場するのが無作為抽出という方法です。これはデータをでたらめにとってくるという方法です。そうすれば、全体を知らずとも全体の状況をうまく反映したデータを確率的に作り出すことができます。そのため、統計的な結論には確率を内包してしまうという側面もあります。実際、選挙における電話調査ではコンピュータがでたらめに選んだ電話番号にかけ調査することによって、この無作為抽出によるデータを作り出そうとしています。

確率を基準に判断

ここでは統計的な判断基準について説明させて頂きます。さきほども説明しましたが、統計的な結果には確率の要素を含んでいます。そのため、統計的な判断はその確率を割り出すことによって行われます。具体的には、仮説検定による母集団の判定方法があげられます。まず仮説検定とは、母集団つまり全体の状況を調べるために仮説を立てて検証する方法です。

このとき、立てられた仮説が正しいのか間違っているのかを、その仮説の下で得られたデータが出現する確率を計算することで検証します。簡単にいいますと、出現する確率が低いのであれば、その仮説は間違っているという判断を下します。

また、どのグループに属しているかが分かっているデータを使ってデータの判別規則を作る判別分析という手法があります。このとき、どのグループに属しているかが分からないデータを判別する方法の1つとしてその確率を用いて行います。具体的には、その出現したデータが発生したと仮定したもとでの各グループの確率を求めます。このとき、出現したデータは最も確率が高かったグループに判別されます。

このように統計学では確率を基準にして判断を下していきます。つまり、統計学では奇跡は起こらないものと考えます。そのため、「たとえ勝つ確率が1%以下だったとしてもオレ達は諦めない」といった熱い展開は起こりません。

最新の統計学 ビッグデータの効用

最後に最新の統計学について触れて終わりにしたいと思います。最近の統計学のトレンドは、ビッグデータまたは高次元データ解析が注目されています。これらのことが注目されるようになった背景にはコンピュータやインターネットの進化が影響しています。

いままでの統計学では、データを取るときにデータの個数を増やすことに注目してきました。そのときに調べる項目はそんなに数は

大きくないことが暗に仮定されていました。それは、ひとつに計算能力の問題があり、そんなに多い項目を含んだデータは扱うことができなかったためです。ですが、現在ではコンピュータやインターネットが発達したことにより多くの項目を含んだデータを扱うことができるようになりました。

調べる項目を増やす

このときの利点を図5のように表すことができます。例えば、図5の左のような平面において赤と青をわける直線を引くことを考えてみて下さい。先ほど出てきた判別分析がやっていることは、直感的に言えば、このような線を引くことです。

しかし、ちょうど赤と青を分ける直線は引くことができません。赤と青を分けるためには図にあるような曲がりくねった線でないとうまく分けることはできません。ですが、ここで1つ軸を加えた図5の右を見て下さい。このように赤が上側に青が下側に配置されているような状況なら図にあるような平面によって赤と青を分離することができます。このように調べる項目を増やすということは直感的には軸を増やすことにつながります。判別分析では軸が多くあればあるほど、うまく分離することができるようになります。

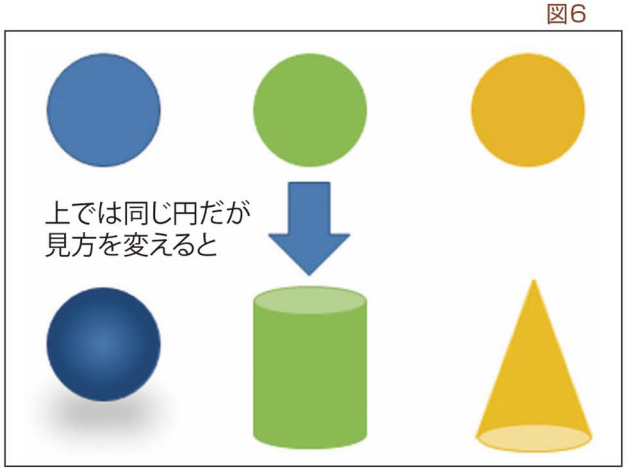
理論的な側面では、いままではデータの個数を増やすことに焦点を当てていたため、調べる項目数を増やすことはあまり考えられていませんでした。ところが、データの個数を増やしたときに導いた理論を改良する必要が出てくるようになりました。データの個数も調べる項目数も両方とも増やしたもとでの理論を構築するのです。求められた理論が実はより幅広い理論であることが一部の結果から得られています。

多面的に調べられる

また、データの個数が増えるだけでは備わっていなかった有効な性質が、項目数も両方とも増やしたもとでは備わるようになることがわかってきました。これは私たちの直感的な感覚と一致しています。なぜなら、データの個数を増やすということは全数調査に近づけていることであり、調べる項目を増やすということは色々な側面からそれを観測していることになります。

例えば、図6のような状況になります。上では同じ円だが、角度を変えれば球、円柱、円錐と異なる図形になります。このように調べる項目を増やすことでより多面的な側面から調べることができるようになります。

図6



データの個数も調べる項目数も両方とも増やすということは、より詳細な調査を行っていることを表してます。現在は、その詳細な調査からうまく情報を取り出す方法がいろいろと開発されています。

このように統計学も他の学問と同様に日々進歩しています。それらの結果は、論文や教科書または計算ソフトとして発表されています。最新の結果を知る機会は限られていますが、5年後10年後にはエクセルの表計算ソフトなどに組み込まれて誰でも使えるようになっているかもしれません。

前回、今回と2回に渡り、筆者の専門分野に関連する話題について解説させていただきました。いよいよ来月号で連載も最終回となります。どうぞ、最後までお付き合いいただければ幸いです。

参考文献

[1] 田栗 正章、藤越 康祝、柳井 晴夫、C.R.ラオ、やさしい統計入門、講談社、2007

[2] 西内 啓、統計学が最強の学問である、ダイヤモンド社、2013