統計学とは何か

私たちの周りには、数限りないデータがあります。データとは「何らかの目的のために取得されたまとまった数値や符号の集合体」ですが、それらの集合体を漠然と見ても、そこからは何も得ることはできません。データの数を数えたり、平均を出したり、傾向を見たり、分類をしたりと、何らかの手を加えることによって、初めてデータの性質や意味を知ることができ、活用することができるのです。

ある程度の数のデータには、必ずバラツキ（不確実性）が伴います。もし、ある学校のテストの点数が全員同じであったら、平均点や順位、偏差値を出すことに全く意味はありません。一年中天気や気温が一定であったとしたら、天気予報は要らないし、気温をグラフに描く必要もないのです。
しかし、実際には、学年やクラスによって点数は異なりますし、地域や日時によって天気も気温もばらつきます。それゆえ、クラス別の平均点や気温のグラフなどを描いて、クラスの特性を把握したり、明日の気温の予測をしたりします。

統計学とは、ある程度以上の数のバラツキのあるデータの性質を調べたり、大きなデータ（母集団）から一部を抜き取って、その抜き取ったデータ（標本）の性質を調べることで、元の大きなデータの性質を推測したりするための方法論を体系化したものです。

統計学の体系

統計学には、大きく分けて 2 種類あります。あるデータを集めて、表やグラフを作り、平均や傾向を見ることでデータの特徴を把握するという統計を「記述統計」といいます。一方、母集団からサンプルを抜き取って、そのサンプルの特性から母集団の特性を推測し、それが正しいかどうかを検定する統計を「推測統計」といいます。

※参考文献:『統計学とその応用』 田栗 正明，日本放送出版協会日経 BP 社，ISBN:4-595-30556-7

記述統計

標本データにバラツキがなければ、標本特性は 1 つの値を示せばすべてを表すことになるのですが、データには例外なくバラツキが存在するため、複数の集団の特徴を表すには様々な統計的指標が必要になります。最もよく使われるのが平均値です。平均値はバラツキのある集団の値を代表する値であり、「A 組の英語の平均点は 60 点、B 組の英語の平均点は 55 点だったから、A 組のほうが優秀だ」という使い方をします。

しかし、本当に代表値でクラス全体が優秀かどうかを判断してよいのでしょうか。例えば、A 組には極端に優秀な生徒が数人いて全員が 100 点を取っていた。しかし、この数人を除いた生徒の平均点は 53 点だったらどうでしょう。代表値がそのクラスの全体の特性を表していない可能性もあるということです。
こういう時に活躍するのが、点数のバラツキ（分布）を示すヒストグラムです。バラツキの様子を知ることで、より詳しくクラスの特徴を知ることができます。
クラスの特徴を知ろうと思ったら、英語だけではなく、国語や数学、理科の点数も知りたくなるでしょう。このように、ある集団の特性をより詳細に知ろうと思うと、非常に多くの項目についてのデータを集めなくてはならず、では数学と理科の点数には関係があるのだろうか、どういう生徒は英語ができるのだろうか、どのようにすれば平均点が上がるのだろうか。このような複雑な課題を解決する統計が、「記述統計」といえます。

推測統計

一方の「推測統計」は 1920 年代に生まれたため、記述統計よりはかなり歴史が浅いです。プリミティブな統計方法は、基本的に全数調査であり、母集団と標本という考え方はありませんでした。
調査対象が多くなると全数調査は物理的にも時間的にも難しいので、標本抽出（サンプリング）という考え方が出てきます。アンケートで代表性を確保するための「層化無作為二段抽出法」などの標本調査論や実験計画法などは、母集団から抜き出すサンプル数が少なくても、より正確に母集団特性を把握するためのデータ収集の方法論といえます。

選挙の出口調査というものがあります。これは開票前に開票結果を予測するためのもので、代表的推測統計です。どこの投票場で何人に対して出口調査を行なうかなどは、各新聞社や放送局のノウハウになっているようですが、標本調査論に基づく標本抽出が行なわれています。有権者数が約 1 億人、投票率が 50%だとすると、投票の母集団は 5000 万人。出口調査は 20 万人程度の有効回答数があるそうなので、20 万人で 5,000 万人の推測をすることになります。
選挙の場合は、開票は母集団の全数調査ですので、標本調査の正しさが、調査後 1 日も経てば完全に検証されてしまいます。しかし、多くの標本調査は、このような検証ができません。従って、標本調査で得られた結果が、本当に母集団の特性を表しているか、またどの程度の確率で正しく表しているかの検定をすることが、極めて重要な関心事になるのです。

テレビ視聴率がよく話題になります。調査対象世帯数は、関東地区・関西地区・名古屋地区で 600 世帯、それ以外の調査地区は 200 世帯です。先の出口調査と比較すると、かなり標本数が少ないと思うことでしょう。推測統計的には、600 サンプルで調査をした時のサンプリング誤差というものが、明確に定義されています。例えば視聴率が 10%だったとしましょう。この 10%には±2.4%の誤差があります。すなわち、母集団の視聴率は、95%の確率で 7.6%～12.4%の間に入っているということになります。これだけの誤差があるのですから、視聴率が 10%を切って 9%になってしまったという議論には意味がないことがわかります。統計学を知らない人は、そのような誤差について何も考えずに議論を進めてしまうことになり、極めて危険だと言わざるをえません。統計学において、この推測統計は非常に重要な位置を占め、近年発展してきました。しかしながら、ビックデータ時代を迎えこの推測統計の位置づけは大きく変容することになります。

ビッグデータの統計学

ビッグデータの登場で統計学が注目を集めています。理由は、統計学を駆使してビッグデータを分析することで、経営戦略やマーケティング戦略の立案、新商品・新サービスの開発などで大きな成果が得られることがわかってきたからです。勘や経験や度胸ではなく、データに基づく科学的な分析によって意思決定をすべきだということは、何十年も昔から誰もが分かっていたことでしょう。
にもかかわらず、歴史的には確固たる”学“としての体系を作ってこられなかったといわれ、日本の大学には統計学部が存在しません。統計学は地味だし統計で嘘をつくなどといういかがわしい印象があるとか、大学で統計学概論を勉強したが「ある集団とある集団に差があるかを知りたいのに、差がないという反対の仮説（帰無仮説）を立て、差がないことは滅多に起きないので差がないという仮説は棄却された」といった、非常に意味がわかりにくい日本語に接して、統計が嫌いになった人も多いことでしょう。

そもそも統計学がうさん臭いと思われ、“学”としての発展が遅れた背景には、「数学」との対比があります。統計と数学は似ているように思えるのですが、真逆の学問だといってもよいでしょう。
なぜなら、数学は公理があり定理があり確固たる解答がある場合がほとんどですので、演繹的論

理だといえます。一方、統計学はいくつかのバラツキのあるデータから母集団の本質を見抜こうという帰納的な推論であるため、このような人を煙に巻くようないかがわしいものを、学問としてみなすことはできないと思われていたのではないでしょうか。

歴史的に統計学が日の目を見始めたのは、イギリスのジョン・グラントやハレー彗星で有名なエドモンド・ハレーによる、人口の推測や死亡の規則性の発見だといわれ、その後確実な成果を上げてきました。そして、近年、不確実性の時代を迎え、急速な情報技術の進化があいまって、バラツキのある大量のデータ（ビッグデータ）を収集、分析し、意思決定に活かすことが、企業経営に必須だという考えが台頭し、統計学が一躍脚光を浴びたのです。

ビッグデータ時代の変化

ビッグデータ時代を迎え統計学はどのように変化してきたのでしょうか。先に述べたように、母集団特性は、母集団全体を調査できれば、標本抽出をする必要はありません。選挙は母集団全数の開票結果で決まるのですから、当選者を決定するという目的を達成するには、一部のサンプルを抽出し全体を推計する出口調査はなくても問題ありません。しかし、マーケティング課題を解決するための市場調査においては、国民全体に対して調査をしたり、その商品を購入したユーザー全員に調査を行なったりすることができなかったので、標本調査が行なわれてきました。ユーザーを性年代別にその特性を調べたり、購入状況や価値観質問によっていくつかのクラスターに分けたりし、市場全体を把握しようという努力がされてきたのです。

しかし、このタイプの市場調査には決定的な欠点がありました。例えば 1,000 人の調査をしてその母集団特性である市場が把握できたとします。その結果をもとに、商品開発を行なったり、プロモーション戦略を立案したりすることはできます。しかし、CRM（Customer Relationship Management の略で顧客関係管理）の要諦でもある One to one マーケティングを実現しようとすると、ほとんどを占める、抽出したサンプル以外のユーザーが、どんな特性かを個別に知ることができないのです。高度成長期のマスマーケティングの時代においては、よい商品を安く大量に生産し、テレビ宣伝をすれば売上は右肩上がりに上昇しました。しかし、ユーザーニーズが多様化し、市場をセグメントし、ターゲットを絞らなくてはモノが売れない時代に突入し、さらにインターネットの普及により生活者の購買行動が変化したことにより、企業のマーケティング戦略は大きくその方法論を変えなくてはならなくなったのです。

ビッグデータ時代を迎え、ID 付 POS や Web サイトの閲覧履歴、購買履歴が簡単に取得できるようになりました。マーケティング的な興味は、どのユーザーは何を欲しがっており、何を買いたいと思っているかを知ることです。ユーザーの嗜好を知る方法として、従来は性年代や居住地、可処分所得などの比較的変化の少ないハードな属性とも呼ばれるデモグラフィック特性、価値観やライフスタイルなどのサイコグラフィック特性が用いられましたが、これらのデータでは充分にユーザーの嗜好をとらえることはできず、一人ひとりが次に何を購入するかを予測することは、ほとんど不可能でした。

じつは、ユーザーが何を購入するかを知る上で、最も信頼できるデータは、その人の過去の行動履歴です。過去の来店記録や閲覧履歴、購入履歴などは、そのユーザーの嗜好を直接的に表現しているからです。ユーザー全員の行動データが取得できるようになったことで、広告の世界も劇的に変貌を遂げつつあります。誰にでも同じコンテンツを提示するマス広告の時代から、個別のユーザーの行動履歴を分析することで、同じコンテンツでも人によって興味を持ちそうな広告を個別に出し分ける「行動ターゲティング広告」や、広告主だけでなく提示する商品まで出し分ける「レコメンドバナー広告」が台頭してきました。真の One to one マーケティングが実現する時代になったのです。