

## 続々 やさしい統計のお話

(クォータリ-No.7 より)

前号と前々号の二回にわたって、統計の基礎的なことについて述べてきました。本号でこのシリーズも一先ず終わりへと致しますが、今回は現在も広く利用されている「統計的手法」の代表的な2, 3のものについて、先駆者たちがどのような背景のなかで、どんな考え方で開発したかを簡単に紹介したいと思います。何か勉強したいとするとき、先ず基本的な思想を知ることが大事で、そのためには、どのようにしてそれが生まれたかの歴史を探ることは意義のあることと思います。

また、記述統計学と推測統計学について少し説明し、最後に、統計的な考え方などについてまとめて見ることとします。

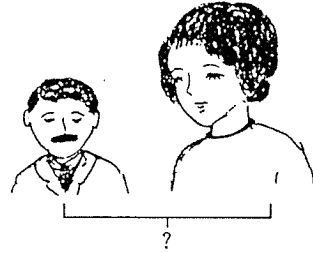
## 1. ガウスの誤差分布

誤差の概念や正規分布は、ドイツの偉大な数学者ガウス(C.F.Gauss 1777~1855)が発見し、理論づけをしたものです。彼は天文学、測地学、物理学にも数々の業績を残しております。当時地球の形状が問題となっていた時代で、天文台長のとき、ある地点間の距離を測量する必要があり、自らもその作業に従事しました。何回も測定すると、同じ数値が得られる筈なのに、少し違うデータが出てきました。恐らく始めは何かの間違いと考えたかも知れませんが、数多くのデータを解析し、研究の結果、測定には偶然的な変動がつきまとうものであること、それらの変動を全体として集約すると、ある規則性があり、つり鍾型の分布をすることなどを発見し、数学的に体系づけて発表しました。これがガウスの誤差分布で、今日の正規分布そのものです。

彼はまた、現在多くの技術者が利用している最小二乗法の理論を開発し、完成したことで著名です。(分析関係でも、検量線の数式化はこの最小二乗法によっています。)

## 2. ゴルトンの相関分析

データの処理で、二つ以上の変数間の関係を知りたい、という場合が多いのですが、そのために相関分析や回帰分析の手法が使用されます。相関分析はもともと遺伝学の分野で開発され、よく使用されたようです。生物遺伝学の研究に大きな足跡を残したイギリスの遺伝学者ゴルトン(F.Galton 1822~1911, 進化論で有名なダーウィンの従兄弟)が考え出した手法です。親が大きいと子も大きい、ということがよく言われますが、当然バラツキがあり、それらの関係を定量的に把握したいと考えたのが動機ようです。グラフのX軸に親の身長、Y軸に子の身長をとって多くのデータをプロットし、数学的に種々検討した結果、相関分



析なる手法を案出しました。親子の身長の関係では、明らかに相関が見られること、変動の中にも規則性があること、そして更に何代にもわたって調べて見ると、遺伝学的に面白い傾向があることも発見しました。それは二代三代と続くと、身長はグラフ上の平均身長の方角に帰って来る、ということです。つまり親が大きい(あるいは小さい)と、その子供は大きい(あるいは小さい)という傾向がある、しかし何代も続くと、どんどん大きくなる(あるいは小さくなる)ということではなく、平均の方角に戻って来る、これを回帰する、と言いました。XとYの関係式を回帰式と呼びますが、これはゴルトンが以上のような関係から名付けたものです。

## 3. フィッシャーの実験計画法

今日の推測統計学の礎石を築いた人はイギリスの生物学者フィッシャー(R.A.Fisher 1890~1962)です。彼がロンドン近郊の農事試験場に勤務していた時、農事の試験、研究について、最小の費用と時間で最大の意義のある情報を得て有効な結論を導くための実験方法の研究を続け、その成果をまとめた「研究者のための統計的方法」を1925年に発表しました。当時としては、極めて革新的な考え方と手法として、大変注目され、多大の評価を受けました。この文献は、現在でも推測統計学を深く勉強しようという人達にとっては、その後発表された彼の2, 3の文献も含めて必読のものとされています。

農事研究の実験が多くの困難を伴うことは、われわれ素人にも容易に想像出来ます。新品種開発のための実験でも、結果がでるまでには長い時間がかかるでしょうし、天候、日当たり、水はけなど人間が制御することが困難な条件も多いと思います。何とかして客観的に信頼出来る結果を、しかも効率的に得たい、という切実な必要性があったに違いありません。それまでの古典的な考えによれば、何も農事に限りませんが、実験はなるべく均質化されたところで、一つの要因について二つか三つの異なった水準で、それらの影響を比較して見よう、とする態度であったわけです。(今でもそのように考えている人がいます。)フィッシャーはその考え方を打ち破り、現実的な面から、ブロック化、確率的な誤差や変動、ランダム化など実験の効率化、精度向上のための具体的な実験方法を理論づけて提案し、今日各方面で活用されている実験計画法の骨組みを築き、少数のデータに基づく推測統計学の提唱者となりました。その後、多くの学者や実務家が、彼の思想を受け継ぎ、サンプルの観察による母集団の性質の推定や検定などの問題を次々と展開して来たのです。

#### 4. 記述統計学と推測統計学

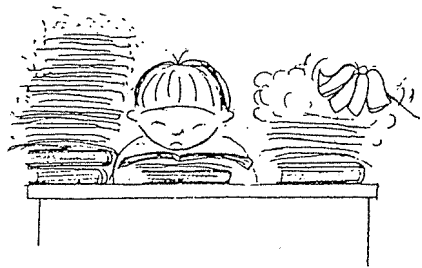
統計の起源が、主として国の行政のための資料の作成にあったことは、前々号でも触れました。国状学とか政治算術と言う言葉があるくらいです。人口、年齢構成、収入、財政状態その他多くの状態を対象として大量のデータを処理する術で、「記述統計」と呼ばれています。これに対し、前項に述べましたように、20世紀の始めのフィッシャーの研究と提唱に端を発して、少数データによる推定、検定などの数多くの手法が開発されて来ました。これらの統計を「推測統計」と言いますが、その結果、われわれ工業分野特に現場での技術活動にも極めて有益かつ実際の、数多くの統計的な考え方、手法が提唱されて来ました。アメリカのウェスタンエレクトリック社の電気技師 W.A. シューハートが案出して、1920年代に発表した「工業製品の統計的品質管理図法」もその一つです。

ところで、記述統計は古典的であり、推測統計は近代的で、より役に立つものだと単純に考えないで下さい。現在では、国、地方、あるいは業界などのレベルで数多くの各種の統計がとられて発表されており、いろいろの面に大いに役立っています。

一つの例として、ある人達には少しショッキングな統計資料を紹介しましょう。日本全体で見ますと、生

まれる人の男女の比率は、女100人に対し男は105人位とのこと、ここ100年近くの間、この数値は103~106人の範囲に入っているそうです。ただし、ひのえうまの年とその前後の年は若干違うとのこと、その理由は皆さんのご想像にまかせます。この数字は世界中どこでも大体同じだそうで、何故そうなるのか、神のおぼしめしという外はありません。独身男性にとって、早めに手を打っておいた方が良いですよ、という一つの警告かも知れません。

大量データの処理の問題は、われわれの身近にもあります。現場には、例えば始めから目的ははっきりしないが、とにかく記録しておこう、あるいは記録することが目的だ、という場合があります。その結果、何時の間にか作業日誌やデータ表がうず高く積み、ほこりをかぶっている、という情景も珍しくありません。しかし貴重な情報を含んでいるかもしれない大事なデータベースです。何とか整理解析したいと考えて、統計家に相談しても、なんだかんだと言って、簡単にはやってくれません。実はそのやりかたも難しく、大変な作業を伴います。記録の取り方も再検討しなければなりません。何れにせよ、データの背景を良く知っている実務家がやるしかないでしょう。コンピューターに頼るのが極めて効率的ですが、ごく最近このような大量データの処理にも役立つものとしてパソコンのソフトが市販されているようです。(例えば ASCII社の LOTUS 1-2-3) これから勉強していかなければならない実務上の課題だと思います。



#### 5. むすび

最後にまとめとして、統計的な考え方について、筆者の私見もまじえて書き加えたいと思います。

もともと統計学なる学問分野が存在するかどうか、という議論が長年にわたってなされているようですが、われわれにとってはあまり興味のないことでありまして、要は統計が仕事の効率化にいかに関与するかです。しかし統計の場合でも単に「手法」だけ使えば良い、ということでは極めて危険です。基本的な精神、統計

的なものを見かたを十分理解しておく必要があります。もっとも、始めから観念論を読んでもなかなかわかりにくいと思いますから、実際問題について自分で計算もして見て、ある程度手法に慣れた段階で、また基礎的な考え方に戻って見るのが良いでしょう。近頃は、統計処理の必要性を痛感していながら、基本的な考え方どころか、手法そのものもあまり勉強せず、市販の統計処理のソフトに頼ってパソコンの使い方さえわかれば良い、という若い技術者がいますが、困った風潮です。（コンピューターの利用を否定しているのではありません。今や統計処理には欠かすことのできないものです。）

統計の精神は「データを見て考える」ことです。いくつかのデータから、これは何を示しているのか、どう処置するかを考えるのは人間です。データを統計処理して、コンピューターに計算させ、有意差ありのマークがついたからそれは採用しない、という結論を出して得意になっている技術者がいます。つまり考えるということまで統計的手法とコンピューターにやらせているわけで、恐ろしいことです。例として、異常値の問題をとり上げて見ます。一連のデータのなかに、1つか2つの飛び離れた値があったとします。統計手法の一つである異常値の棄却検定をして、有意差あり、と出たので捨ててしまう、生はんかにこの手法を知っている人がよくやる手です。捨てる前に、サンプリングや測定の方法に問題は無かったか、あるいは対象とする母集団は一つでは無く、二つ以上の山が混じっているのではないかなど、調べなければならないことが

ある筈です。あるいは有効な情報が含まれているかも知れません。一般に、統計手法の「検定」には、よほど気をつけなければならない場合が多いようです。正しいサンプリングや測定をしたとしても、検定の結果には必ず危険率何%と但し書きが付きまゝ。問題の性質によっては、大きなリスクを犯すことになりかねません。

また統計的なものの考え方には、いくつかの特徴があります。要約して見ますと、具体的であり現実的であること、あるがままの姿をとらえて数字で表わすこと、問題としている対象は常に変動、誤差を伴うものであり、集団として見るとき規則性あるいは傾向があるということ、などに集約されるようです。

たとえ話でわかり易く説明しましょう。「樹を見て森を見ず」という格言があります。統計的な考え方とは、「樹を見て森を見よう」という態度です。何本かのサンプルとしての樹木を調べて、その結果から森という母集団を対象にして、広さ、樹木の構成、成長度などなどの情報をできるだけ的確に判断しよう、ということ。サンプルの個々の樹木にはあまりこだわりません。対象は総て具体的であり、数字で表わせること、情緒的文学的な表現は苦手です。そして変動、誤差に注目して、森という集団の規則性を追及しよう、ということでもあります。

長々と駄文を書いて来ました。どうかこれを機会に統計に興味を持たれて、一層勉強され、日常の仕事の効率化に役立てて頂ければ幸いです。