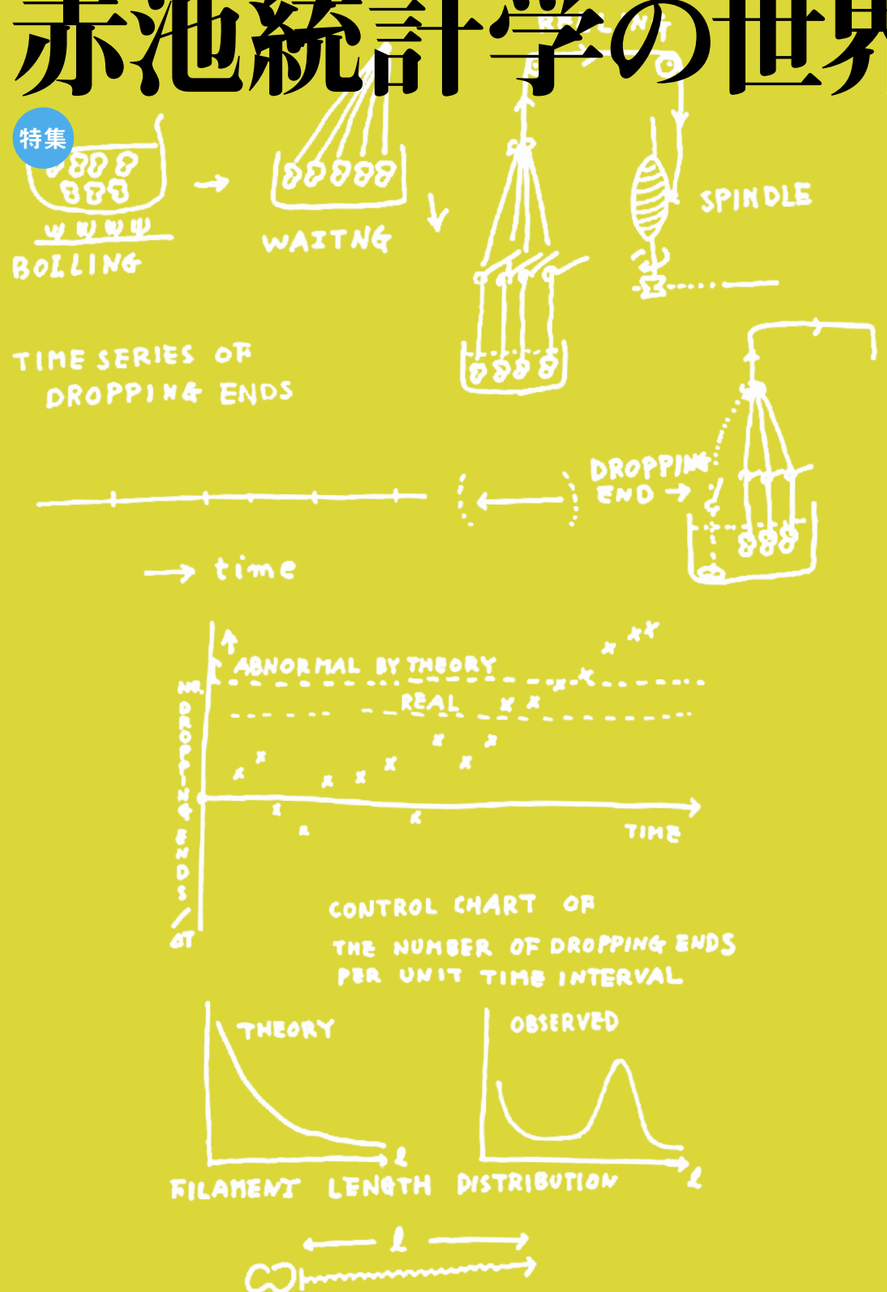
統計科学は今日、データを扱うあらゆる研究分野に浸透し、科学的研究のための方法論を提供している。多様なデータと知識を結びつけてモデルを構成し、モデルにもとづいてデータを生み出したシステムを理解し、予測や新たな知識発見を行う。「モデリング」と「予測」。この重要な概念を統計学の中心に据えたのが赤池博士（総研大名誉教授）であった。以来30年余、大量データの生成に伴って科学研究のスタイルが変化しつつあるなかで、赤池統計学はさらに応用分野を広げている。2006年12月には、その功績に対して京都賞が贈られた。この特集では、Part1「赤池統計学の源流」で研究の足跡をたどり、Part2「赤池統計学の展開」では赤池博士の薫陶を受けた研究者たちの活動を紹介。Part3「赤池弘次博士に聞く」では、堀田凱樹・情報・システム研究機構長との対談を通して、赤池統計学の真髄に迫る。

赤池統計学の世界

特集



生糸操糸工程の管理図法を記した赤池博士のメモ（21ページ参照）

Part 1 赤池統計学の源流

実世界との接点が生み出したパラダイム転換

北川源四郎
総合研究大学院大学教授 統計科学専攻／情報・システム研究機構 統計数理研究所長

赤池情報量規準AIC、ベイズ型情報量規準ABICに代表されるように、赤池統計学は統計学にパラダイム転換をもたらした。数多くの研究は現実の問題を解決するという必要性から生まれた。その思想と研究の流れを追う。

記述統計学から推測統計学へ

過去のデータや経験にもとづく将来予測や意思決定。われわれが日々何気なく行っているこのような行為は、人類がその進化の過程で獲得してきた知的な情報処理機能である。統計学はこのような人類のきわめて知的な営みを定式化したものといえる。しかしながら、確定的世界観にもとづく知的な営みがニュートン力学によって数理的方法として確立したのとは対照的に、複雑で偶然を伴う実世界をデータにもとづき科学的に把握するための方法論の歴史は比較的浅く、確率的思考は遅れて世に出てきた。

ゴルトンが遺伝の研究から類似性の指標となる相関係数の概念を見いだし、K.ピアソンがあらゆる現象が科学の対象となりうることを主張して「科学の文法」を提唱したのは19世紀も末のことである。K.ピアソンたちは、観測データからさまざまな現象を分布としてとらえる記述統計学を確立した。これに対して、20世紀に入るとフィッシャーたちは、現象を表現するモデルを仮定し、厳密に設計された少数の実験データからモデルを得る推測統計学を進めた。この実験にもとづく科学的方法論の確立によって、生物、医学、薬学、経済、心理、調査、品質管理などの複雑な現象の解析や管理において著しい成果が得られてきた。こうして近年に至るまで、理論科学と実験科学が科学的方法論の双璧を成していたといえる。

情報量規準AICへの軌跡

20世紀後半になると、現実の問題が複雑化・多様化する中で、「真のモデル」の存在を前提とする、従来の統計的推論の枠組みはしだいに現実にそぐわないものとなってきた。1973年、赤池氏は将来のデータを予測する状況を想定し、もっとも良い予測値を与えるモデルを求めるための規準AIC（Akaike Information Criterion）を提案し、統計学の歴史に偉大な足跡を残すこととなった。

情報量規準へ至る道には3つのポイントがあった。まず、第一は「予測」の視点である。従来の統計推論が、自然科学の目的とする「真理の探究」に対応して、「真の」モデルの推定をめざしたのに対して、将来の予測のために「良い」モデルを求めることをめざしたのである。真のモデルをめざす立場と、予測のための良いモデルをめざす立場には大きな隔たりが存在する。真のモデルの推定をめざして得られたモデルが、予測のために良いモデルとはいえないのである。

第二は、予測の問題を「分布」としてとらえるという立場である。赤池氏は1968年には予測誤差分散の推定量としてFPE（最終予測誤差）規準を提案し、時系列モデルの次数選択の自動化に成功していた。しかし、予測誤差の大きさに拘るかぎり、時系列モデルの推定は実用化できても、一般の統計的モデルの評価規準は得られなかった。赤池氏は、予測の問題は「値」ではなく「分布」としてとらえるべきことに気づき、モデルの良さを予測分布の近さで評価することにした。

第三は、その分布の近さを測る尺度として、カルバック-ライブラー（K-L）情報量を用いたことである。K-L情報量はボルツマンのエントロピーとも密接に関連する。ただし、K-L情報量には真の分布とモデルの分布が必要であり、そのままでは統計的モデルの評価には利用できない。赤池氏は、K-L情報量（の本質的部分）をデータによって不偏推定したものが、

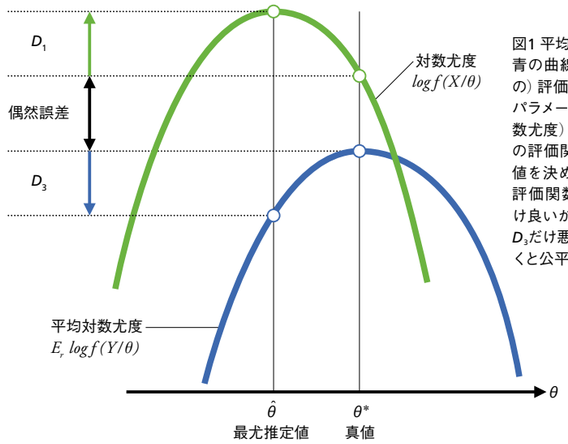
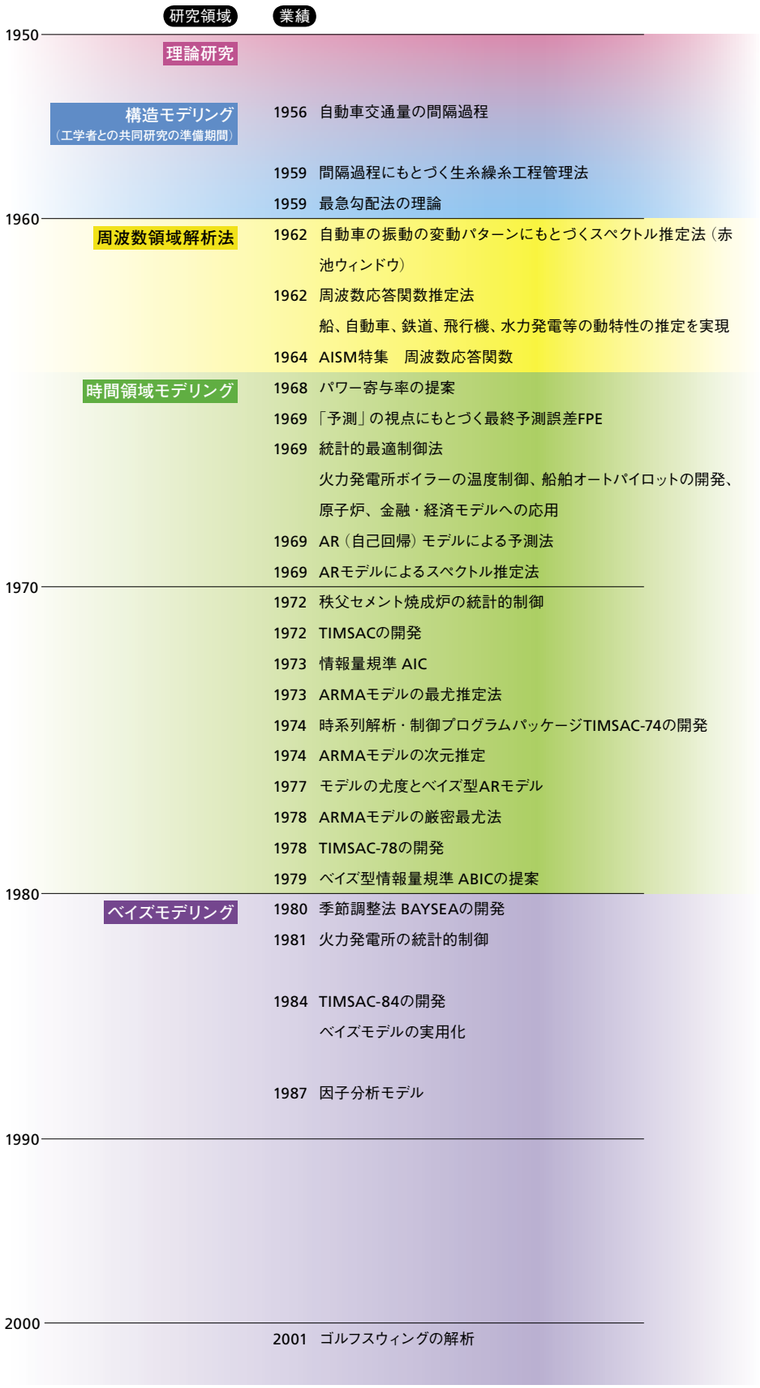


図1 平均対数尤度と対数尤度の関係
青の曲線（平均対数尤度）は（未知の）評価関数。その最大点が、最適なパラメーターを決める。緑の曲線（対数尤度）はデータから推定した見かけの評価関数。その最大点が最尤推定値を決める。最尤推定値は見かけの評価関数によれば、最適値よりD_1だけ良いが、本当の評価基準によればD_3だけ悪い。D_1+D_3を対数尤度から引くと公平な評価ができるようになる。

図2　赤池弘次博士の研究史



対数尤度になることに気がついた（図1）。この解釈によって、対数尤度最大化によりパラメーターを推定する最尤法は、実はK-L情報量の最適化をめざしていることを明らかにしただけでなく、原理的にはさまざまなモデルの良さを対数尤度の大小で比較できることがわかった。数理統計学の重要な概念であった尤度に関して、不思議にも従来この視点が欠落していたのである。

対数尤度がモデルの良さを表すとすれば、候補となるモデルが多数ある場合には、対数尤度最大のモデルを探せば、最適なモデルが決まると期待できる。だが、現実はそれほど簡単ではなかった。未知のパラメーターをデータから推定した場合には、対数尤度は正の偏りを持つ。その補正を行わないかぎり、公平なモデル比較はできない。赤池氏はこの偏りが、パラメーター数に比例することを見いだし、それを補正することによって情報量規準

$$\mathrm{AIC} = -2(\text{最大対数尤度}) + 2(\text{パラメーター数})$$

を導いた。

AICは統計的モデルの良さをデータにもとづき客観的に評価する。したがって、多項式の次数やフーリエ展開の項数のようにモデルが未知の「次数」を含む場合には、AICを最小にする次数を選ぶことによって、客観的に次数選択を実現できる。最高次数の係数が有意かどうかの検定を繰り返し適用する、従来の統計的方法に比べれば、格段に実用的になったことはいうまでもない。情報量規準AICの利用により、原理的にはすべての統計的モデルを同時に評価し、相互比較することが可能となる。

しかし、AICを便利なモデル選択基準と見なすのは適当ではない。AICの導入は、20世紀初頭以来の実験科学のための検証の統計学から知的情報処理のためのモデリングへと、統計的パラダイムの転換をもたらした。

AICの導出から明らかなように、情報量規準には「真のモデル」は不要であり、これがいくつかの重要な問題を提起した。第一に、われわれがなしうるモデル選択は相対的なものであり、常により良いモデルが存在する可能性が残されている。したがって、特定のモデル族の中で、最適なものを探すことにより、良いモデル族を提案することのほうがはるかに重要である。これはモデリングの重要性、科学研究における仮説提示の重要性を示している。

次に、いったん「真のモデル」の推定をめざす客観的な推論という立場を離れると、必然的により「良い」モデルを求めるという方向に進むことになる。従来の統計的推論においては、データにもとづく客観的推論をめざすことが主流であったが、いまや、観測されたデータだけでなく、対象に関する理論や知識、これまでの経験などのすべての情報を用いて「良いモデル」を構成することが肝要となった。情報量規準はそのような主観的に提示されたモデルに関しても客観的な評価を可能にした。情報量規準の提案は、科学研究におけるモデリングの重要性を明らかにし、それを実現する具体的方法を与えたことになる。

ベイズモデル実用化の先達に

社会の情報化が急速に進展し知識社会へ向かおうとする現在、情報技術の飛躍的進展によって、多くの科学研究分野や一般社会で大量のデータが時々刻々蓄積し、データベースが構築されつつある。このような情報化の波が、科学研究のあり方に影響を与えないはずはない。大規模データに基づく予測や情報抽出・知識発見が科学研究に不可欠の方法となり、理論科学、実験科学に続いてデータ科学が新しい科学的方法論として確立しようとしている。

問題はこの新しい科学の方法において中核となる技術である。赤池氏はAIC提案直後の1976年にはすでに、知的情報処理におけるベイズモデルの重要性を見抜き、その実用化の研究に着手した。それまでの統計的モデルでは、パラメーターの数を規定してきた。パラメーター数を増やすと、モデルの記述能力は向上するが、将来の予測能力は減少する。この問題に対して、パラメーターについても統計的モデル（事前分布という）を想定するのがベイズモデルである。ベイズ推論の方法は、その理論的優越性は認められながらも、哲学的論争、事前分布設定の困難、事後分布の計算困難性の問題から、実用化に至っていなかったのである。

1979年、赤池氏は経済時系列の季節調整に関連して、パラメーター数がデータ数の2倍以上という驚くべきモデルを提案した。言うまでもなく、従来の最小二乗法や最尤法では意味のある結果は得られない。赤池氏はペナルティ付き最小二乗法がベイズモデルから得られることを示して、ベイズ型情報量規準ABICによって事前分布を決める方法を提案し、ベイズモデルの実用化に大きな貢献をすることとなった。さらに、その後の計算機の高速化とモンテカルロ法に基づく統計計算法の急激な進展によって、計算困難性の問題も大きく緩和され、現在ではベイズモデリングは情報化時代に即した知的情報処理の主流としての地位を占めるようになっている。ここにおいても赤池氏の貢献は大きかった。四半世紀前にこのような知的情報処理の時代が到来することを予見し、3世紀にまたがる懸案であったベイズモデルの実用化を先導した慧眼には驚くばかりである。

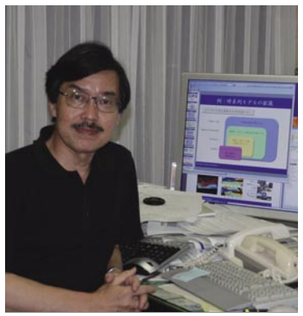
赤池統計学の原点は現場主義

情報量規準AICは、統計科学に限らずデータを扱うあらゆる研究分野で利用されてきた。実際、AICを提案した2つの論文の年ごとの被引用数は減少するどころか増加の一途をたどり、30年以上が経過した現在では年間1000件近くに及んでいる。一般に被引用数が少ない統計科学の論文としては驚異的な記録である。

このような偉業を成しとげた背景には、常に現実の問題を直視し、その解決に資する方法を開発しようとしてきた赤池氏の一貫した姿勢がある。1952年に東大数学科を卒業して統計数理研究所（統数研）の研究員となった赤池氏は、それからの10年近くをさまざまな分野の工学研究者との共同研究のシステム作りに費やしたといわれる。その交流のなかで、1960年代には、統計解析には個別の構造に立ち入ったモデリングが不可欠という、当初の考え方を封印し、線形定常モデルに基づく時系列解析に移行した。さらに、セメントの焼成炉のフィードバック解析を機に、1960年代後半には、周波数領域解析から時間領域モデリングへと転進し、ARモデルの実用化の要請から次数選択基準FPEとモデル評価規準AICを提案した。さらに、1980年前後には、新しい季節調整モデルの提案を機にベイズモデルの実用化に成功した。

このように、赤池氏の研究には何回かの大きな方向転換と飛躍的発展が見られるが、これらは現実の問題の解決の必要の中から生まれたものであった。しかも、それを単なる問題解決に止めず、常に統計的方法の発展につなげ、最終的には統計的パラダイムの転換にまで至ったのである。

赤池氏は常に、データを用いる現場の研究者にとって有用な方法の開発をめざしてきた。このような現場主義を離れては、これだけの偉業達成はあり得なかったのではないかと考えられる。



北川源四郎（きたがわ・げんしろう）
大学院では数学を専攻し統計数理研究所に就職したが、赤池さんの勧めで船舶の統計的制御の問題に挑戦したのを機会に時系列解析に転進。以後30数年、地震データ自動処理、経済時系列解析などを中心に統計的モデリングの研究を行ってきた。とくに、非定常・非線形時系列の解析のためのフィルタリングの方法とその応用に力を入れている。