情報科学講座 A·5·4

A. 共通基礎理論
A.5 統計理論
北川敏男編

情報量統計学

坂元慶行
石黒真木夫 共著
北川源四郎

共立出版株式会社

情報科学講座 A・5・4

A. 共通基礎理論
A.5 統計理論
北川敏男編

情報量統計学

坂元慶行
石黒真木夫 共著
北川源四郎

共立出版株式会社

著者紹介

坂元 慶行（さかもと よしゆき） 統計数理研究所付属
統計技術員養成所主事

石黒 真木夫（いしぐろ まきお） 統計数理研究所研究員

北川 源四郎（きたがわ げんしろう） 統計数理研究所研究員

情報科学講座 A・5・4

情報量統計学

検印廃止

定価 3500 円

昭和 58 年 1 月 15 日 初版 1 刷発行
昭和 58 年 3 月 25 日 初版 2 刷発行

編集代表者 北川敏男 喜安善市

発行者 南條正男
東京都文京区小日向 4 丁目 6 番 19 号

印刷者 大久保絢史
東京都新宿区市ヶ谷本村町 27 番地

NDC 417

発行所 東京都文京区小日向 4 丁目 6 番19号
電話 東京 947 局 2511 番 （代表）
郵便番号 112 振替 東京 1-57035番

共立出版株式会社

印刷・新日本印刷 製本・関山製本 Printed in Japan

社団法人
自然科学書協会
会員

ISBN 4-320-02171-1

“情報科学講座”の序

情報科学は機械・生体および人間社会における情報の生成・伝達・改造・蓄積・利用についての一般原理を攻究する基幹科学である．情報理論の建設，情報現象の解明，情報方式の開発の三つの領域にわたり，相互間の緊密な協力によってその発展をはかることは，現代科学技術の進歩にとって，必須の要請となっている．

21世紀の人類のビジョンは，情報革命のもとに築かれなければならない．20世紀前半における物理科学の進展は，やがて生物科学，人文科学，社会科学に大きな影響を及ぼし，これらの科学分野の飛躍的発展が期待されている．この時代において，それらの相互間の連結は，情報科学を通じて行なわれるべきものであり，情報科学の進歩は，これらの諸分野の発展に強力な推進力となるものと期待される．科学・技術の研究が，人類社会に占める役割は年とともに加重しているが，科学・技術の研究のために共通の基盤を提供するものは，計算機といい，ドキュメンテーションといい，情報科学の負担すべき任務に属する．

情報科学の組織的研究体制を整備するとともに，情報科学について系統的な学習を行ない，広範な教養を培い，わが国における情報科学の水準を世界のそれに遅れないようにすることは，今日の急務である．

このような趣旨から，情報科学講座（刊行当初全 63 巻）を刊行しようというのである．講座の意図するところは，情報科学の体系的集成であるから，理論・素子・組織・生体情報・装置の全分野にわたり，基礎的な解説から，第一線の研究の紹介にまで及ぶように努めた次第である．

わが国の情報科学の水準が世界をリードし卓越する日の来ることを待望しつつ，この目的のために，情報科学講座がいささかなりとも寄与できることを，心から念願するものである．

(1966 年 9 月記，1977 年 1 月一部変更)

編集委員一同

序

情報科学の生誕において，統計科学が一つの重要な開通路を用意したことは，N. Wiener のサイバネティックス，C. Shannon の情報理論において，これを見ることができる．生誕だけでなく，その後の発展においても，パターン認識，学習理論などにおける統計的決定関数，多変量解析などに見られるように，情報科学は近代統計理論の応用に負うものが少なくない．確率過程論と制御理論との関連についてはいうまでもない．本講座が A·5 統計理論を設けたのは，この関連を重視したからに外ならない．この部門では，すでに数巻を刊行したが，本書は，A·5·4 としてこの部門に属する最近の注目すべき成果の一つを新たに加えるものである．

本書は，基礎的な統計的解析の全般にわたって情報量を基礎にした接近を，入門から実例応用まで懇切に解説した，極めて特色ある著述である．情報量の概念を，統計学において導入しかつ重視したのは，外ならぬ推測統計学の創建者 R. A. Fisher その人であるが，そのエントロピーとの関連に着目してこれを統計的接近の諸問題に，体系的に適用する仕事は，わが国の統計数理研究所第5研究部長赤池弘次博士の提唱した AIC (Akaike Information Criterion) 理論の発展によるところが多い．博士が，この理論および応用を開発し，優秀な協力者を得て，種々の実際的応用において卓抜な実績をあげてこられたことは，国の内外において広く認められている通りである．

本書の3人の著者達は，このような研究活動に従事しつつある赤池グループの方々である．本書の刊行を通じて，この理論，方法の理解者，利用者さらには研究者の層を広範にする機縁がここに提供されたことを，私達は，深く欣びとするものである．

データ解析に苦闘する統計家のきめ細かな努力をよく体験したうえで，情報

量規準で統計的手法を統一的にまとめようとする本書は，入門解説書であると共に，問題提起書でもある．本書には，検定論の放棄とか，標本分布論の回避とか，統計数値表の無用とか，推測統計学の築いた伝統に対する批判が，底流しているようである．編集者はこれらの意見については，著者達の意見に直ちに全面的に同意するものでないことを，特に断っておきたい．しかしながら，情報科学と統計科学との連結領域には，datalogy はじめ，多くの課題群が山積みしつつあるのが現状である．R. A. Fisher, J. Neyman, A. Wald らの統計学を超えようとする時代の要請に対して，統計数理研究所の研究者の提起した上述の率直な批判は貴重であり，稔り多い発展の契機を包蔵しているものと思われる．

1982 年 11 月

編集者　北 川 敏 男

自　　序

情報量規準 AIC が統計数理研究所の赤池弘次氏によって導入されてから既に 10 年近く，その間の発展にはめざましいものがある．AIC の応用やそれに言及した論文の数は国の内外を問わず今では厖大なものになっている．このような論文を目にされた読者も多いであろう．この展開をつぶさに見られる位置に身を置くことができたのは統計学を学ぶわれわれにとって最大の幸運であった．本書を執筆するに至った最大の動機も，実は，われわれ自身が AIC を用いてその有効性を身にしみて感じたことにある．複雑な現実のデータには，データ解析の手を加えられて，その内包する構造をいきいきと語り始める一瞬がある．われわれは AIC を使って実際のデータを解析しながら幾度となくこのみずみずしい発見のよろこびを味わった．われわれが本書の狙いとしたのは，つたないながらも，この感覚を伝えることである．説明の便宜上，一応教科書の体裁をとり一般的なモデルをとりあげてはいるが，これらは代表的な応用例のつもりである．もちろん，これらのモデルを用いてデータの一応の解析をすることは可能であり有効でもあるが，われわれの本意は，読者自身が自らの目的に応じてモデルを開発しデータの構造を探るための手助けをすることにある．読者自身が自分で作ったモデルによってデータの構造を描き出すことに成功し，“データに語らせる”という感覚を味わわれたときはじめて本書の目的は達せられたといえよう．

本書の執筆にあたっては多くの方々の御指導と御援助をいただいた．統計数理研究所の林知己夫所長に，まず，心からの感謝を申しあげたい．林所長には，著者の遅々とした研究の進展にもかかわらず，暖かく見守っていただき，統計学の全般にわたって御指導・御援助をいただいた．赤池氏には統計学の初歩から懇切丁寧な御指導をいただいたのみならず，御本人をさしおいてこのよ

うな本を書く著者の非礼を心よくお許しいただいた．ここで断っておかなければならないが，本書は赤池氏の考え方に倣ってはいるが，氏とは独立に執筆されたものである．浅薄な理解や曲解による部分があるとすれば，いうまでもなく著者の責任である．AIC 導入の背景や AIC についてのさらに深い理解を得たい読者には，赤池氏自身の著述を直接読まれることをおすすめしたい．また，九州大学名誉教授北川敏男先生には，AIC に関する理論に興味を示され，本講座の一冊として執筆するようおすすめいただいた．先生の御好意と御激励に深謝したい．

さらに，統計数理研究所の清水良一，尾形良彦，濱田義保の各氏には原稿に目を通していただき有益な御指摘をいただいた．田辺國士氏には日常の討論の中で多くの示唆を受け，桂康一氏にはプログラムの作成に当って多大な援助をあおいだ．参考文献の収集等については，日本銀行の浪花貞夫氏，緯度観測所の大江昌嗣氏，統計数理研究所の尾崎統氏の御協力を得た．併せて謝意を表したい．

本書の執筆を決意したのは 1977 年の日本統計学会大会で情報量規準が共通テーマとして討論された日のことであった．以来，著者の怠惰のために，5 年近くを経過してしまった．この間辛抱強く本書の執筆をお勧め下さった共立出版の佐藤邦久氏にお礼を申しあげたい．

1982 年 10 月

坂　元　慶　行
石　黒　真　木　夫
北　川　源　四　郎

本書の視点と構成

すべての科学の目的は特定の現象に内在する規則性・法則性を見い出し将来に対して有効な推論を行うことにある．数理統計学の目的は，観測されたデータに基づいて，不確実な現象の特性を確率によって表現し，将来の観測値の確率分布を推定し，予測や制御に資すことにある．

この目的を果すためには，1）データの特性を表現する確率分布（統計モデル）を，現実に生じるさまざまな分析目的に応じて的確に構成すること．2）考え得るいくつかのモデルの良さを評価・比較しうる基準を提示すること，が不可欠である．従来の統計学がその厖大なモデルや理論の集積にもかかわらずなお実用性に乏しかった原因はこれら2点に関して未成熟であったことにある．わけても従来の統計学で仮説（統計モデル）の評価の手続きとして用いられてきた統計的仮説検定は，想定される数多くのモデルの評価・比較という実用上の要請に対してあまりにも無力であった．

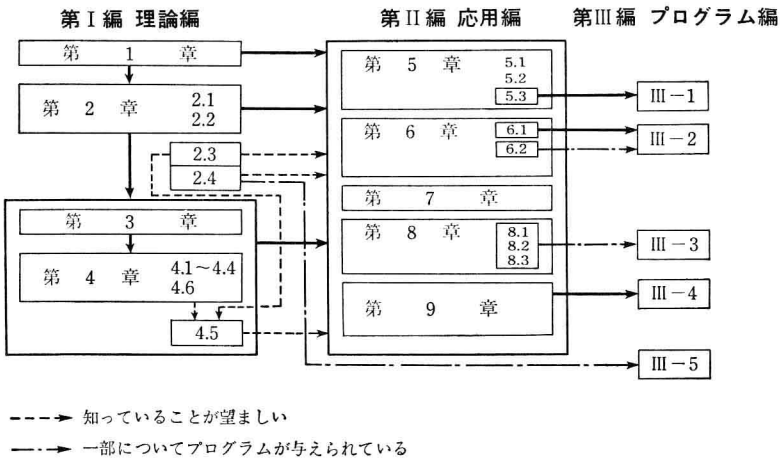
赤池は，尤度という概念を情報量の視点から見なおすことによって，異なるモデルも尤度に基づく客観的な規準によって比較可能であることを見い出した．この規準がAIC（赤池情報量規準，Akaike Information Criterion）と呼ばれるものであり，これによってはじめて想定された数多くのモデルの良さを統一的に比較することが可能になった．重要なことは，この統一的な評価規準の導入によって，不適切なモデルを不適切とみなして排し，より適切なモデルの開発を促進し，よって諸科学の発展に一層寄与しうるようになった点にある．

本書は，従来推定や検定，あるいはデータの記述法としてとりあつかわれてきた数理統計学の諸問題をモデルの構成と情報量による評価という一貫した視

点から見直したものである．その結果，とりあげられているモデルは従来の統計学書とほとんど同じものでありながら，個々の標本分布論を排し，統計数値表を不要にした点など，趣きを異にする本になっている．われわれの視点に興味をもち，理解しようと思われる読者には通読されることをお勧めするが，実用的な手法をともかくも使ってみたい読者は，本書の第Ⅱ編のどの章でもよい，必要なところから読みはじめられたい．

〔**本書の構成と読み方について**〕

目次からわかるように，本書は3編から構成されている．各章節の関連の概略は下図のとおりである．



第1章と第2章の大部分は従来の統計学のテキストに依拠してその結果を簡単にまとめただけで独自のものは含まれていない．これに対し，第3章以降は終章まで本書独自の考え方に基づいて執筆されている．特に第Ⅱ編では主な統計的解析法について新しい見地から例題と解法が示されている．したがって，一通り統計学を学んだことのある読者は第3章から読みはじめられるとよい．また，初心者は，必要に応じて他のテキストを参照しながら，第1章から読み進められるのがよい．なお，目次中の章節の番号の右肩の＊印はやや詳細な説明にわたることをしめしており，読みとばしてさしつかえない．

なお，第Ⅰ編は3人が共同で執筆し，第Ⅱ編以降の各章は3人が手分けして執筆した．執筆分担は下記のとおりである．

坂元　第5章，第6章，Ⅲ-1，Ⅲ-2

石黒　第7章，第8章 §8.4，第9章，Ⅲ-4，Ⅲ-5

北川　第8章，Ⅲ-3

目　　次

第Ⅰ編　理　論　編

第Ⅱ編　応　用　編

第Ⅲ編 プログラム編

第Ⅰ編　理　論　編

第1章　確率と確率変数

われわれは，日常生活の中で「十中八九」とか「万が一」というような表現を用いることがよくある．これらの言葉は，ある事象の起こる可能性の程度を表わしている．また，同じ意味あいで，確率という言葉を用いて，「お年玉つき年賀はがきが1等に当る確率」という表現をすることもある．数理統計学では，この確率という概念が重要な役割を果す．つぎにいくつかの例をあげよう．

［例 1.1］　当りくじ1枚，空くじ2枚の中からでたらめに取りだした1枚のくじが当りくじである確率

［例 1.2］　硬貨を1回投げたとき表が出る確率

［例 1.3］　歪みのないこまをまわして，こまが倒れた点と基線のなす角が30 度以下である確率

［例 1.4］　正しいサイコロを投げたとき1の目が出る確率

［例 1.5］　52 枚のカードから1枚をひいたとき，エースが出る確率

これらの例に共通な事実は，いずれも，考え得る限り同じ条件のもとで幾度も幾度も同じ実験や試行を繰り返したとき，着目した結果がどの程度の割合いで現われるかを問題にしていることである．統計学ではこのような意味で確率を問題にすることが多い．この章ではこの確率の基礎的な事項について述べる．

1·1　事象と確率

ある偶然を伴なう実験の結果が $\omega_1, \omega_2, \cdots, \omega_s$ のどれかになるとするとき，これらの結果すべての集合を**標本空間**と呼び Ω で表わすことにする．たとえば，前出の［例 1.1］では，当りくじが出る，空くじのうちの1枚が出る，空くじのうちの別の1枚が出る，という3通りの結果がそれぞれ $\omega_1, \omega_2, \omega_3$ で表わされ，Ω は $\{\omega_1, \omega_2, \omega_3\}$ で定義される．標本空間の部分集合のことを**事象**と呼び，E と表わすことにする．そして実験の結果，事象 E の元のひとつが観測されたとき，事象 E が起こったということにする．前出の［例 1.1］では，当りくじが出るという事象 E_1 は $\{\omega_1\}$ を意味し，空くじが出るという事象 E_2 は $\{\omega_2, \omega_3\}$ を意味する．このように事象 E は，

$$E=\{\omega|\omega \text{ に関する条件}\}$$

と書ける．ここで，ただ1つの結果から成る事象を特に**根元事象**と呼ぶことにする．また，どの結果も含まない事象も事象の1つとみなし，**空事象**と呼び，ϕ で表わす．事象 E の**余事象** E^c とは，E に属さない根元事象の集合のことである．

2つの事象 E_1 と E_2 の**和事象**とは，E_1 か E_2 の少くとも一方に属する根元事象の集合のことで，記号 $E_1 \cup E_2$ で表わす．また，E_1 と E_2 のいずれにも属する根元事象の集合は事象 E_1 と E_2 の**積事象**と呼ばれ，$E_1 \cap E_2$ で表わされる．そして特に，$E_1 \cap E_2=\phi$ という関係があるとき，すなわち，2つの事象 E_1 と E_2 が同時には起こりえないとき，事象 E_1 と E_2 は**排反**であるという．

これら3つの用語は無限個の事象 $E_1, E_2, \cdots$ が与えられている場合にも同様に定義することができ，それぞれ，$\bigcup_i E_i$，$\bigcap_i E_i$，$E_i \cap E_j=\phi$ 等で表わすことにする．

ある事象 E_1 に属する根元事象がすべて別の事象 E_2 に含まれるとき，E_1 は E_2 の部分事象であるといい，$E_1 \subset E_2$ で表わす．また，このとき同時に $E_1 \supset E_2$ も成り立てば E_1 と E_2 は等しいといい，$E_1=E_2$ で表わす．