コンピュータサイエンスシリーズ

情報処理と統計数理

林
樋口
駒沢

情報処理と統計数理

林　知己夫
樋口伊佐夫著
駒沢　　勉

産業図書

<著者略歴>

林　知己夫（はやし　ちきお）

昭和17年　東京大学理学部卒
昭和22年　統計数理研究所入所
昭和30年　理学博士
　　同所第2研究部長
昭和49年　同所長
　　現在に至る．

樋口　伊佐夫（ひぐち　いさお）

昭和22年　名古屋大学理学部卒
　　統計数理研究所入所
昭和30年　同所第2研究部第2研究室長
昭和46年　理学博士
昭和49年　同所第2研究部長
　　現在に至る．

駒沢　勉（こまざわ　つとむ）

昭和34年　早稲田大学理工学部卒
　　統計数理研究所入所
昭和46年　同所第4研究部第1研究室長
　　現在に至る．

情報処理と統計数理　　定価 3200 円

昭和45年5月30日　初　版
昭和54年4月12日　第7刷

著　者　林　知己夫
　　樋口　伊佐夫
　　駒沢　勉

発行者　森田勝久

発行所　産業図書株式会社
東京都千代田区外神田 1-4-21
郵便番号 101-91
電話東京 (253) 7821(代)
振替口座東京 2-27724 番

文弘社印刷・関口製本

序にかえて

——統計とコンピュータ処理の諸相——

統計は統計的方法によってつくり出される．データ解析を究極の目的とする統計的方法は，データ処理とは切っても切れぬ関係にある．データ処理の道具が貧弱であった場合には，統計的方法論もそれに応じたものが考えられていたわけである．統計量の計算法などにおいてもそうした工夫がなされていた．しかし，データ処理の道具の能力が飛躍的に増大してくると，これをもとに考えた統計理論が発展してくることになる．また，こうした方法を使えばこうしたことがわかってくるということが明らかになり，このことの実現の隘路がデータ処理能力の不足になってくれば，その道具を新たに開発することになってくる．こうして新しくできたものによるデータ処理は，また，新しい統計理論の開発を促すという相互作用が見い出されてくる．これが現状である．データ処理能力の増大はコンピュータの能力増大に負うところが多い．コンピュータといっても特にディジタル・コンピュータの能力の飛躍的増大が与える影響は大きい．しかし，アナログ・コンピュータの性能の向上やディジタル・コンピュータとの連繋によるハイブリッド方式の開発による能力の増大も見逃せない．コンピュータの中心部の性能がいくらあがっても，その入出力が便利なものでなくては統計におけるデータ処理に有力なものとはならない．計算能力があり，いくら演算処理速度が大であり，記憶力が大きくても入力が不便であったり，結果の表示が不便なものであったりすれば，思うように使いこなせるものではない．現状では，我慢しつつ，不便な入出力を使ってデータ処理をしているわけであるが，これでもかなりの成果をあげている．ここの能力がさらに開発されれば，ますます統計的解析には有利なものになってくる．この入出力はしだいに開発されてきているように思われる．現状では，まだ，パンチ・カード入力が主力，ライン・プリンタ印刷が主力といったところであって，ためにあたら多くのデータが死蔵されたり，解析結果から有効な情報を引き出すために多大の努力を必要とするような情況である．統計処理の立場からいえば，コンピュータ本体の能力の増強とともに入出力の開発にセンスを生かしてもらいたいものである．

いまのところ，統計によるデータ処理は，入出力の問題に始まり，集計や複雑な計算の処理，シミュレーションといったところが中心であるが，さらに応用が広まることが望まれる．もう少し順を追って統計におけるデータ処理をながめてみよう．まず入力のところであるが，統計では諸々様々のデータを扱うものであるから，得られたデータをパンチ・カードに打って入力するだけでは能がなさすぎる．一番単純には，調査した結果（これは調査票に記入されているコードを意味する）をコード票に写しこれをパンチする．しかし，その間の誤りや時間のむだが大きい．つぎがマーク・センスといわれる周知の方法でパンチをするかわりに，カードにマークをつける．このマークのついたカードを機械にかけると穴があいてこれが入力となるといった方法である．これだけのことであるが，その信頼性などに問題があるようである．もっと直接にしたいとなると，調査票自身を入力に使うのである．コード票転記をやめるのである．調査票に記入したもの（もちろん一定の個所をマークする）そのものを入力として，調査情報を記憶に蓄えるのである．こうなると中間の人手が省けて能率的になる．調査票そのものとは，質問と回答欄が記入してあるものを指す．質問が別の紙であったりすれば，調査員が記入誤りすることが多い．質問と回答の記入欄とがついていなければならない．こうした調査票の型式でありながらそのまま入力となるところが狙いである．また，地図など書いてあって，どこからどこまでいったか（交通における OD 調査といわれるもの）を質問し回答を地図上にチェックする．これがそのまま入力となり（地図上の地点チェックがそのままコードとなるように仕組まれてある）記憶に蓄えられることになる，という様式も考えられている．これははなはだ便利なものである．地点をきき，コードを調べ記入するのであれば，時間もかかるし，誤りも多い．地図（地名は書きこんである）をみせてチェックしさへすればよいとなるとはなはだ便利である．また，日本の文章を入れたいようなとき，いちいちコードに直してカード化するのは大変である．ローマ字に直して入れるのも大変である．漢字仮名混り文をそのまま打って，それが自動的にコード化されて記憶装置にはいるようにすることも便利なことである．さらに便利であるためには，文字（数字）そのものを自動的に読み取り，これが記憶装置にはいるのが望ましい．こうしたこともやがて実現されることになろう．調査や実験結果が時系列的曲線で得られている場合などはどうであるか．いままでは曲線の座標を物差しで読み取り，これをカード・パンチして入力をするといったことが行なわれていた．ここでは大量のデータを処理することはまず不可能なことである．この入力のところで行きづまってしまう．このようなときは曲線自動読取り装置にかけ，アナログ量に変換，

さらに A-D 変換機によって，ディジタル・コンピュータの記憶に入れることが望ましい．時系列曲線でない一般の曲線のときは，曲線判読装置（自動的ではなくマニュアルで曲線を追跡する）などで曲線情報でコンピュータに入れることができれば便利である．ライトペン装置と称するものがあり，一定のブラウン管上に書いた任意図形をディジタル情報に直して記憶に格納することもできるが，入力としては便利なものである．こうした図形自動判読装置も漸次実用化されてきている．書かれたものではなく，取ってきたデータが磁気テープである場合がある．たとえば，自動車の振動の記録，心電図，心音図，調査の記録（コードを磁気テープに記録）などある．また必要な情報が測定されたアナログ量のままコンピュータに直送されてくることもある．A-D 変換機にかけ，自動的にディジタル情報化して記憶装置にしまいこむのである．こうすれば，調査・実験データが，このままディジタル・コンピュータの中にはいりこみ解析が可能となる．また磁気テープをアナログ・コンピュータにかけ事前処理を行なって，アナログ・コンピュータに A-D 変換機を接続して，アナログ情報をディジタル化してディジタル・コンピュータ記憶にしまっておくことも望ましいことである．かわったところでは，音楽の曲の分析にあたって（たとえば，弦楽四重奏曲のバイオリンの部分(重音は除く)）これを入力するのにいちいちコード化していたのでは大変であるので鍵盤を用い曲をたたく，これがディジタル情報に直されて（音譜の高低，長さ）ディジタル・コンピュータの記憶にはいるような装置，レントゲン写真の自動判読装置（レントゲン写真を1本の光が走査しこの光の反射あるいは透過の度合をディジタル化して記憶装置にしまいこむ）など工夫されているが，入力としては有力なものである．さらに，また，いわゆるデータ通信網によって測定結果がそのまま通信されてくることもある．これが直接コンピュータにはいりこむようにするシステムの開発が望ましいものである．このほか，音声入力などあればさらに望ましいものになろう．出力の方では，印刷形式の工夫はもちろんであるが，このほか，図形表示，漢字仮名混り文表示などのできる機械印刷が容易にできる装置，音による表示，データ通信（制御通信）など考えられる．これは，計算され，解析された結果を必要情報として送り出すこと，見やすい形に表現すること，結果の本質的部分がはっきりするような形に表現することが目的となる．また，こうした入出力装置がそなえられ TSS 方式を介して人がコンピュータと自由に応答し，試行錯誤を繰り返すことも大事なことになる．

以上長くなりすぎたが入出力の装置が，統計におけるデータ処理において大切なことを示したつもりである．さて，つぎがデータの処理であるが，ここでは，計算，判断などの

操作が大事な働きをもつ．統計理論にもとづく諸操作が行なわれるのである．統計操作にしても，よいサンプルから情報を引き出すばかりでなく，いわゆる歪んだサンプルから，妥当な情報を引きだす方法も開発されてこなくてはならない．さて，以上のような諸操作が一貫した形で行なわれるようになるのが望ましい．たとえば多次元解析において，もとのデータがはいれば，これから成分分析が行なわれて，ある大きさ以下の特性根が切りすてられ次元が決定すると同時にウェイト・ベクトル(特性ベクルト)が決定され，これが図示されるとともに，もとの各個のものの数値がそれぞれ次元で算され，これもまた図示される，ということが一貫して行なわれるようになる．いわゆる数量化の計算分析でも，もとになるデータがはいれば，集計され，データの扱い方が決定され，計算され，判断され，情報が選別され分析のための適切な諸表示がされるという一貫操作が行なわれる．またこういう例もある．2つの標本平均があったとき，これに有意な差があるか否か検定され，95%の信頼度で有意差がないと認められたとき，2つの標本が合わされて1つの平均が算出され，この精度が95%の信頼度のもとにプラスマイナスいくつという形で表示される，という一連の操作が続いて行なわれるようにする．集計分析においても，単純集計，関連集計(このときの製表はいかに見やすいものにするかを工夫する必要がある)尺度作成(調査結果にあるウェイトを与え物差しをつくる，この妥当性の検討もコンピュータによって自動的に行なう)，尺度を用いてのいろいろの分析が一貫して行なわれるなどのことがある．関連分析ではすべての表を打ち出すのであれば表の数が多くなりすぎ，検討できなくなるので χ^2 検定方式は適切な関連性の物差しをつくりこれによる判定を行なって，関連性のあると見なせる表だけ打ち出し，残りは関連性がなかったとして項目名だけをあげておくというような処置をすれば利用しやすい形になる．ここでは，ある程度の整理までコンピュータにやらせるわけである．時系列分析では，季節調整の出し方なども，もとのデータを入れれば，趨勢・サイクル，季節指数，不規則部分が分解されてグラフ化されるという一貫したことが行なわれる．自動診断などにおいては諸計測データがはいれば，自動的に判読され，一定の論理に従って診断名がつけられ，この精度が何%という形で打ち出されるということになる．たとえば心臓の疾患などにおいては，磁気テープにとられた心電図や心音図やレントゲン写真による心臓の形の読みが入力されれば，これが一定の方式に従ってディジタル化され，一定の論理によって病名がある精度表示のもとに打ち出されるということになる．こうしたことはパタン認識のひとつである．全体の総合判断というこのパタン認識はディジタル・コンピュータには最も不得手のものであるが，統計の情報

処理ではきわめて大事なのである．また，シミュレーションにおいては，モデルが命令として与えられれば，乱数を用いて（乱数発生機によってつくりだされたものが直接ディジタル・コンピュータの記憶にはいるようにする．あるいはつくりだされた擬似乱数による，ともにそれぞれ一定方式で検定されたもの）結果が計算されその分布が図示され，その統計量が計算されるというようなことにもよく用いられる．たとえば，それぞれ大きさや形の異なる粒子のパッキングの様態を実現することができるし，交通量と交差点信号系の制御の最適化ということも研究できるし，情報伝達の様相，伝染病伝播の様相を諸条件のもとにつくってみることもできるし，不規則入力にもとづくフィードバック系の状態の検討もできるし，天然林の発生模様，すなわち何年後かの樹木位置図，各樹木の胸高直径，樹高が一定のモデルに従ってシミュレーションにより求められ，これが図示されるといったことも行なわれる．また一定領域内に樹木位置，胸高直径が与えられていれば，この中にランダムに点をおとし，ビッテルリッヒ法による胸高断面積の推定を行なう実験も行ないうる．立木位置や常数の大きさを変えてみて，推定がどうなるかも現地調査をやらずにコンピュータの中でみることもできる．いろいろの条件のもとで（これは応用分野によって異なる，たとえば，ある動物の行動のシミュレーションなど）ランダム・ウォークの実験を乱数を用いて行ない，この結果を図示したり，行動範囲の分布を求めたり（一定時間後の位置あるいは，一定時間後までランダム・ウォークしたとき，出発点から最も離れた位置の分布など）することも行なわれる．このほか，数えればきりがないが，調査対象の層別，標本割当て，抽出などの操作も一貫したプログラムで実行することもある．このときは機械的に命令をくだすのでなく，種々の制限条件を設けこの条件のもとでの層別を行ない，しかる後，一定の確率に従って乱数を用い標本を抽出，これを割当て標本とともに印刷する，といったことを行なうのである．これまでのものは統計だけに限ったものであるが，ある事象がシステム化されたとき，全体を結べる考え方に以下述べる統計的考え方が主軸をなし，全体の流れの一部に上述の統計操作がすっかりはめ込まれたり，各所各所にさまざまな統計的方法がちりばめられたりすることも多いのである．このシステムがフィードバック系であったり，オンラインのこともあるし，オフラインのこともあるが，それに応じて統計的方法が活用されるわけである．

いずれにせよ，統計においては，データの獲得からあと，人手を加えずコンピュータに情報がはいり，所定の統計的方法によって解析され，判断され，予測され，決定された諸結果が，見やすい形に表示されて出てくるといったことが最も望ましいのである．これへ

向かって努力をしていることになる．このとき，可能なかぎり一貫した一連の操作が切れることなく行なわれ，しかもそれらが妥当性をもって行なわれ必要とする情報を常に取り出して数字化し，文字化し，図表化し，容易に人目にふれるようにできる，といったことが大事なことになる．統計では，非常に種類の多いデータの解析が行なわれるので，コンピュータ操作も多岐にわたらざるをえないのである．なお，数値計算法においても統計的アイディアが重要な働きをもつこともある．きわめて元の多い行列式の計算など標本抽出の考えから繰返しなくランダムに行，列を選び，積をつくり，偶・奇置換の判断から符号をつけて，計算を行なうこともできる．

最後になったが，情報検索のことに少しふれておく必要があろう．情報検索そのものが，統計の問題として意義をもつことも当然ありうるが（検索に対して後述する判別や多次元解析，数量化の考えが有効に使われる場合もありうる）多くは，データ供給源としての情報検索である．欲するときに欲する範囲のデータを取り出し，突き合わせ，統計的考えから解析するということである．情報検索そのものは統計の問題ではないが，情報検索の仕方に対する方法を提供し，統計的解析に有用な情報の検索方法を開発するように問題を提示し，情報検索されたデータをもとにして統計的分析を行なうといった相互関連性をもち，お互いに稗益しあうべき関係にあるものといってよかろう．

林　知己夫記

目　　次

1. 基本的な統計的方法とコンピュータ処理

1.1 統計的方法の基礎概念

ここでは，統計的なものの考え方の基礎的部分について述べることにする．本書のように，コンピュータとの関連における記述を中心とするものにおいては脇道のようにも考えられるが，現象解析では統計解析の理論をそのまま公式のように適用し，それにもとづいてコンピュータでデータ処理をしても，妥当な結果が必ずしも出てくるものではない，ということを銘記していただきたいので，あえて記載するわけである．現象解析における統計的方法においては，その基礎概念が非常に重要な働きをするわけである．この考えにもとづいて統計的解析の方針がきめられ，統計的方法が開発され——応用され——データ処理がなされるという手順である．

統計数理では，現象に立ち向かうとき，まず，何が肝要であるか，妥当性・有用性はいかにして得しめるべきかを考えるのである．これはきわめて漠とした言い方である．しかも当り前のことであると思われるであろうが，そこが見失われるのが不思議といえば不思議なことである．漠とはしているが中心的な位置を占め，方向づけを与えるエネルギの根源であるといえよう．さて，これから基礎の概念のうち大事なものいくつかを説明してみよう．

1.1.1 集団について

まずもっていえば，統計学は集団というものを取り扱うのであり，また集団というものを背景としてものを考えるのである．もちろんここでは，いろいろの現象を「集団の現象とみなせるようにして取り扱う」ということも含んでいるのを忘れてはならない．集団とは，即物的には，2つ以上の要素からなるものということができる．なお，1つの要素を取り上げたとしても，この背景に，そのよってきたる論理的な（思惟的な，こうはいって

も以下述べるような立場からそうすることが可能であるとみなされなければならない）集団を想定することもあるわけである．集団といっても，これを広狭自在に考察し，利用する必要がある．しかし，広狭自在にするとき，その段階的区別は十分意識されねばならない．ここに，集団の要素は，ふつう人とか物とかそういうものでもよいし，「事象」というような抽象的な要素であっても差し支えないのである．要素のとらえ方のところも問題のフォーミュレイションによって，有効ならしむればよいわけである．さて，背景としての論理的集団とはいかなるものであるか，といえば，つぎの節の確率という考えを積極的に利用し，母集団なる概念によって表現されるものである．これについては後述する．

1.1.2　操作的な考え方

統計学で大切な考え方のひとつの柱ということができる．測定なくして，標識(後述)はないのである．「われわれの測定」によって「標識」をわれわれが与えるのである．即物的に測定されえぬ（可能性のない）ものは取り扱わないし，測定操作を媒介せずして，標識は存在しないと考えるのである．このことは，後に確率を考えるときにも大切なことになってくる．たとえば，経済学でいう概念の需要とか供給とか国富とかいうものは実際に測定できるものではない．したがって標識とはならない．しかし，購買量，出荷量，生産量といったものは単位を定めれば測定できるものである．もちろんこの単位の定め方は大事なことで，今後の解析における妥当性を左右するものである．この「操作」という意識がないと，「統計の嘘」といわれるものが出てくるのである．いわゆるオペレイショナリズムの考え方との関係はきわめて深いというべきであろう．

1.1.3　統計における2つの立場について

これは，寺田寅彦の言葉を借りれば「最初の言葉」，「最後の言葉」に相当するものである．1つは最初のさぐりを入れる消息子の働きをするものであり，もう1つは，われわれの知りたいこと，それを知ることが妥当性のあること（外的基準といおう）を端的に指し示すもの，目的とするものを直接的に指し示すものであるといえよう．この2つは，はなはだ異なる機能をもつのである．それぞれの特色があるので，取り違えては，大変な誤りとなってしまう．

例で説明してみよう．病気の分類，すなわち病名を与えることは最初の言葉となる．病名を与えることは，取り上げられたいくつかの計測値群の似たものを集めて分類すること

に対応する．それぞれの病名に応じ完全に同じ計測値パタン(病状ということができよう)が出ていれば文句のないところであるが，実際にはそうはいかない．この点では同じだが，この点では異なっているというような場合があるし計測値の隔り具合も異なっている場合もあろう．計測値といったが，計測されたものが妥当性ある意味で，必ず量で与えられるものでもない．このようなとき，似ているということを統計的にきちんと表現しておかねば，似たものを集めるということはできない．似ているということを観念的にいうことはできるが，これを操作的——測定したものにもとづいて明確に定義できるもののみを取り扱うこと——に一義的に定義づけることはできるものではない．操作的には，いくつかの定義の仕方があろう，このうちで，「よさそう」なものを決断をもって選ぶのである．「よさそう」とは，その定義が，可能なる限り客観性をもちかつ統計分析の処理が妥当性を示しつつ，容易であることを意味するのである．

この定義に従って，計測値の同じ傾向を示すものを集めて分類を行なうのである．この分類に名がつけられたものが病名と考えられるのである，この分類は，他の『よさそうな」定義にすれば，当然変わってきてしまうこともおこるのである．2つの定義で決定的に一方がよいとは，このままではいいきれないのである．いずれか一方がその点でまさるというためには別の検証を経なければならないのである．したがって，ある定義に従って行なわれた分類は，最初の言葉を与えるもので，これをもとに研究を進め知識をふやしつつ第2の立場に立って，検討を重ねなければ明確なものとはならないのである．

一方，第2の立場を述べた「最後の言葉」の方は，ある計測値を示すものに対して，甲という治療法がよい結果を与えるか，あるいはよい結果を与えないか，うんぬんの計測値パタンをもつものは予後がよい，あるいは予後が悪いか，などをいおうとするときに用いられるものである．これは病名の決定というものではなく，「どうすればよいか」を指し示すものとなるのである．したがって，よい結果を与えるかどうかという「はっきりした」ことを可能なる限り精度高く予測することが目的となるので，どうすれば，精度よい予測ができるか，どのくらいの的中率があるか，が示される方法を考えることになるのである．

第1の方法で得られた結果は，解釈をすることが非常に必要となるのである．しかし，解釈そのものはデータからそのまま得られるものではなく，われわれが用いた，ひとつの方法によって的確に描かれた結果から，仮説をたてつつ考えるべきものなのである．このため，必ずしも解釈は一義的にきめられるべきものではないのである．いくら解釈をしても，これは「推量」である．決定的推論の結末ではないのである．この方法によって得た

解釈が壮麗に書きあげられた論文をよくみるが，これはひとつの想像であり，類推なのである．これを確かめるためには，得られた仮説をもとにして，再び調査・実験を組み目的に対する検証（これには「最後の言葉」による方法が用いられる）をしなくてはならない．解釈は人々に誇示するものでなく，机の引出しの中にしまっておくべきものである．一方「最後の言葉」の方は，たとえばうんぬんの条件をもつ候補者は選挙の結果当選するか，落選するか，うんぬんの条件をもつ人は甲という商品の購買層と考えられるか（どの程度の購買層となるか）などを明らかにしようとするものである．

このような2つの立場によって，なされる統計的方法は，それぞれの特色をもつものである．前者は，混沌とした事象に斧を入れ，「ああであろうか，こうであろうか」と考え学ばれる過程に有効な道具なのである．われわれの知恵の閃きを誘うものであり，手控えなのである．得られることは間接的な知識であり，予想であり，教養なのである．この意味で大変大事なものなのである．しかしこれだけでは実際の役に立たないのである．一方は，「こうであるぞ」との止めを刺すための方法なのである．

もう少しこれらの2つの方法の使い方を説明してみよう．実際の例でいおう．稲の種の分類を計測値パタンから行なうことを考えよう．いま，稲の学問が進んでいて――実際そうなのである――その分類がこれまでの学問の蓄積でわかっていたとしよう．分類がわかっていれば外的基準として取り扱い，これらの分類に対して計測値たる要因がどういうふうに関係し，要因のうちどれが分類に有効にきいているか確かめることができる．しかし，このままであれば，結末は稲の問題だけに終わってしまう．そこで外的基準がないとして，ある仮説に従って類似度を設定し，計測値たる要因パタンの同じものを集めるような統計的操作を行なってみることにする．そして，これによる分類がすでに得られている分類と同じかどうかを検討するのである．もし一致しているとすれば，われわれの行なった「類似度」の定義，統計的操作が妥当性あるということになる．このように，あらかじめ分類のわかっているものに適用して，われわれの方法の妥当性が確かめられていればますます強い．

こうしてわれわれの方法が成功したとき，分類のわかっていないもの，たとえば，ある種の昆虫の分類や化石の分類といった，似た現象に対して，計測値パタンからわれわれの方法によって分類を考えることに意味が出てくる．われわれの方法によって得られた分類をもとにして，仮説をたてそれを深めていくことが可能となるのである．こうしたことは，外的基準のある方法を適用したのでは得られないのである．拡張応用がきくというと