統計数理 第35巻 第1号 (1987)

"Electronic Journal of Data Analysis" の構想

慶應義塾大学理工学部 渋谷政昭・柴田里程

(1987年5月 受付)

統計学の応用に関心をもつ研究者の仕事の環境が，近い将来に先端技術により大きく変化する．その好機をとらえて，統計学応用を扱う学術雑誌の形態を根本的に変革し，データと解析とソフトウェアとを統合した電子ジャーナルを創刊し，新たな研究，発表，交流の環境を創造しようというのが "Electronic Journal of Data Analysis" の構想である．

本稿は提唱者たちの最初の覚え書きである*．統計学術雑誌の問題点，技術変革の評価，電子ジャーナルの意義，その実現可能性，人工知能研究との関連，実現の段階について現在理解していることを報告する．

1. 統計学術雑誌の制約

統計学応用の研究論文における悩みはデータの扱いである．できるだけナマのデータが望まれていても紙数の制限により，小規模データか，要約した値しか掲載できない．また，データが大きければ解析結果をレフェリーがチェックしたり，他の解析を提示したりすることもできない．読者からの反応もデータの解析に立ち入ることまでは望めない．

統計学の教科書でも悩みは同じである．たとえばCox and Snellの良い教科書 (Cox and Snell, 1981)は，130ページの中に24の事例と15のデータを載せている．当然各事例は小規模なものとなり，著者たちもその制約をボヤいている．大学の演習用としても，自分で入力し，チェックし，プログラムを作り，計算を確めながら解析する，という目的には良いが，統計パッケージを利用した演習には小さすぎる．Andrews and Herzberg (1985) のデータはそれほど大規模ではないが，現在のところ磁気テープの形では入手できない．日本オペレーションズ・リサーチ学会のデータ収集 (森口繁一，1976) は先駆的であったが継続しなかった．奥野 他 (1986)は工業における8つの事例の説明と解析で，興味深く貴重であり，再検討が望まれる．

学術雑誌の製作はまた技術的，経済的な諸困難に面している．著者と編集者の立場からは，誤植が多く，英語綴りと数式を正しく印刷してくれる業者が少なくて困る．印刷業者の立場からは，能力の高い植字工，タイピストを雇うほど割の良い仕事ではないであろう．ワードプロセッサが普及して入力の費用が下がっても，質の向上とはならない．著者自身がワードプロセッサを使用しても，写植機との連結が不十分で，入力の努力が効を発揮しない．良い雑誌には原稿が集中して出版が遅れ，原稿の貯まっていない雑誌では印刷所への入稿が安定しないために出版までの月日が長くなる．

---

* 応用統計学会年会 (1987年4月24日 東京) での報告に加筆した．

2. 研究環境の急変

ごく近い将来にどのような商品が利用可能になるだろうか．まず，32ビット・ワークステーションが学科単位で購入可能となり，大規模計算以外の多くの仕事を手近に処理できるようになる．ワークステーションにはレーザ・プリンターやグラフィックス機器が付随していて机上出版が可能となる．第2に光ディスク等の実用化によりギガ・バイトの位（くらい）のデータを机上に保管できるようになる．我々が一生の間に書くことができる論文，プログラムはもちろん，一生の間に読むことができる量の論文も机上に収まるようになる．第3に良質な基本的ソフトウェアが大衆製品となり適正な価格で市場に出回るだろう．最後に，これがもっとも重要な点であるが，データ通信が安価となり，計算機ネットワークが大学間を継ぎ，各種の水準のネットワークが研究室に入り込んでくる．一言で言えばパーソナル・コンピュータの第3世代が始まりつつある．

好むと好まざるとにかかわらず，商業主義の力により，新技術の波が侵入してくる．一方古い技術の死滅も早い．5年前に大学でかなりの台数のタイプライターを購入していた．今なおタイプライターを使用している人々の比率はどのくらいであろう．8ビット・パーソナル・コンピュータはほとんど使われていない．〝消費者〟となっていつも受け身になるだけでなく真にわれわれが必要とするものを整理し，要求し，実現する努力が必要である．

3. 電子ジャーナルの利点，目標とデータベース

すべての情報が媒体と独立に作成，転送，保存されるようになりつつある．雑誌，書物も紙を離れて存在し得るし，すでに作成され，市販されている．ランカスター（1984）の原著が出版されて10年近くとなり，必要な技術が十分に身近なものとなっている．

電子学術雑誌EJDAとは：投稿は書込み専用ディレクトリへの転送と編集長への電子メイルである．査読依頼は投稿ディレクトリのパスワードの送付である（公開鍵システムを作らないならばパスワードは別送となる）．論文採択は，読出し専用ディレクトリへの転送とニュース欄（news, bulletin-board）での告示である．つまり発刊（publication）と同時で不定期である．

紙の雑誌と違って論文の本文，つまり解析の部分と，データとその説明，ソフトウェアとその説明，の3部門を独立したものとみなすことができる．データ中心，ソフトウェア中心の投稿も，論文中心の投稿と同様に評価されるだろう．それぞれの索引があれば利用者は独立に〝読む〟ことができる．充実した検索システムの可能性と，利用の便利さもEJDAの魅力である．だれが〝読んだか〟を著者に知らせることも（著者・読者が希望するならば）できる．ひとつのデータについて複数の解析を積み重ねることによる研究の交流も進むであろう．

EJDAが豊富なデータを蓄積するといっても，これは特定の情報システムのためのデータベースとは違う．経済分析，地震予知，癌研究，新素材開発などのために，各所で多量のデータが蓄積されている．これらは情報システムとしての目的が明確に設定されているかどうかは別として，かなり詳細，包括的，大量のものであろう．EJDAに自然に蓄積され，あるいは意欲的に収集されるのは，むしろ〝統計的課題に関して典型的なデータ〟である．たとえば，R. A. Fisherのアヤメのデータ（Fisher (1936) が線形判別関数の例に使用して以来，多変量解析でしばしば引用されている．Andrews and Herzberg (1985) に原データの解説がある）のように，解析法を考える契機となるデータである．大規模データ（すでにデータベースにあるデータ）については，アクセス法をデータの代わりに記述しておけばよい．

学術雑誌である以上，EJDAの目的は専門家のためのものである．モノがないと業績として

評価されるだろうか，という疑問が出されたが，"業績報告"に添付するためにはハードコピーを作ればよい．"公刊の学術雑誌" という概念に合うか，という疑問と，またこれと関連して，計算機設備の乏しい人々から，"どのように利用できるか" という疑問が提出されている．

一般的には，もよりの主要大学計算センターまで何とか計算機間通信を設置していただきたい．徐々に大学等の図書館が計算センターと融合して "情報センター" となるに違いない．少額の費用で雑誌にアクセスできる端末が公共の場所に備えられて，初めて真の公刊であろう．暫定的には，要約と解析部門だけを従来の形態で発売することも必要であろう．もっと広い読者層のための EJDA が必要であろう，という提案もあるが，当面は総合報告，解説などの論文分類項目を設けておくだけでよいだろう．それが増加すれば "特別号" の発行，つまりアクセスの異なる別のファイル，あるいは別のシステムに移すことになるだろう．

計算機は意志疎通（communication）の道具である．これを用いて著者，編集者，読者の間の交流をより滑らかにする可能性を追求するのが EJDA の計画である．小さなことであるが，諸種の研究集会の告知，講演申し込み，プログラム発送などは EJDA のニュース欄を用いて能率良く連絡できる．ハードウェアとソフトウェアが普及し，その費用を無視できるようになれば，電子ジャーナル発刊は従来の雑誌よりも経済的となる．

4. 知識ベースの構築

人工知能ないしエキスパート・システムについて，統計学からの評価と期待はさまざまである（たとえば，Billard, 1985 と Gale, 1986（以下，この章ではこれらの本の中の論文を年号とページだけで引用する））．誰でも思いつくこと，期待することを R.A. Thisted (1985, pp. 276-284) を参考に整理すると次のようになる．

(a) 計算機の対話的な使用中に，何をして良いか分からなくなったときの，いわゆる help 機能を的確にすること（特に統計学に限ったことではない）．もっとも必要になるのは，操作ミス（キーの押し間違い）で予期しない使用モードに入り込み，元に戻れないときである．望ましいのは，マニュアルを詳しく読めば理解できることを，端末から調べられる機能である．簡単な自然言語処理とメニュー選択などで，むだなく調べられることが重要である．もちろん良く書かれた，索引の完全なマニュアルの存在が前提条件である．文献検索を能率良くできるとか，データベースに関する照会を使い易くするとか，一口で言えば情報システムを親切にすることである（Guide for the perplexed）．

(b) 統計ソフトウェア・パッケージを利用して解析をするときに，諸手法のどれを使うか，結果の数値をどう眺めるか，についてシステムが案内し，警告し，援助するものである．利用者の問題とデータを用いて実地教育をするようなもので，現在のソフトウェア・パッケージの出力や，メッセージを親切にする，という程度の実現しやすいもの（Guardian of the novice）もあるだろう．基本的な解析法で必要なものを示唆して，専門家に相談を受けるための準備を整える（初診患者の予備問診に対応するだろう）ためのもの（Intelligent assistant）もあるだろう．

以上の機能は人工知能と呼ぶこともなく，これまでの技術でも実現可能であるが，実現が容易になり，本来の目的の機能とうまく連動して，利用者も手軽に使えるようになることが進歩である．P. Huber (1986, pp. 285-294) が言うように "ハンマーが人間の力を強めてくれる程度のものしか期待しない" ならば，期待外れになることはないし，各種の商品も現れるだろう．しかしエキスパート・システムにたいする期待は，もう少し高い所にある．

(c) 統計ソフトウェアの利用者を案内するにしても，いくつか用意した〝バック・ツアー〟のひとつを選んで乗せるようなものが，(b) とするならば，本当の問題解決，つまり利用者の新しい未定形の問題について，何等かの援助をするようなもの（Apprentice consultant）が欲しい．あるいは利用者が学習意欲をもつものの，当面の緊急の問題を解決するのに必要な事項(何が解けており，何が解けていないかも含めて）だけを能率良く学習したいときの〝目的別，能力別，実践的〟計算機学習（Computer assisted instruction）も望ましい．

(a), (b)にたいしては (c) は，より不定な状態をより動的に扱う．しかも利用者の直接の要求に受動的に答えるだけでなく，作業過程を見て能動的に介入する．これを実現するものは，

アルゴリズム（Algorithm)＋データ構造（Data Structure)＝プログラム（Program）

の模式になぞらえて，

知識（Knowledge)＋推論（Inference)＝専門的判断（Expertise）

であると言われている．知識は facts, heuristics, strategies より成り，論理式，述語，状態遷移図などで形式的に記述されねばならない．

Gale (1985, pp. 115-120, 1986, pp. 173-227) の REX は，その実験のひとつで，回帰分析に限定し，S システムの上に LISP で書いた推論エンジンを乗せたものである．strategy を作るのに，回帰分析の 6 つの事例からネットワークを作り，それを 36 の事例で試して修正している．この解析例から strategy を作り修正する部分も，システムに組み入れた Student システムは，いわば学習機能をもったエキスパート・システムであり，これによりそれぞれの現場に合わせたシステムの構築が可能である，と主張している．

同じように P.J. Cowley and M.A. Whiting (1985, pp. 121-127) は，データ解析例のロギングにもとづいた知識の集積を試みて，その結果〝失敗に終わった（dead end）解析の経路〟が重要な知識源であることを強調している．

以上のような知識の定式化，形式化が進むとして，その構築の材料となる解析事例の蓄積が前提条件であり，もっとも時間と手間のかかる部分である．EJDA の役割はこの部分を担うものである．現在の日本で，コンサルティング・システムが強力でない点を補うことにもなろう．学術雑誌の形式をとることにより，

(a) heuristic として，文章だけで記述されているものも含めることができる．

(b) 同じデータと目的にたいし，複数個の path があり得るし，失敗の path もあり得る．これらを，まとまった形に定式化することを準備することなく，自然な蓄積に委ねることができる．

(c) Gale も指摘しているように，複数の専門家の知識を併合すること，異なる分野での知識の共通部分を抽象化することなど自動化が全く期待できないことは，雑誌の上での議論により初めて可能である．

(d) 知識の質のある程度の保障も，雑誌の形態で初めて可能である．

5. 標準化の可能性

もちろん最大の障害は標準化である．各メーカーがそれぞれの機器，オペレーティング・システム，応用ソフトウェアを競って開発し，ソフトウェア，諸媒体上のファイルについての互換性はなかなか望めない．互換性のあるのは，RS-232C モデムにより送受信する ASCII テキ

スト・ファイルぐらいであろうか.

人間の生み出す文化が多様であり，地域ごと，言語圏ごと，宗教ごとに異質の芸術があるように，歴史と発展速度の違う技術が混在するときに様式の標準化には当然限度がある．利用者側としては，自分の目的にあった，できるだけ広く使われているものを若干個選ぶことしかできない．技術的理由よりも経済的，政治的理由で選択を左右されることも多い．

EJDA では，著者ができるだけ多くの読者を期待して形式の選択に留意する．編集者はできるだけ著者と読者の要求に応じるように努力する．つまり，ある程度世の中の流行に従うことは仕方ないし，流行しているいくつかのものの間の変換はある程度可能であろう，という楽観に立たざるをえない．ワードプロセッサの機能は制限して使い，数式はできるだけ簡単に書くことが必要であろう．計算言語はもちろん，もっとも標準的な仕様に従う．図は，タイプライターの記号で描けるものに精度を落とすか，論文から切り離してファクシミリで送受信する，とかの制限も考えられる．

雑誌そのもの，つまり論文，データ，ソフトウェアの記述の標準化に移ろう．より具体的に言えば，投稿規定の作成方針である．応用研究を中心に考えるならば，規定の中で本質的なのは，"データ構造の記述" である（計算化学で分子構造式の文字綴りによる表現が重要課題であることと類似している）．解析法やソフトウェアの分類，検索も，どのような構造のデータについての解析であり，計算であるかが主要な手懸かりになる．もちろん方法論での分類や鍵語句での検索が可能な範囲では，これらを用いるのが簡便であるが，大規模になったときの的確な検索は，データ構造を主対象としなければならない．

知識の蓄積のためにも，検索のためにも，データ構造の記述は形式化する必要がある．その方法論は統計学と計算機科学の接点の主要課題のひとつであり，本構想の科学的内容の部分である．当面は簡単のために，多重配列と，関係形式（n 項述語列（predicate set），あるいは記録（record）の集まりであるファイル，とみなしてもよい）を中心に考える．

さて論文は，データとその記述，解析，ソフトウェアと記録の 4 部門より成る．最初の 3 部門すべてが備わっている必要はない．第 1 部門を他論文の同じ部門の引用ですますこともできる．部門の概略は付録 1 の通りである．解析手続の記述は，計算機言語によることになろう．

データ部門は，言葉による記述，形式的な記述，データ本体より成る．詳細の案は付録 2 の通りであるが，もちろん検討すべき点が数多く残っている．データの実質科学的な意味は言葉により説明するしかないが，統計解析に直接利用される情報はできるだけ形式化したい．

ソフトウェアの発表形式については，いくつかの雑誌における発表様式があり，それを参考にして定める．典型的なものとして ACM Transactions on Mathematical Software (TOMS) 誌がある．統計計算については，Applied Statistics 誌での経験が Griffiths and Hill (1985) によりまとめられている．これと EJDA の違いは，単なる試験プログラム（test driver）だけでなく実験的な規模のデータと計算結果が別項目に備わっていることである．

解析部門の叙述は，通常の論文の形式となるが，数式の表現の標準化が必要である．ハードコピーを作るときは美しく出力したい．そのための制御記号はシステムごとに異なっているが，ポストスクリプト言語のように，どのシステムからでも変換できる中間言語が普及すれば，それほど大きな問題ではないと考えている．

最後に，独立した節で議論すべきほどの内容を含むが，運用上の問題を注意しておく．著作権は当然，印刷雑誌と同様に考えられるべきである．投稿から採択までの審査期間には，論文の独創性を重んじるための安全策が必要である．"一般購読"，"ハードコピー作成" にたいする課金の問題なども解決しなければならない．

6. 当面の計画

以上述べた構想は本年度(1987-1988)より，統計数理研究所の共同研究計画のひとつとして実験を始めている[課題番号 62-共研-6]．現在の研究所で利用可能なシステムに基づくために限定されたものとなり，アクセス可能な参加者もそれによって限定される．その意味で閉鎖的なシステムとなるが，アクセス可能な人すべてにニュース，事例などを公開する．

諸兄姉の支持と援助を得て，研究費が増すならば，来年度以降はこれを全国主要大学と結び，複数個のシステムの下でも稼動できるようにしたい．既存，新設のネットワークに参加し，利用することになる．実際に，投稿，審査，採択の過程を試したい．これを1，2年続ければ雑誌の形態についての目処(めど)が得られるであろう．

当面参加しないし，参加できそうにない，と思う多くの読者にお願いしたい．パソコンを購入するよりワークステーションに注目していただきたい．計算センターが核となって計算機ネットワークが作られる時には積極的に参加して，何ができるのか，できないのかに注意し，議論に参加していただきたい．

ひとつの研究所にいろいろの機能が集中することは弊害を伴うが，国立大学共同利用機関としての充実と発展のためには，このような構想が貢献すると期待している．統計諸学会が現在多過ぎる情況にあるが，このような新形態のものが連合誌として協力，交流の役割を果たせることも希望している．

7. あとがき

本稿は上述の共同研究の準備段階での諸議論に基づいている．これらの議論から多くの刺激を受けたことを感謝している．計画に予想される困難の指摘はもっとも貴重であるので，早期の厳しい批判をお願いしたい．

参 考 文 献

Andrews, D.F. and Herzberg, A.M. (1985). *Data; A Collection of Problems from Many Fields for the Student and Research Worker* (Springer Series in Statistics), Springer-Verlag.

Billard, L., ed. (1985). *Computer Science and Statistics; Proceedings of the Sixteenth Symposium on the Interface*, North Holland.

Cox, D.R. and Snell, E.J. (1981). *Applied Statistics; Principles and Examples*, Chapman and Hall (医学統計研究会訳，応用統計実践教本，1985，MPC).

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems, *Ann. Eugenics.*, **7**, II, 179-188.

Gale, W.A., ed. (1986). *Artificial Intelligence and Statistics*, Addison-Wesley.

Griffiths, P. and Hill, I.D. (eds.) (1985). *Applied Statistics Algorithms, Royal Statist. Soc.*, Ellis Horwood Ltd., Chichester, England.

森口繁一 他 (1976). オペレーションズ・リサーチのためのデータとプログラムに関する研究，日本オペレーションズ・リサーチ学会，報文シリーズ，**T-76-1**.

奥野忠一 他 (1986). 工業における多変量データの解析，日科技連出版社.

ランカスター著，植村俊亮訳 (1984). 紙なし情報システム，共立出版(原著，1978，Academic Press).

〔付録 1〕

1. データ部門（データの加工段階の違いにより複数の形式で構成されることもある）
   叙述（Narrative description）
   形式的記述（Formal description, Meta data）
   データ本体（Data itself）
2. 解析部門
   叙述（Narrative description）
   図，表
   参考文献
   手続きの形式的記述（Formal description）
3. ソフトウェア部門
   叙述
   形式的記述
   プログラム
4. 記録部門（Logging）
   投稿日，審査過程，変更，登録番号　など

〔付録 2〕

1. 付録 1 のデータ部門の詳細
   叙述
     登録番号
     表題，著者
     テキスト
   記述
     登録番号
     データ形式＝配列｜関係形式（フラット・ファイル）
     軸情報
     補助情報
   データ
     次元ベクトル＝各軸の大きさより成る整数ベクトル
     添字ベクトル＝空｜整数ベクトル
     要素ベクトル＝添字ベクトルが空なら均等配列，空でなければ添字ベクトルに従った順序に並ぶ．関係形式のときは全て文字型
2. 上記 1 の軸情報の詳細
   軸名
   軸識別子（Axis identifier）＝空｜順序番号｜群，水準番号｜時間　など
   軸ラベル＝軸識別子に対応した文字別ベクトル
   値 の 型＝整数｜実数｜文字；範囲；精度；単位；確率的｜非確率的