探索的データ解析法の考え方

美添 泰人（青山学院大学経済学部）

1　はじめに

今回のエストレーラでは探索的データ解析法 (Exploratory Data Analysis, 略して EDA) の特集が企画されているという．最近はこの分野も常識的になりつつあるので，具体的な手法については他の記述に任せることとして，ここでは「なぜ，探索的データ解析なのか」という基本的な問題を検討してみたい．

探索的データ解析として知られている手法の大部分は，1960年代から1980年代にかけて，当時プリンストン大学統計学部の教授であり，AT&Tベル研究所の主任研究員でもあったチューキー (J. W. Tukey) を中心として開発されたものである．主要な手法を解説したものとして，1970年代にタイプ原稿を製本したテキストが主としてプリンストンとハーバード周辺で出まわっていたものに，Addison-Wesley, 1970と記されていた．その後の正式出版の前に教材として作成されたものと聞いている．最終的に1977年に市販本の形で出版されたのが，オレンジ色の表紙で有名になった EDA というタイトルの本[9]である．同時に，EDAの手法のうち，主として回帰分析における諸問題を扱った薄緑色の本[6]も出版された．この本の著者はチューキーと，当時ハーバード大学統計学部の教授であったモステラー (F. Mosteller) である．

筆者は，出版当時は大学院生としてハーバード大学に在籍しており，月に1,2回開催されるモステラーとチューキーのセミナーで，具体的なデータの分析について，この分野における主要な貢献者であるホーグリン (D.C. Hoaglin) などとの議論を聞いていた．当時の筆者の拙い英語力でどこまで理解できたかはさておき，大多数の大学院生にとって，統計的データ解析における本質的な考え方を学べる貴重な場所だったと思われる．

オレンジ色の本は，極めて難解であった．筆者は，この本のタイプ印刷版を見ていたし，出版された直後の1977年の秋学期に，モステラーの担当するデータ解析の講義を履修した．そこでは，緑色の本を教科書，オレンジ色の本を参考書として読みながら他の学生とともに質問を重ねたにも関らず，オレンジ色の本については難しいという感想は変わらなかった．記されている手法そのものは，実は，簡単である．難しいのは，何故，そのような手法が必要なのかという点と，そもそも英語が難解で，どこまでが本気で，どこからが冗談なのか分からないような筆致なのである．アメリカ人の友人たちも難しいと言っていたのだから，これは筆者の英語力の問題だけではなさそうである．

EDAの手法については，読者もすぐに気づかれるように，その用語が独特であり，学術用語というより，日常用語に近く，半分冗談ではないかと思われるような命名がなされている．EDAの本では本文全体がその調子なのだから，読んでいて疲れることは間違いない．

なお，難解さについての上記の判断が私だけのものではない証拠に，1977年の記念碑的な「オレンジ色の本」はアメリカですら広く受け入れられなかったという評価がなされている．その後の変遷について，父親がチューキーの親友で，子供のころからチューキーのひざの上で可愛がられたという統計学者（現在，カーネギー・メロン大学の教授である）が，チューキー自身から聞いたと，筆者に話してくれたことがある．それによると，チューキーはEDAの本が出たのだから，当然その内容が理解されるだろうと考えていたところ，その後数年たって振り返ってみると，ほ

とんど誰も後をついてきていないことに気づいた．そこで，1980年代になってからいろいろな形で易しく書いた本 [2, 3, 4] を出版したところ，今度は，大勢の読者の理解を得ることができたというのである．なお，具体的な手法を身につけるだけなら “ABC of EDA” という魅力的な書名をつけた [10] も，比較的早い時期に出版されている．

ちなみに，チューキーは初めは数学を専攻したそうで，集合論における基本的な命題である「ツォルンの補題」の変形として知られるチューキーの命題は，大学院生のときの仕事とのことである．ついでながら，次の形である．

定理　C を集合 X の部分集合に関する有限的な条件とし，C を満たすような X の部分集合 Y が少なくとも 1 つ存在するとする．そのとき，C を満たす X の（包含関係の意味で）極大な部分集合が存在する．

大学院時代に化学に転向し，多数の実験データを解析しながら統計的データ解析を専門とするようになったチューキーは，1950年代にはさまざまな数理統計的手法の論文を著している．そのチューキーが「数理統計学は役に立たない」として，新しくデータ解析という分野を研究の主題にすると宣言したのが，早くも1962年に発表された “The future of data analysis” という論文[8]である．この原稿は資料を確認しながら書いているわけではないので，正確に歴史を表していない可能性はあるが，データ解析という用語が用いられたのは，恐らくこの論文が最初ではないかと思われる．

今から考えれば，これが探索的データ解析の必要性を説いた画期的論文として位置づけられるのだが，筆者がこの論文を読んだ1972年頃のわが国で，そのような評価を聞いた記憶がない．大多数の人は，統計的手法のうち，後にロバスト統計学と呼ばれるようになる手法を記述した論文とだけ受け止めていたのではないかと想像している．

筆者のデータ解析の理解は，チューキーから直接学んだというより，モステラーから教わったり，ロバスト統計学を学んだ際に関連して身につけた部分が多いが，次節以降で本来の問題意識を紹介してみたい．

2　数理統計学誕生の頃

現在の統計学は，1920年代にフィッシャー (R.A. Fisher) によって，その基礎が与えられたというのが，一般に受け入れられている見方である．そこで，まず一つの例を紹介しておきたい．

分布の散らばりを計る尺度として，現在は標準偏差

$$s_n = \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{\frac{1}{2}}$$

が広く用いられている．しかし，今世紀のはじめには天文学者のエディントン (Eddington) との間で平均偏差

$$d_n = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

と標準偏差のいずれがより優れた尺度であるかについての議論があったと言う．論争に決着をつけたのがフィッシャーであり，その議論は次のようなものであった．

データ $(x_1, \ldots, x_n)$ を正規分布 $N(\mu, \sigma^2)$ からの無作為標本と仮定する．n が大きいとき，s_n は真の値 σ に近づき，一方 d_n は $\sqrt{2/\pi}\,\sigma \doteqdot 0.80\sigma$ に近づく．さらに，これらの極限の近くでは 2つの尺度は近似的に正規分布に従い，その分散はいずれも n の逆数に比例することが確かめられる．これは大数の法則と中心極限定理から導かれる結果である．そこで d_n の代わりに $d'_n = \sqrt{\pi/2}\, d_n$ を s_n と比較して，同じ n に対してどちらの推定量の分散が小さいかを調べることにする．実際に分散の比を理論的に求めることは可能であり，それによると

$$ARE = \lim_{n \to \infty} \frac{\mathrm{var}(s_n)}{\mathrm{var}(d'_n)} \doteqdot 0.876$$

が得られる．n が大きいときの分散の比は「漸近相対効率（Asymptotic Relative Efficiency）」と呼ばれるが，$ARE = 0.876$ ということは，d_n の代わりに s_n を用いれば d_n を用いるときの標本の 87.6% の大きさで同程度の精度を達成できる，ということを意味する．従って平均偏差よりも標準偏差の方が優れていることになる．

以上の議論は 1920 年頃のことであるが，このような議論によって初めて精密な議論が可能になったと言われるものである．実際，この例に見られるような「厳密な仮定の下での最適性」の議論は，統計学を精密科学に高めたと評価され，最近までの数理統計学の主流であった．同じ頃スチューデントとフィッシャーによって導入された t 分布も，似たような考え方に基づいている．すなわち，観測値の分布が正規分布であることを前提として，母集団平均に関する厳密な推論を導いているのである．このような視点から t 分布の発見の歴史的な価値を評価するのがモステラーであるが，詳細をここで紹介する余裕がない．ただし，このような視点から t 分布の真の意義を理解している統計学者は，数が少ないように見える．

フィッシャーのような議論がなされる前の統計学においては，大量のデータが集められることを前提としている．その分析は大数の法則と中心極限定理の応用によって理解できる，というのがモステラーの解説である．そうすると，フィッシャー以前と以降の決定的な差は，データの分布について厳密な仮定をおき，最適性に基づく議論を行うかどうかにあると言っても過言ではない．

もちろんフィッシャー自身は正規分布の仮定が成立するように実験をコントロールして理想的なデータを得ようとしたのであり，それが実験計画法と呼ばれる分野となっている．なお，次の節以降で解説する EDA ないしその発展した形態とも考えられるロバスト統計学の立場から言うと，このような意味で実験計画法を理解しているのは応用統計学者が中心であり，有限群論にもとづく美しい理論に熱中している数理統計学者の問題意識は低いということになる．

3　正規分布の仮定の意味

ところで，d_n と s_n の比較は 1960年代になってもう一度取り上げられた．チューキー[7]は，現実にわれわれが分析の対象とするデータは，たとえ十分に管理されたものであっても，厳密な正規分布とはいいがたいことを認めて，一つの近似として混合正規分布

$$(1-\varepsilon)N(\mu,\sigma^2)+\varepsilon N(\mu,k^2\sigma^2)$$

を想定した $(0<\varepsilon<1,\ k>1)$．この分布ではデータのうち $1-\varepsilon$ の割合で質のいい観測値が得られ，ε の割合で（未熟な技術者の測定のように）多少精度の低い観測値が混ざっている，という状態が表現される．チューキーによれば，自然科学における良質なデータは $\varepsilon = 0.01 \sim 0.1,\ k = 3$ 程度の混合正規分布で近似されるという．この論文では混合正規分布の仮定のもとで，さまざまな統計量の性質を論じているが，先程の例に戻ってみよう．いま ε の値を変えて，改めて s_n と d'_n の ARE を計算してみると，

$$ARE(\varepsilon)=\frac{\big(3(1+80\varepsilon)(1+8\varepsilon)^{-2}-1\big)/4}{\pi(1+8\varepsilon)(1+2\varepsilon)^{-2}/2-1}$$

と，展示 1 のような結果が得られる．ちなみに EDA の流儀では図や表の代りに，すべて「展示 (exhibit)」として統一番号をつける．展示 1によれば，1000 個の観測値のうち，たった2 個の観測値の精度が低いだけで，s_n の優位は失われてしまい，また $\varepsilon = 0.05$ の近くで ARE は 2を超える最大値をとる．これから，現実的な ε の値に対しては s_n は d_n に比較すると，かなり性能の劣る統計量であるということがわかる．

この例から学ぶべき最も重要なことは，厳密なモデルの下での最適性は，モデルがわずかに変化しただけで失われることがある，と

いう点である．われわれが正規分布にもとづいた議論をするときでも，その根拠は「現実のデータが従う分布が正規分布で近似される」ということであって，理論の想定と現実の分布とに多少の違いは認める必要がある．そうすると，仮定するモデルがわずかに違うときに結論が大きく変化するような手法は好ましくないことになる．実際，混合正規分布の下で標準偏差の効率が小さいのは，わずかな外れ値に対して s_n が大きく変化するからである．なお，現在のロバスト統計学の議論によれば，d_n は決して優れた統計量ではなく，他にもいい性質を持った推定量が存在することを注意しておきたい．

それまでの数理統計学，特に多変量正規分布から導かれる分布論などは厳密な仮定にもとづく議論が中心であり，実際のデータ解析にはほとんど役に立たないということになる．現実的なデータ解析について豊富な経験を持つチューキーが数理統計学を否定して，新たにデータ解析という名の下に EDA の手法の研究を始めたのは，このような背景があった．

初期の EDA では，データの洗練 (data cleaning) とグラフによる手法，外れ値の影響を受けにくい手法の開発が中心であった．このうち，データクリーニングは，次のような考え方によるものである．

先の標準偏差の例では，極めて少数の悪い観測値が混入することによって，正規分布の下での優れた性質が失われてしまった．その主要な原因は，極端に大きいか小さい観測値が頻度が少ないとはいえ，発生することにある．したがって，正規分布との違いとして注目すべきなのは，このような意味での外れ値 (outliers) と言える．データクリーニングでは，何らかの意味で他の観測値の従う法則に従わない外れ値を発見して除去することを考える．外れ値が適切に除去されたデータセットであれば，標準的な手法を利用することができるだろうということである．なお，現在のロバスト統計学の視点では，もう少し問題設定の範囲が広い．すなわち，データクリーニングのような処理の基本的考え方は正しいとしても，外れ値の除去方法を論ずるだけでは不十分であり，外れ値であることが疑われるような観測値をどのように扱うかも問題とされる．

このように，数理統計学は役に立たない学問として否定されるべきであると主張された時期があった．ただし，一時のプリンストンの統計学部はその方向に行き過ぎたと批判されたこともある．たとえば，1970年代後半の時期には，プリンストンから統計学の博士号を貰った学生には2変量正規分布の相関係数の分布（母集団が独立の場合）が導けないという逸話もあった程である．

実際は，チューキーが否定した数理統計学は，狭い意味の数理統計学と考えるべきであり，その真意については次節で明らかにする．

4　ロバスト統計学との融合

1970年代に入ってから，ロバスト統計学と呼ばれる分野で大きな進展が見られた．この分野における最初の体系的な本[5]を著したフーバー (P.J. Huber) によれば，ロバスト統計学の基本的な視点は，モデルとして想定している仮定のすべての妥当性を疑い，モデルが厳密に成立しない場合の統計的手法の安定性，有効性を検証することである．たとえば，母集団の位置の尺度（平均やメディアンなど）を推定する問題では，観測値の分布として正規分布を想定した場合の最適性が，正規分布とわずかに異なる分布でどのように変化するかが扱われる．同様に，回帰分析では回帰式が線形でない場合に，線形性の下で導かれた推定方法ないし予測方法の最適性がどう変化するかも扱われる．

すぐに分かるように，このように整理した形では，ロバスト統計学の目指すところは，究極的にはEDAと同一である．

ところで，ロバスト統計学の考え方は，数理統計学の否定ではなく，その新たな拡張で

ある．標準偏差の例では，純粋な正規分布を拡張して混合正規分布を想定した．外れ値は正規分布ではほとんど生じないが，混合正規分布ではときどき発生することになる．このようにモデルの仮定を緩めて，より広い範囲のモデルを考えたとき，各種の推定方法の比較を行なおうというのがロバスト統計学の基本的な考え方である．なお，混合正規分布だけでは，正規分布に近い分布がすべて表現されているわけではない．ロバスト統計学ではさらに広い範囲の，データ解析にとって意味のある現実的な分布を想定する．

たとえば，位置の尺度として算術平均は有用だし，平均所得などは平等に分配することを考えれば実際にも十分意味のあるものである．しかし，算術平均は外れ値の影響をきわめて受けやすいという性質を持っている．心理学や社会科学における応用では，かわりの尺度としてメディアンなどがよく用いられるが，その理由の一つには，メディアンが外れ値の影響を受けにくいというロバスト統計学からの性質があげられる．

統計的なデータ解析における最も重要な視点は，モデルは現実の近似に過ぎないということを十分に理解することである．チューキーが否定した数理統計学は，このような基本的な視点を理解せず，形式的数理を展開する数理統計学である．EDAが受け入れられる前の時点では，このような批判がなされて当然と思われるような論文が，多数の専門紙に掲載されていたのが実情である．

一方で，チューキーが，ロバスト統計学のように整理された形の新たな数理統計学に反対していないことは，1980年代にまとめられた本[2]の表題に「ロバスト」という表現が用いられていることや，そもそもロバスト統計学の初期の研究の集大成である[1]の著者の中にフーバーとともにチューキーの名があることからも明らかである．

最後にまた個人的な経験を記すと，フーバーの著書[5]は，フーバーがハーバード大学に客員教授として滞在した1977年から78年の間に執筆され，その手書きの草稿にもとづいて講義が行われた．素晴らしい内容の一方，最初は20名以上いた受講生が次第に減っていき，最終的に単位を取得したのは筆者の他にもう１名だけ，という難しい講義でもあった．筆者の博士論文は，このような経緯からフーバーを指導教授として，ロバスト統計学とベイジアン統計学の両方に関わるものとなった．EDAとともに，これらは筆者の「統計的なものの見方」の原点である．

この短いスペースでロバスト統計学の手法について解説することはできないが，統計的データ解析の基本的な視点として，近似的なモデルという理解が決定的に重要であること，そしてこの視点こそ EDAを理解するために必須であることを強調しておきたい．

参考文献

[1] Andrews, D.F., P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers, and J.W. Tukey (1972) *Robust Estimates of Location*, Princeton U.P.

[2] Hoaglin, D.C., F. Mosteller, and J.W. Tukey (1983) *Understanding Robust and Exploratory Data Analysis*, Wiley

[3] Hoaglin, D.C., F. Mosteller, and J.W. Tukey (1985) *Exploring Data Tables, Trends, and Shapes*, Wiley

[4] Hoaglin, D.C., F. Mosteller, and J.W. Tukey (1991) *Fundamentals of Exploratory Analysis of Variance*, Wiley

[5] Huber, P. J. (1981) *Robust Statistics*, Wiley

[6] Mosteller, F. and J.W. Tukey (1977) *Data Analysis and Regression*, Addison-Wesley

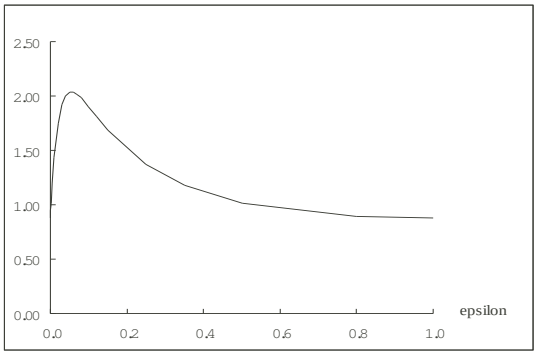
[7] Tukey, J.W. (1961) “A Survey of Sampling from Contaminated Distributions,” in I. Olkin *et al.* (eds.), *Contributions to Probability and Statistics — Essays in Honor of Harold Hotelling*, 448–485.

[8] Tukey, J.W. (1962) “The future of data analysis,” *Annals of Mathematical Statistics*, vol. 33.

[9] Tukey, J.W. (1977) *Exploratory Data Analysis*, Addison-Wesley

[10] Velleman, P.F. and D.C. Hoaglin, (1981) *Applications, Basics, and Computing of Exploratory Data Analysis*, Duxbury Press

展示 1. s_n と d'_n の漸近相対効率



1999年7月
Estrela 原稿