文献紹介

―― 私の研究方法から ――

探索的データ解析の方法

―― 教育行政学研究に対する数量的アプローチのひとつとして ――

大学入試センター　池　田　輝　政

はじめに

教育行政学が社会科学として発展していく上で、統計学の分野で確立されてきた数量的アプローチを有効に活用する必要がある。数量的アプローチの有効な活用法として次の2点が考えられる。

(1) ある仮説に基づき収集したデータを、その仮説の検証のために解析する。

(2) 所与のデータについて、それが意味するものを探り事実の説明に役立つ理論を発見するために解析する。

数量的アプローチと言っても目的に応じた解析の方法論があるので、われわれユーザーはそのことに十分注意する必要がある。前者(1)の目的に応じた解析の方法論は既に普及をみているが、後者(2)の目的に応じた解析の方法論についてはそれ程知られていないと思われる。教育行政学の研究にとっては、この方法論を知ることが数量的アプローチを有効なものにすると考え、本欄で紹介することにした。

1. 探索的データ解析の方法とは

Exploratory Data Analysis　John W. Tukey (Addison-Wesley, 1977)

本書はその標題の頭文字をとってＥＤＡ（エダ）と略称されている。しかも、ＥＤＡの略称はその書名を表わすだけでなく、統計学における1つのスクール名（データ解析学派とも称する）をも同時に表わすものとして使われている。

従来の統計学の教科書には、観測・収集した標本のデータを記述し、その標本のデータからもとの母集団データ（母数）を、一定の精度でもって推定するときの原理と手続が叙述される。ＥＤＡでは、この仮説検定を主とする統計データ解析を確定的データ解析 (Confirmatory Data Analysis) と呼ぶ。

確定的データ解析の方法論は、今世紀の知的活動の所産として現在では不可欠のものとなっている。とくに、理論的体系化が進み、分析される変数がその理論により合理的な方法で選択されるような領域では、仮説検定の方法は有効性を発揮する。しかし、そうした段階に達していない領域では、仮説検定のレベルに至る前にデータをよく検討してデータの語るものを探究する作業が要求される。

標本のデータがある程度規模の大きい計数データ (counted data) や、計量データ (measured data) の場合、その要約・記述には度数分布表や平均値・分散などの統計量が普通に利用される。そ

して、これらの方法を利用して標本データの分析を行う訳であるが、確定的データ解析においては標本の要約・記述の情報が仮説の検定に直ちに適用される傾向になり、分析のための要約・記述というよりも単なる要約・記述に終わる場合が多い。

母集団からの標本という性格にこだわりすぎると、データのもつ情報を読みとることに失敗し、その結果はデータの一面的な活用とひいては一面的な理解につながる場合がある。これはデータの要約・記述のレベルにおける確定的データ解析の限界とみることができる。なぜなら、確定的な性質のもの以外はものを言わないという前提が基本的な考えとしてあるからである。

社会科学の中でも教育行政学は、理論的サポートをもった仮説をたて、それを統計的手続に則って検証するために、データを収集し記述・推定し検定を行うという形のデータ解析を忠実に実行することが困難な領域である。従って、収集データを標本データとして扱うことから一旦離れて、データの語りうる色々の可能な情報を探るという哲学と方法論に立脚したデータ解析が、こうした領域では必要かつ有効である。

ＥＤＡすなわち探索的データ解析は、上述の課題に応えるために従前のデータ解析の技法を再編・結合し、内的一貫性をもたせた１つの方法論であると同時にデータ解析の哲学でもある。

著者のトゥーキイは、米国のプリンストン大学の統計学の教授であり、統計学の発展に大きな寄与をしているベル電話研究所の副所長をも兼任する、この分野での権威である。彼のデータ解析に関する哲学は、この本の序文に述べられている。

「この本は探索的なデータ解析、すなわちデータが語るものを理解するためにデータを吟味することを述べたものである。 このため簡単な数式と容易な作図を強調する。そして、 我々が知り得たものは全て部分的なものであると考え、さらに新しい洞察を得るためにそれら表層の下にあるものを探るよう試みる。つまり、その関心は確定(confirmation)にではなくて発見 (appearance)にある」（ p. Ⅴ ）

簡単な数式（中学校レベルの数学的知識で充分）と容易な作図を強調するその背景には、データに関する基本問題が（ p. Ⅴ ）、

(1) 記述をより単純化することがそれを容易に操作可能にする。

(2) 既に記述したものを更に深めることがその記述を更に有効なものにする。

の２点にあると考えられている。つまり(1)描写を単純化することと、(2)１つの層を更に深く記述することがデータ解析の基本姿勢になっている。

この基本姿勢は本書の至る所で貫ぬかれている。しかも、この姿勢を一貫していくために独特な解析のテクニックを導入しその展開をはかるのだが、そのためには解析の意味について考えること（例えば「こんな解析を行ってどのような意味があるのかというような問」など）を一旦停止することを要求している。極言すれば、研究目的にとらわれず眼前のデータを単純化し、それを入口にしてデータを更に深く解析していくという手続に徹底することを説いていると考えられる。

2. 各章の内容

全編の構成は２１章と巻末の用語解説からなっているが、第１章から第６章までがデータ解析の手続に相当する基本的テクニックを説明している。また第１０章から第１２章は、いわゆるクロス表形式のデータに関する解析テクニックが触れてあり、活用可能性の観点からは特殊なテクニックが述べてある他の章に比較して重要度が高い。

文献紹介という目的から判断すれば、上記の各章を要約することで役目は終わるのであろうが、本書の中味は実際のデータ例に基づく解析テクニックの説明・展開であるので、言葉だけでは要約しきれないところがある。また各章にわたって全てを要約するのは筆者の力量の範囲を出るものである。従って今回は本書の基本的テクニックが述べられている第１章から第６章に要約を限定し、その後とくに第１、２章のテクニックを使って実際のデータに適用した例を示して紹介責務を果したい。

第１章 ： 冒頭の部分には犯罪の捜査にたとえて、探索的データ解析と確定的データ解析の役割分担の関係が述べられている。探索的データ解析の仕事は手掛りを発見しそれを明らかにすることであり、確定的データ解析ではこれを更に進めてそれを証拠として評価することになる。

その他に数字や記号に関すり取り決めが導入部分に述べられているが、この章の中心的事項は「幹葉表示」(the stem-and-leaf display) の説明にある。これは分布の情報を得るために、あるデータの集まり（バッチ batch と表現され、同じ意味をもった数値のまとまりをさすが、必ずしも標本のデータをさすものではない）を大小の順序に並べかえて図示する方法である。分布の情報としては(i)多峰性(separation) (ii)非対称性(asymmetry) (iii)はずれ値(irregularities) (iv)中心(centering) (v)広がり(width)をみることである。

幹葉表示は通常よく利用されるヒストグラムに相当するものであり、ＥＤＡの解析テクニックの出発点に位置する。その例を下に示すが、ａの欄はあるデータ値の集まり(a batch of data)であり、ｂおよびｃがそれに基づく幹葉表示の例である。

ｂの幹葉表示は簡便 (conventional) タイプで、ｃの幹葉表示は数字保存 (digit-reminder) タイプで通常はこれが使用される。縦のラインが幹に、そして横のラインが葉に各々相当する。

ｃの幹葉表示の分布に関する情報としては、(i)単峰性であり、(ii)非対称形で高い数値の方向に漸減しており、(iii)１６６、１７７、１７７の３つの数値ははずれ値と考えられ、(iv)中心は７７の数値あたりに位置し、(v)広がりはかなり大きい、などが認識できよう。

a. a batch of data

N = 18 : 33, 33, 44, 44, 55, 55, 66, 66, 77,
77, 77, 77, 77, 99, 111, 166, 177, 177

b. 簡便タイプ

3	××
4	××
5	××
6	××
7	×××××
8	
9	×
10	
11	×
12	
13	
14	
15	
16	×
17	××

c. 数字保存タイプ

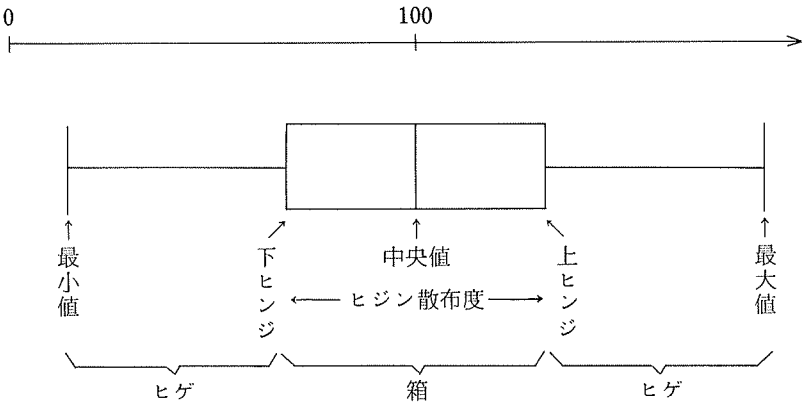
3	33
4	44
5	55
6	66
7	77777
8	
9	9
10	
11	1
12	
13	
14	
15	
16	6
17	77

(Tukey, J. W. p. 7)

第２章 ： 先の幹葉表示に基づき、データの分布に関する行動あるいは型 (behavior or pattern) を要約する手続が述べられる。要約には数値に基づくタイプと視覚に訴えるタイプがある。

数値要約は５つの要約数 (a 5-number summary) が基本である。中心の位置を示す代表値に(i)中央値（メディアン）、ほぼ1／4分位数および3／4分位数に相当する (ii) 下ヒンジおよび(iii)上ヒンジ、そしてextremesと呼ばれる (iv)最小値、(v)最大値が５つの要約数である。

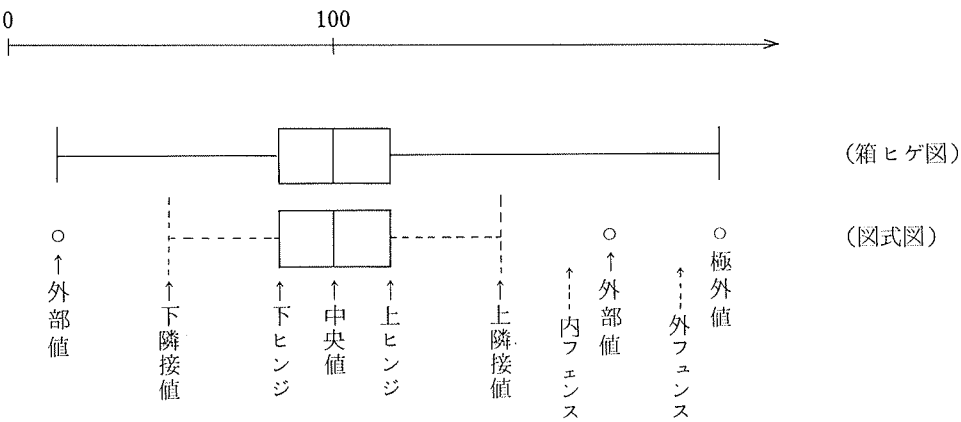
視覚に訴える図示要約は箱ヒゲ図 (a box-and-whisker plot) が基本である。箱ヒゲ図の作図要領は簡単で先述の５つの要約数を用いて下記のような形の図を描く。



上記の図中、箱ヒゲの箱の部分は上ヒンジと下ヒンジの各々の値と中央値から構成され、上ヒンジと下ヒンジの差はヒンジ散布度 (H-spread) と称される。ヒンジ散布度は理論的にも実際にもデータのほぼ５０％を含む広がりの測度である。

図示要約はＥＤＡの解析テクニックの大きな特色の１つである。先の箱ヒゲ図が分布の全体的な行動に注目するために活用されるのに対し、分布の異常な行動に注目するために活用されるのが図式図 (a schematic plot) と呼ばれるものである。

図式図といっても箱ヒゲを作図するのであるから、箱ヒゲ図のバリエーションの１つにすぎないが、データの分布の裾の行動を強調して作図する点に特色がある。その作図の例を箱ヒゲ図と対照させて描いておく。



上側が箱ヒゲ図の箱ヒゲ、下側が図式図の箱ヒゲとはずれ値（外部値および極外値）である。上下の図を比較してみると、箱ヒゲ図の中の「ヒゲ」の部分が強調されて図式図が出来上っているのがわかるであろう。

データの分布の行動において正常な(usual)部分と異常な(unusual)部分を分けるために、図式図の上下の隣接値 (adjacent values) の範囲内を正常な部分とみる。隣接値は（ヒンジ散布度）×1.5を上下の各ヒンジに加え、その値に最も近い中心側のデータ値である。従って、残りの分布の裾にあたる部分、すなわち外部値 (outer vaues) および極外値 (far out values) を異常な分布の行動とみなす。これらは一括して、はずれ値 (unusual values) と呼ぶ。

はずれ値の見分け方は、（ヒンジ散布度）×1.5＝1ステップとおき、それを上下の各ヒンジに加えた点を内フェンス、これに更にもう1ステップを加えた点を外フェンスとすれば、外部値は内フェンスと外フェンスの間にあるデータ値をさし、極外値は外フェンスより外側にあるデータ値である。

この見分け方の規則は厳密なものではなく、著者のデータ解析に関する経験的知見によるものと考えられる。探索的データ解析に厳密性を求めるのは、この解析の狙いとする「発見」という目的からしてさほど重要視されていない。

第3章 ： データの尺度を再表現 (re-expression) する方法が述べられている。この再表現のテクニックとして、$y = x^p$ という「べき乗変換」(power transformations) が採用されている。

再表現の必要性に関する最も単純な指針としては例えば、（最大値）／（最小値）の比が100を超える場合とし、通常は $p = 0$ の対数変換、$p = 1/2$ のルート変換あるいは $p = -1$ の逆数変換を推めている。ただし、この章で述べられているルールは、データの尺度が分数やパーセントあるいはカテゴリカルな順序量の場合には妥当しない。

データを再表現することもデータを要約するうえで重要な方法であると考えられている。例えば、データの分布が非対称形 (asymmetry) である場合は、対称形(symmetry)にして要約するのがより有効であるとされる。EDAにおける再表現（変換）の目的は、分布を非対称から対称にするという要約の目的に止まらず、比較などのさらに重要な目的に関係していくのである。

第4章 ： 前章までは主として単一のデータ群（単一のバッチ）を前提に解析テクニックが展開されてきた。この章では複数のデータ群を前提にしその比較という観点から効果的な解析テクニックが紹介される。

データ解析において有効な比較を行うためには (i)各データ群の分布を各々対称形に近づけること、(ii)データ群の各々の広がりを同程度にすること、を挙げている。(i)の分布の対称性よりも (ii) の広がりの一致がより重要とされている。この目的のための解析テクニックが展開されるが、比較のための最も重要な解析テクニックは後章で触れられることになる。

第5章 ： 1つの変量に関する複数のデータ群の比較という観点から目を転じて、2つの変量間の関係を探索する解析テクニックを述べた部分である。

2つの変量は (factor, response) の形で構成され factor が x 軸上に、response が y 軸上に位置づけられる。これは要するに $x-y$ 軸上に散布図を描くことである。

x (factor) に対する y (response) の関係を解析する際に、データに関する基本モデルが提示される。それは、

Data = Fit + Residual
（データ）（フィット）（残　差）

の関係式である。フィットは $x-y$ 軸上にプロットされたデータの予期される関係を示し、残差はその予期できない部分を示す。

フィットはデータのもつ regularity を表わす部分であり、残差はその irregularity を表わす部分とみることができる。 regularity の把握には、単純性を重視するため、直線による表現に徹底する。そのための解析テクニックとして、前章で展開したデータの変換（再表現）が活用されることになる。 この regularity の直線化作業（フィッティング）が効果的に行われると、次にはより深い解析の段階に進み irregularity を表わす残差の解析が行われることになる。

第６章 ： 前章の展開部分にあたり、データのフィットの部分を直線に表現するためのルールが具体例に即して紹介されている。

3. ＥＤＡの適用例

以上、ＥＤＡの基本的な解析テクニックが述べられた章について要約してきたけれども、最後に実際のデータに対してＥＤＡの図式図を用いた表示法の一例を示してみる。

使用したデータは都道府県別に整理された高等学校全日制課程の教育費総額である。データの出所は全国教育調査研究協会編「戦後３０年　学校教育統計総覧」（ぎょうせい）昭５５と文部省「地方教育費の調査報告書（昭和５３会計年度）」である。

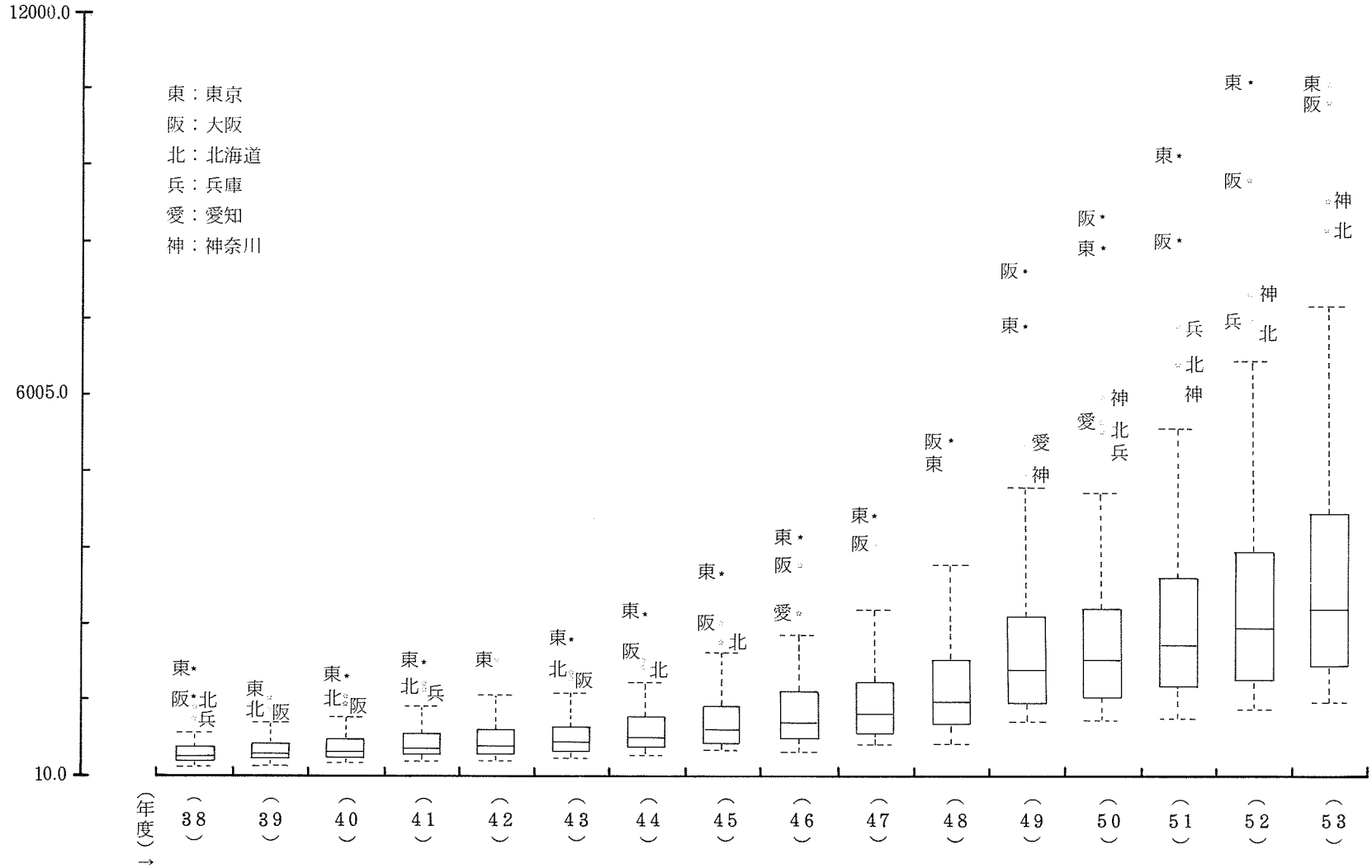
図１はこの高等学校教育費を昭和３８会計年度から昭和５３会計年度まで表示したものである。もちろん、教育費のデータは①公債（国庫補助金・都道府県支出金・市町村支出金）、②地方債、③公費に組み入れられた寄付金、④公費に組み入れられない寄付金（ＰＴＡ寄付金・その他の寄付金）という種々の財源を含む総額であり、そのデータの意味は一義的には確定しえない。適用例としては、多義的な概念のデータを扱うことは避けた方がよいが、しかし、ここでは目的が分析にはないので、 厳密に考えないことにする。

さて図１の中の３８年度における都道府県の教育費（高校教育費）の分布形をみてもらいたい。箱ヒゲの上方にある黒星・白星は他の集団からはずれた特異点であることを示している。東京・大阪・北海道・兵庫がこの特異点に該当し、他の集団にはない別の独自な要因が働いていることがわかる。黒星と白星の違いは程度の差であり、白星は部外値、黒星は極外値であって理論的分布の出現確率はこの順に小さくなる。箱ヒゲの「箱」の部分には分布全体の５０％の情報が入っている。その「箱」の中の横線が分布の中心にあたる。

図１から得られる重要な情報のひとつは、都道府県教育費の年度変動は３層に分けてみることができる点である。東京・大阪などの教育費の規模が大きい層では、年を追うに従って指数関数的な増加をみせている。次に中心付近の中間的な層は直線的な増加をみせ、さらに規模の小さい層は増加の程度が極めて小さい。こうした増加のパターンがどのような要因によるのかは、さらに深い分析を必要とする。

（単位＝10^7 円）

図1　高等学校全日制課程の教育費総額の都道府県分布に関する推移



また、分布全体の変動を追うと、昭和４８年と４９年の間で大きな飛躍があることが一目瞭然である。この両年度の間でインフレが急激に進んだためとも考えられる。

その他、大阪府の動きとして東京を追い越した昭和４８～５０年は注目に値する。また、昭和５２～５３年の東京の動きも注目できる。つまり、全体が上昇しているのに東京だけが下降しているのである。

まだ他にも分析の興味をひく点が幾つかあるが、この辺でやめることにしたい。なお図１のデータ表示の方法は唯一のものではなく、この先まだ変換などの色々な解析テクニックに基づき操作し、その結果を表示することもできるのである。そして、その手続によって新たな情報を得ることが可能になるのである。

ＥＤＡに関する日本語文献は、

① Ｆ・ハートウィグ、Ｂ・Ｅ・デアリング（柳井晴夫・高木広文訳）「探索的データ解析の方法」（朝倉書店）１９８１

があるだけである。ただし、この本はＥＤＡの忠実な解説ではなく、著者らの多少の解釈に基づきＥＤＡの手法を解説・紹介したものである。しかし、ＥＤＡの本を理解するには有用である。

なおＥＤＡとは関係ないが、社会科学の分野に属して、データ解析（あるいは統計学）に少しでも興味を抱く人に対して次の文献をお推めしたい。

② Roderick Floud, An Introduction to Quantitative Methods for Historians, Methuen, 1979 (2nd ed.)

これは書名が示す通り、歴史学の研究者に対する数量分析の手法の紹介である。歴史データを使いながらの説明であるので、興味がもてるし、難解な数式の使用もほとんどない。ただし、入門編であるので説明がくどかったり、逆に不足したりしている所がみられる。しかし、和書ではこうした形での入門書は現時点では期待できない点を考えれば、一読に値すると思われる。

（謝辞）

本編の中で提示した図１は大学入試センター研究部の鈴木規夫助手（情報処理研究部門）および山田文康助手（試験方法研究部門）の共同作成した電算機プログラムを使用して作成した。記して感謝の意を表わしたい。