目で見る統計学

大上慎吾

統計学って何？

将来どのような専攻分野を希望しているかにかかわらず（あるいははっきりとした希望がなくとも），一橋大生の多くが最初の2年間を通して統計学を勉強する機会を持ちます．それは，統計学がそもそも必修であったり，希望する専攻分野に必要であると知っていたり，たまたま友達が何人か受講するのでノートを借りるのに便利であったりするためかもしれません．では，そのように多様な目的意識を持った学生達が勉強する統計学とは，いったい何をするための学問なのでしょうか．

講義の冒頭で，あるいは教科書の第1章の中で，統計学はさまざまな表現を用いて説明されています．統計学とは，「最も広義に定義すれば，数量的データを処理する方法」であったり，「母集団に関する結論を標本から引き出す方法」であったり，また「さまざまな種類の情報を収集し，要約するための有効な方法を企画すること」であったりします [2]．実際，「統計学」という言葉でくくられるものの中には非常に多くの手法が含まれていますから，論理的に正しくしかも偏りのない説明をしようとすれば，いきおい一般的な表現になることは避けられません．統計学とはデータを分析する手法の寄せ集めだと理解しているという人も少なくないかもしれません．しかし表現するのは難しくても，そこには「統計学的思考」というものが確かに存在します．データを要約し，その結果を記述するのにも統計手法独特の「くせ」があるのです．その「くせ」を通して見てみれば，個々の統計手法は，

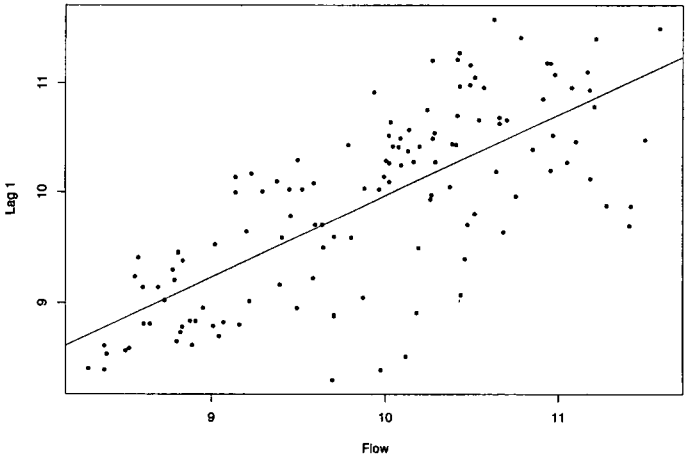
想定されている状況の違いに応じて，ある1つの思考方法が具体化したものであるというのが私の考えです．そしてその様な見方をもつことが，個別の技術的な議論にふりまわされることなく，統計学を使ってデータ分析を成功させる秘訣であると思っています．そこで，個人的な偏向はあるのですが，あえて各方面からあがるであろう反論や非難に耳をふさいで，私なりの「統計学」について以下論じていきたいと思います．

ノイズの分析？

まず統計学は数量的データを取り扱います．数として記録することのできる情報です．こうした数量的な情報を要約するために，まずデータをシグナル（「確実」に説明できる部分）とノイズ（「確実」には説明できない部分）に分解します．ここで大事なのは，ノイズとは分析をする上で「確実」には説明できない部分なのであって，必ずしもデータの測定誤差のみを指すわけではありません．例えば，ある銘柄の株価の動向が，様々な経済指標から各証券会社の社員食堂のメニューまで10億個以上の変数の動きによって正確に説明できるとしましょう．しかし，毎日10億個の変数を（もし可能だとして）収集し，保存するのは大変な仕事です．この株価の動向を予測することがそれだけの労力・資力を費しても意味のあることであれば良いのでしょうが，現実にその様な場合はまずないでしょう．一方で，その10億個以上の変数の内たった3個の変数がその株価の動向の大まかな部分を説明するならば，それは意味のあるシグナルということができます．この場合に，その3個の変数によっては説明されないが，残りの変数をもって説明できる部分はノイズの中に取り込まれることになります．

例として，図1を見てください．このデータはオレゴン州のある川で毎月の平均流水量を記録したものです．ここで，各点のX座標はある月の平均流水量，Y座標はその翌月の平均流水量をあらわしています．川の流水量は季節によって変動しますが，1月ごとに急激に増えたり減ったりするものではありません．実際，図1を見ると，点が右上がりに散らばっているのが

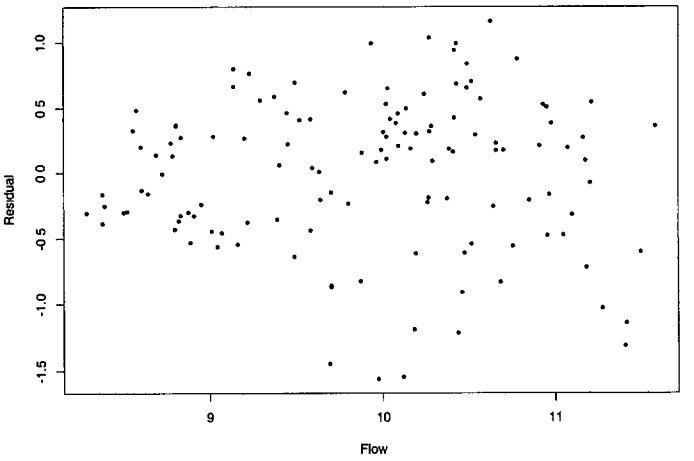
図1　流水量 vs. 翌月の流水量



わかります．一方，ある月の平均流水量から翌月の平均流水量が一意的に決められるわけではありませんから，点は必ずしもある直線あるいは曲線の上にのっかっているわけでもありません．そこでこの「右上がり」の関係を要約するためにデータをシグナルとノイズに分解しましょう．今，図1に描かれた直線をシグナルとして，X座標はある月の平均流水量，Y座標はその翌月の平均流水量から直線によって説明された量を引いたものとしたのが図2です．図1にあったようなはっきりとした関係はもう見られません．よりノイズらしくはありませんか？

データをシグナルとノイズに分解することで統計学は終わりなのかというとそうではありません．確かに統計学の文献をながめてみると，様々な状況を想定した分解のための手法が次から次へと提案されています．しかし，統計学のもう1つのお楽しみは（そして講義を受ける学生を多いに苦しめるものは？）この後に控えています．それはノイズとして「確実」には説明されなかった部分を「ノイズとして妥当であるかどうか」評価することです．この2段階の思考方法が統計学の「くせ」の1つです．

図2　流水量 vs.「ノイズ」



統計学の最大の発明の1つは，ノイズを評価・分析するために確率モデルを導入したことです．そして伝統的に，確率論に基づいたノイズの評価や分析が統計学の講義の中心となるために，統計学といえば確率論であると思う人も多いようです．さて，確率モデルは確率論を使った数学モデルです．これをデータにあてはめるときには，得ようとする結論がその数学モデルを通して表現できるものなのか，また現実にノイズとみなされた部分がそのモデルから大きく隔たっていないかを確認することが大切です．ところが，多くの統計のソフトウエァでは，シグナル・ノイズの分解とノイズの評価を1度にまとめてやってくれます．したがって，分解の過程で得られたノイズについて，ノイズの評価で使われている確率モデルとの整合性を確かめないまま，その評価を受け入れてしまう危険性があります．さきほど私は図1の中に「右上がり」の関係を見つけ，図2を見て「よりノイズらしい」と言いました．このようにデータや統計手法の結果を「目で見て」確認することは，今述べたような失敗を避けるのにとても有効な方法です．一般に，統計学の講義ではあまり時間のさかれることのないこの「目で見る」統計学についても

う少し考えてみたいと思います．

高次元のデータを見る？

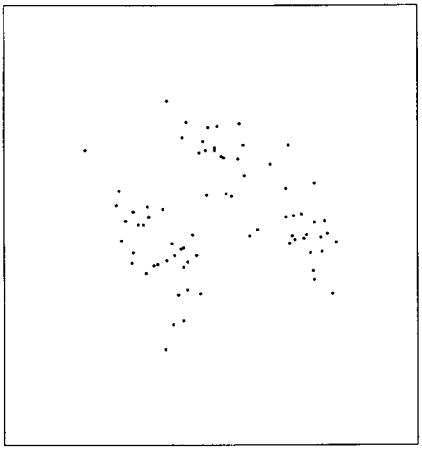
ある対象について，その特徴をどうにかして把握しようとして，多種類の特性が測定されることがよくあります．このようなデータを多変量データと呼びます．例えば，いくつかの企業を比較しようとする際に，収益率や資産効率といった財務指標から消費者の好感度アンケートといったデータまでが測られたりします．この場合にも統計手法の「くせ」は同じです．データをシグナルとノイズに分解し，ノイズの評価をおこないます．ところがデータや統計手法の結果を「目で見て」確認しようということになると勝手が違います．図1では各月の平均流水量と翌月の平均流水量を2次元空間の1点として表現することができました．今，5種類の特性を測定した多変量データがあるとしましょう．例えば，一橋大学の新入生の身長，体重，左右の視力，血圧でもいいでしょう．2次元の場合のアナロジーでいくと，各新入生は5次元空間の1点と考えることができます．これをどのようにして紙の上で表現しようかということになると，少し頭をひねらなければなりません．

多変量データを目で見ようという場合の代表的な手法は，データを2次元の平面に射影してやることです．さきほどの新入生の例で，とりあえず身長と体重だけ取り上げて絵を描こうという場合には，データを身長と体重という変数で決められる平面の上に射影することと同じです．5つの特性から2つを選ぶ選び方は10通りありますが，これ以外にも2次元の平面のとりかたは無数にあります．例えば，

$$\left(\frac{\text{右の視力}+\text{左の視力}}{2},\ \text{血圧}\right)$$

で決められる平面だってあります．このような2次元の平面への射影のすぐれているところは，結果として作られる絵が直観的に理解しやすいという点です．一方で，平面の選択の仕方によっては，データのもつ特異な側面を見失ってしまう危険性ももっています．できることなら，何かおもしろい「関

図3　3種類のノミのデータ



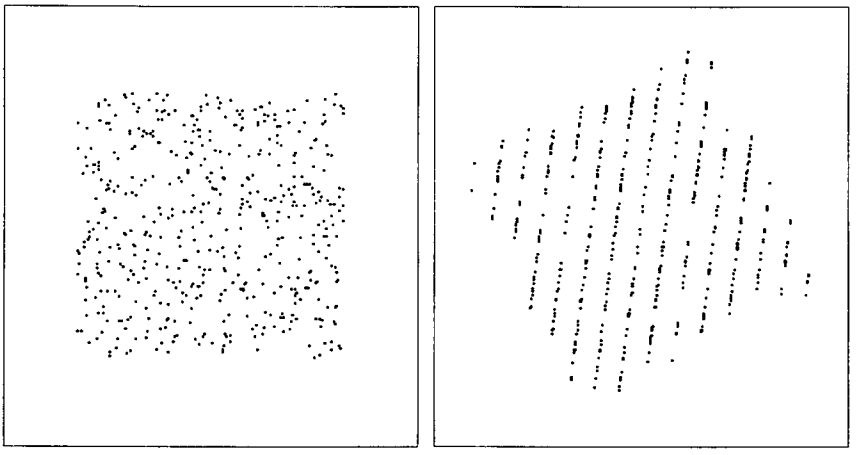
係」が見つかるような平面を選びたいところです．

ところで，図1の中に「右上がり」の関係が見えるというのはどうしてでしょう．点が2次元の平面全体に散らばっておらず，左上と右下に空白があるためです．したがって，それぞれの平面に射影した場合の点のばらつきを「計算」することができれば，おもしろい「関係」を見つけだすための指標として使うことができるかもしれません．このような考えを実現する統計手法の1つに主成分分析法があります．

図3を見てください．これは3種類のノミについて7つの特性を測ったデータから作られた絵です［3］．平面の左下，中上，右下に3つの点のかたまりがあるのが見てとれると思います．この平面は主成分分析法をつかって見つけだした平面で，7つの特性から2つを選ぶ21通りのどの組み合わせにもこのようなはっきりとした「関係」は見られません．

当然のことながら，主成分分析によって全ての「関係」が発見できるわけではありません．図4の左側はある3次元データの，主成分分析によって選ばれた平面への射影です．四角い領域の中に点が一様に散らばっているよう

図4　RANDUデータ



に見えます．ところが，このデータを別の平面に射影してみると，図4の右側のようなはっきりとしたパターンが現れてきます．このデータはRANDUという乱数を発生させるための計算手法を用いてつくりました．本来ならば3次元の立方体の中に点は一様に分布していなければならないのですが，この計算手法には欠陥があって，図4の右側のように平行に走る平面の上に点が集中してしまうのです．このように高次元のデータの中にかくれている可能性のある特異な射影をさがしだすための手法は一般にprojection pursuitと呼ばれ，現在も盛んに研究が行われている分野の1つです．このような手法は，先に挙げたデータをシグナルとノイズに分解する手法やノイズを評価する手法とも異なり，これらの手法を補強する「目で見る統計学」のための手法といえます．

絵を動かす？

これまで私達が見てきた絵は全て2次元の静止画像でした．もし，データを射影する平面を少しずつずらすことによって点を動かすことができるなら，

さらに多くの情報を得ることができるかもしれません．安価で高速なコンピュータの普及にともなって，こうしたダイナミック・グラフィックスをデータの分析につかおうという研究も現在盛んに行われています．また，2次元の射影では把握できないことがわかっている特徴を，データの部分的な射影を組み合わせることで発見しようという研究も進んでいます［1］［4］．統計学っておもしろい研究分野だとは思いませんか？

参考文献

［1］ Furnas, G. W. and Buja, A. (1994) Prosection Views : Dimensional Inference through Sections and Projections (with discussion). *Journal of Computational and Graphical Statistics*, **3**, 4, 323-385.

［2］ P. G. ホーエル著，浅井晃，村上正康共訳（1981）初等統計学（原書第4版）．培風館．

［3］ A. A. Lubischew (1962) *On the Use of Discriminant Functions in Taxonomy*, Biometrics, 455-477.

［4］ S. Oue (1994) Comment on Prosection Views : Dimensional Inference through Sections and Projections. *Journal of Computational and Graphical Statistics*, **3**, 4, 363-367.

（一橋大学専任講師）