教育のためのうまいグラフ表現

鈴木　義一郎

世は正にOA時代，否，鼻をたらした小学生までもが“パソコン”を手にするご時世である．また，教育の現場にコンピュータをといったかけ声のもと，“CAI”という言葉もとみに聞かされるようになった．ところで，世に出まわっているパソコン用のソフトで，中学生の数学といった類のものを眺めてみる．従来のワークブック形式の自習書をそのまま計算機のパターンに移しかえただけで，教育面での配慮の欠けすぎたものが目につく．わざわざディスプレイ表示を用いるわけだから，グラフや図形を多用すべきだし，さらに“動き”も加えたほうが効果的であろう．しかし実際には，安かろう悪かろうの粗製乱造ぶりである．

“目は口ほどに…”といった言葉を導入するまでもなく，視覚を援用して教育することの効果は衆目の一致するところである．ただ，黒板に逐一図を書いたりすることは，存外やっかいなためにかなり熱意のあるタイプの教師でないと，そのような教え方はとらないだろう．しかし現在はスライド，OHP，ビデオ，そしてコンピュータと文明の利器をいろいろと利用できるご時世である．とかく“メカに弱い”とされている中年の現場教師でも，“手軽に扱える”教育のための補助教材が開発されて然るべきではなかろうか．

筆者は日頃より，統計教育の場でなるべく多くのグラフ表現を用いるように心がけてきたつもりである．たとえば相関係数や相関比といった指標とデータの散布パターンの対応([2]，[3]等)，統計データの分析のさいのグラフ表現の効用（[5]等)，分割表のモデルに対するグラフ表現([4])，確率の概念の導入とカルノーマップ([6])といったものを参照されたい．

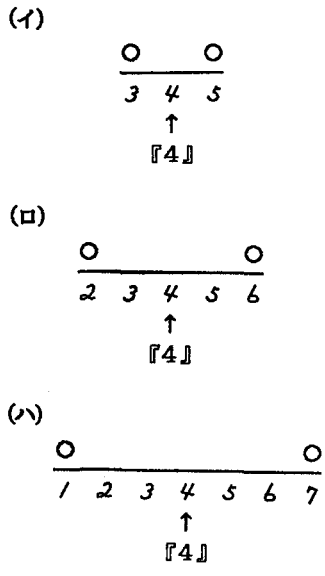
ともあれ教育の場では，教師が主体であってはいけない．ということは，教師の都合により概念を押し売りしてはいけない．なぜ，その概念を導入する必要があるのか，この点を生徒に納得させてから導入をはかるという配慮が大切である．この小稿では，視覚的なイメージに訴えた効果的な概念導入の一端を紹介してみる．また統計手法のなかには，概念を間違って認識したままで利用されるケースというのも珍しくはない．そのような場合，可能なかぎり単純化されたモデルを利用して間違いを正してやる，このことも教育の場では別して大切なことである．

標準偏差の概念の導入

統計的方法のなかで，最も基本的な概念としてまず「平均」が登場する．これは，一群のデータの分布パターンの“バランス・ポイント”というイメージで理解させることができる．

ついで，なぜ平均だけでは済まされないのか？という問題が提供され，そこで，図1の登場となる．(イ)，(ロ)，(ハ)の3つのパターンはいずれも平均が『4』であるが，明らかに異なった様相を呈し

すずき　ぎいちろう　統計数理研究所
〒106　港区南麻布4－6－7



◀ 図 1
平均が同じデータ

ている．

そこで，「範囲」という概念の導入を余儀なくされることになる．(イ)，(ロ)，(ハ)というデータの「範囲」は，それぞれ２，４，６と算出されるから，めでたく識別することができた．

そこで次の図２を提示する．(イ)，(ロ)いずれのパターンも平均が『４』，範囲が「６」と同じであるが，明らかに分布は異なっている．つまり「範囲」には，中間の値が関与してこないので，(イ)の③，

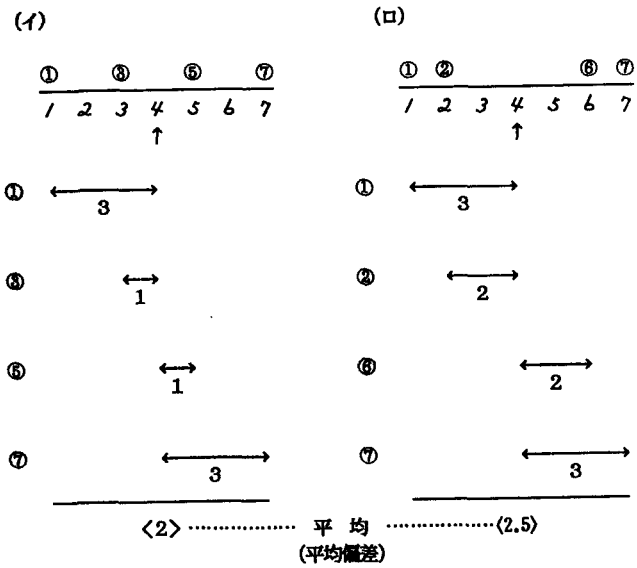


図 3　平均偏差の算出

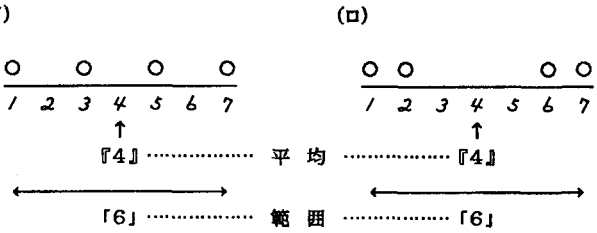


図 2　平均と範囲が同じデータ

⑤と(ロ)の②，⑥というデータを区別することができなかった．ここで初めて，データのすべての値に関係するような“尺度”の導入の必然性が与えられることになる．かくて，図３の登場となる．

(イ)のデータのパターンでは

①は平均『４』から“３”ずれている

③　〃　“１”　〃

⑤　〃　“１”　〃

⑦　〃　“３”　〃

このような“ずれ”の平均値を求めて

$$\frac{3+1+1+3}{4}=2$$

がデータの散布度を表わす尺度として利用できる．いわゆる「平均偏差」とよばれている概念である．(ロ)のデータについても同じような計算を行なって，〈2.5〉という結果を得る．後者のほうの散らばり方が大きいといったパターンを，めでたく識別することに成功した．

最後にいよいよ，図４の登場である．各データの平均からの“ずれ”を，平均とデータのあいだでの“絶対差”の代わりに“差の平方値”で評価した場合で，いわゆる「分散」である．この値は，平均する以前に平方したために，データのもっている単位の２乗になっている．単位をそろえるために“平方根をとる”という操作を加えて，「標準偏差」という尺度が与えられる．

ここで，なぜ平方値を考えたりしたのかを説明することが，存外やっかいである．直感的には，絶対差で“ずれ”を把えることのほうが自然に思われるからである．標

本の変動による挙動を調べるうえでは，平均偏差より分散のほうが扱いやすいという事実は，理論家なら誰でも知っている．しかしそのような理論の展開を理解させることは，相応の予備知識をもった相手でないと困難である．扱いにくい概念に敢然と挑戦をしない理論家の無能さを非難されてもたまらないから，平均偏差（一般には奇数次の絶対モーメント）の標本分布が理論分布のパターンに逐一関係するために，たくさんの数値表が必要となるので汎用性に乏しいといった理由をもちだすことになる．

回帰モデルの適合度

統計手法のなかで，最も多用されているものの1つに「回帰分析」があげられる．被説明変数 y と説明変数 x とのあいだに“線形構造”のようなものを想定し，その直線なり平面なりを推定する．その場合，当該現象が想定した“関係”によってどの程度うまく説明できているかを吟味すること，これが特に肝心なことである．

そこで登場する概念として，残差平方和，これをデータ数で割った残差分散といったものが登場する．これらは，想定した“関係”で説明しそこなった残りの成分とみなされるから，ひたすら小さくすることが至上であると考えがちになる．一般に“関係”を規定するパラメータの数を増やしてやれば，残差平方和を小さくすることができる．当面のデータを記述することだけが目的ではなく，むしろ将来の結果の予測のために利用したいということなので，推定した内容の“安定性”こそが問題なのである．

この種の警告を納得させるために，次の最も単純なケースのモデルを提示すべきである．いま，推定しようとしている真のモデルを

$$y=(0.5)x+\varepsilon$$

とする．ここで“誤差”の分布を

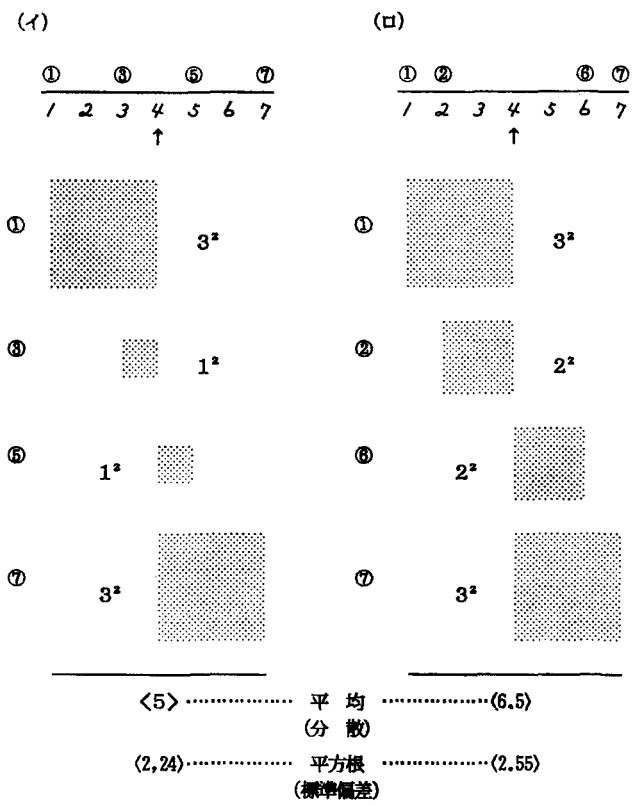


図 4　標準偏差の算出

$$\varepsilon=\begin{cases}-0.5 & \text{確率}1/4\\ 0 & \text{確率}2/4\\ +0.5 & \text{確率}1/4\end{cases}$$

のような三角形の3点分布で与える．さらに，観測すべき x の値としては，$x=1$，$x=2$ の2点だけに限り，$x=1$ のときの y の値を y_1，$x=2$ のときの y の値を y_2 とする．

ε のとり得る値は3通りあるので，(y_1, y_2) の組合せは表1のように $3\times3=9$ 通りあり，それぞれの結果の生じる確率も P 列の値で与えられる．次に，このようなモデルの勾配が0.5であるとは知らされていないので，それを“b”と書き

$$(1,\ y_1),\ (2,\ y_2)$$

という2点の観測値を用いて b の値を推定する．これらの点と“$y=bx$”という直線とのずれはそれぞれ y_1-b，y_2-2b であるから，平方して加えた値（残差平方和）は

$$Q=(y_1-b)^2+(y_2-2b)^2$$

$$=5\left\{b-\frac{y_1+2y_2}{5}\right\}^2+\frac{(2y_1-y_2)^2}{5}$$

となる．これを最小にするbの値として

$$\hat{b}=\frac{y_1+2y_2}{5}$$

これが“最小2乗推定”である．

表1の$\hat{b}$の列に，(イ)から(リ)までについてのこの値が記述されている．真の値“0.5”を的中させているのは(ホ)の場合だけで，その確率は1/4にすぎない．真の値から0.1の範囲だけずれているケースとなると，(ロ)，(ハ)，(ト)，(チ)も加えられ，的中率は5/8と高くなる．

さて表1の最後の列の値をみてみよう．これは最小2乗推定を用いたときの残差平方和の値である．(ホ)のケースは確かに真の値を言いあてているし，Qの値も0である．ついで(イ)，(ニ)，(ヘ)，(リ)といったケースでもQの値はかなり小さい．ところが(イ)や(リ)の場合の$\hat{b}$の値は真の値から0.3もずれた推定を行なっていて，9つのケースで最悪である．これと反対に(ハ)や(ト)のQの値はかなり大きいのに$\hat{b}$の値が真の値に近い推定結果を与えている．

このような事情は図5を眺めてみると一目瞭然である．つまり最小2乗直線とは与えられたデータに最も近い“直線”を推定しているだけでそれがそのまま真の“直線”に近いとは即断できないのである．さらに9つのケースのQと$\hat{b}$とを同時にプロットした図6を眺めてみる．$\hat{b}$とQとが異相関であるといった事情が読みとれる．これは残差平方和Qの値の大小が最小2乗推定$\hat{b}$の良否に対応するとは限らない，という事実を示している．

教育の現場にもっとグラフ表現を…

今から30年前のJRSS(A)誌(引用文献[1]参照)に，E. S. Pearsonによる英国統計学会での会長講演の内容が紹介されている．

まず冒頭で彼の父 K. Pearson がロンドン大学で行なった講義内容に触れ，特に幾何学的な説明に重点を置いていた事実を指摘している．たとえば“Geometry of Statistics”と題する講義テーマの要約として，次のように述べている．

表1

	y_1	y_2	P	$\hat{b}$	Q
(イ)	0.0	0.5	1/16	0.2	0.05
(ロ)	0.0	1.0	2/16	0.4	0.20
(ハ)	0.0	1.5	1/16	0.6	0.45
(ニ)	0.5	0.5	2/16	0.3	0.05
(ホ)	0.5	1.0	4/16	0.5	0.00
(ヘ)	0.5	1.5	2/16	0.7	0.05
(ト)	1.0	0.5	1/16	0.4	0.45
(チ)	1.0	1.0	2/16	0.6	0.20
(リ)	1.0	1.5	1/16	0.8	0.05

「物理現象や社会現象を扱ううえでの科学的手段として，幾何学的方法と算術的方法とはともに大切である．ところが，幾何学的方法は大衆的な表現の一手段にすぎないといった，誤った意見もある．統計的素材を開発し解析していくうえで，幾何学的方法こそが最も基本的なものである」

E. S. Pearson自身も，データのパターンを視覚的に検討することが利用しようと考えているモデルの適否を判断するのに効果的であると述べている．ここで，統計図表の意味合いを検出する能力の認められる統計家にとっては，というただし書きも附されている．ところが数理統計学のコースで，数学的によく訓練を受けた平均的学生が，視覚的なイマジネーションをもつようには仕向けられていない．このことは，大学レベルも含む一般の学校教育で，かなり深刻な問題であると指摘している．

これを受けて，M. G. Kendall は次のような問題提起を行なっている．

「大学教師は，職業訓練を一切受けなくてもなることのできる唯一の専門職である．これは，経験によって教育法を学ばなければならないということでもあり，しかも生涯それを学ぼうとしない者もいる．このような状況を考えるならば，統計的なアイディアを学生たちに授けるための技術を

いろいろ提供し合う機会を作るべきである．そのような努力が一部では払われてきたかもしれないが，より一般に普及するところまでには至っていない」

この問題提起に対する1つの答えを用意することがこの原稿をしたためた目的であるとして，E. S. Pearsonは自分が統計学を学びはじめた頃に重要な鍵を与えてくれたという，図形にまつわるエピソードを披露している．1つはK. Pearsonが，物理的な意味での関係と統計的な相関関係とのあいだの相違を示すのにうまい図的表現を用いた事例．そしてもう1つは，R. A. Fisher の1915年の論文に挿入されていた図が平均や標準偏差そしてステューデント比といった概念の関係を把握するのに非常に有効であったという事例である．

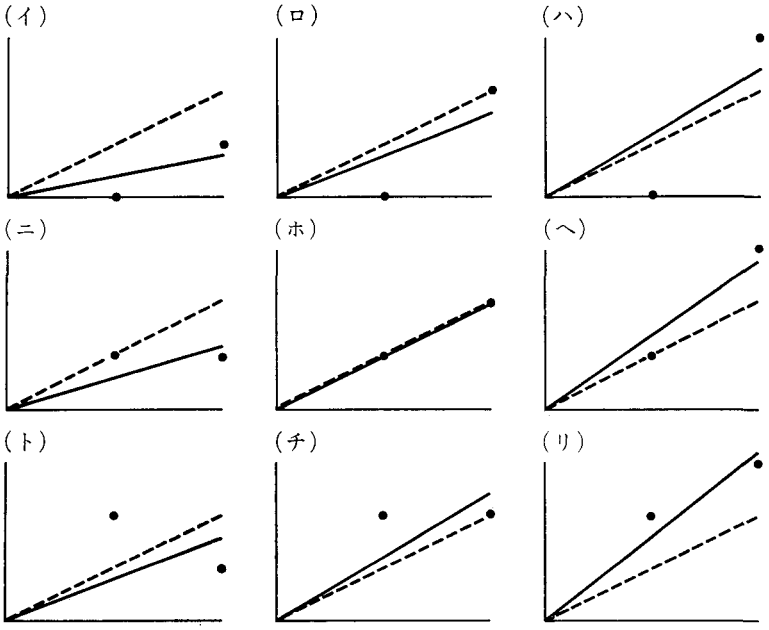


図 5　回帰直線の分布パターン

E. S. Pearsonの論文にはこの他にも含蓄のある提言が行なわれているのだが，この小稿の主旨からして割愛せざるを得ない．ともあれ教育の現場特に入門書の記述にはもっと図表をとり入れていくべきであるとの主張は，30年後の今日でもなお耳を傾けるべき警鐘として受けとめるべきであろう．

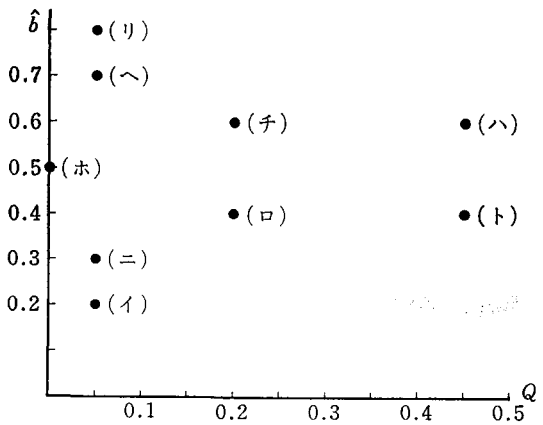


図 6　Qと$\hat{b}$の散布図

引 用 文 献

[1] Pearson, E. S. : Some Aspects on the Geometry of Statistics. *J. Roy. Statist. Soc.* Series A Vol. 119, 1956, 125-149.

[2] 鈴木義一郎：統計のイメージ，数理科学 No.165, 1977, 41-45.

[3] ——：データ解析術，実教出版1977, 33-34.

[4] ——：2×2分割表に対するモデルの図式表現と潜在構造分析，統計数理研究所彙報，32, 1984, 173-195.

[5] ——：統計におけるグラフ表現の効果的利用法，応用統計学，14-1, 1985, 27-37

[6] ——：統計学で楽しむ，講談社，1985