2005/07/14

データ分析におけるグラフの価値について

0000000000 神奈川太郎

要旨

データを統計的手法で分析するための手法を説明する教科書などでは、グラフを使ってデータを大づかみにすることは「不正確」として軽視されているのが実態である。しかし、実際のデータを分析する際には、そもそものデータがどのような性質を持っているかについてのアタリをつけなければ適切な統計手法を選択することができず、闇雲にコンピュータを回してもいい結果は得られない。本稿では(Anscombe 1974)の数値例を用いてグラフでデータを見てから分析することの重要性を示す。

1. なぜグラフが重要なのか

データを統計学的に分析する手法について説明した本や統計分析パッケージソフトウェアの解説書では、データをグラフに図示するという「原始的」な手法は極めて軽い扱いをされているケースが多い。グラフが軽視されている理由はいくつかあるようだが、よく聞くものとしては

1. グラフは雑でいい加減である。数値計算をした方が「正確」である
2. あるデータを分析する際に「適切な」手法は一つしかない
3. 実際のデータを「見て」いい結果が出るよう分析手法を変えるのはずるい

などがある。しかしながら実際のデータ分析は理論モデルにそのままデータを投げこんで統計パッケージで計算して結果が出ればいいというものではないことも事実である。たとえば、多くの統計的手法はデータが「望ましい」性質を持っていることを仮定しているが、実際のデータは統計手法の都合とは関係なく発生しており、普通はそのような仮定を完全には満たしていない。数値的に難しい計算をするまえに、大づかみにデータがどのような性質をもっているかを知るツールとしてグラフは大変便利である。また「変なデータ」がどの程度混じっているかを直観的につかむためにもグラフは欠かせない。もし、最初に考えていた統計的な手法では実際のデータをうまく分析できそうもないとわかったときも、それを回避できる分析手法を選ぶためにもグラフは便利であることは

言うまでもない。

2. データ例とその概要

表1は Xと Yという 2 つのデータの組み合わせを 4 セット(X_1と Y_1, X_2と Y_2, X_3と Y_3, X_4と Y_4)示したものである。よく使われるデータの特性値である平均と分散を見ると、X_1～X_4 では平均が 9、分散が 10 とすべて同一となり、Y_1～Y_4 では平均が 7.5、分散は 3.75 とこちらもすべて同一となっている。X_1～X_3 は同じデータであるから当然であるが、Y についてもほぼ同じようなデータと考えていいのだろうか？

表1. Anscombeの数値例

観測番号	測定1		測定2		測定3		測定4	
	X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
1	4	4.26	4	3.1	4	5.39	8	5.25
2	5	5.68	5	4.74	5	5.73	8	5.56
3	6	7.24	6	6.13	6	6.08	8	5.76
4	7	4.82	7	7.26	7	6.42	8	6.58
5	8	6.95	8	8.14	8	6.77	8	6.89
6	9	8.81	9	8.77	9	7.11	8	7.04
7	10	8.04	10	9.14	10	7.46	8	7.71
8	11	8.33	11	9.26	11	7.81	8	7.91
9	12	10.84	12	9.13	12	8.15	8	8.47
10	13	7.58	13	8.74	13	12.74	8	8.84
11	14	9.96	14	8.1	14	8.84	19	12.5
平均	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
分散	10.00	3.75	10.00	3.75	10.00	3.75	10.00	3.75

Anscombe, F.J. 1973, ″Graphs in Statistical Analysis″, *The American Statistician*より引用
ただし、変数の並びはXが昇順になるように並べ替えてある。

さらに、Xと Yの 2 変数の関係が、i番目のペアの変数 x_iと y_iについて

$$y_i = a + bx_i + \varepsilon_i \qquad (1)$$

という関係を持っていると仮定してみよう。つまり、実際に観察されている y_i は、aとb で決まる直線の方程式と x_i から計算される値に残差ε_i を加えたものだと考えるわけである。このとき、「最適な」aとb を決める方法はいろいろ考えられるがよく使われる手法は OLS(Ordinary Least Square)と呼ばれる手法である。この手法では、データに含まれる x_i, y_iの組に対してε_i の 2 乗(square)し

たものをすべて加えたものを最小化(least)するようなaとbが「最適な」aとbであると考える(もちろん、このようなaとbが本当に最適であるためにはいろいろな仮定が必要であるが[1]、つまり、

$$\sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \{y_i - (a + bx_i)\}^2 \qquad (2)$$

を最小にするようなaとbが最適であると考えるわけである。この計算自体は Excel も含めていろいろなソフトで計算できるが、実際に計算してみるとX_1とY_1, X_2とY_2, X_3とY_3, X_4とY_4ではすべて

$$y = 3 + 0.5x \qquad (3)$$

となる。(3)式の関係からｙの推定値を計算してみると表2のようになる。

表2　OLSによる推定値

観測番号	X_1～X_3	$3+0.5x_i$	X_4	$3+0.5x_i$
1	4	5	8	7
2	5	5.5	8	7
3	6	6	8	7
4	7	6.5	8	7
5	8	7	8	7
6	9	7.5	8	7
7	10	8	8	7
8	11	8.5	8	7
9	12	9	8	7
10	13	9.5	8	7
11	14	10	19	12.5

つまり、グラフを描かないでの分析では「X_1とY_1, X_2とY_2, X_3とY_3, X_4とY_4は大体同じようなデータであり、2 変数の関係も大体同じようなものである」という結論が得られた。めでたしめでたし……ではない。

3. グラフを描いてみよう

X_1とY_1, X_2とY_2, X_3とY_3, X_4とY_4をそれぞれ散布図として図示し、(2)式で表される直線を追加したものが**エラー! 参照元が見つかりません。**～図4である。X_1 と Y_1 の関係を表した図1は(2)式

[1] (Amemiya 1985)によるとそのような仮定は実際のデータで満たされやすいからではなく、話を簡単にするためにおかれているらしい。

の直線が大体のところデータの関係を表しているように見えるが、図2は放物線のように見えるし、図3は変な値が一つあるため図からみてもっともらしい直線より傾きが大きくなっている。また、図4は大部分のデータが縦1直線に並んでいて1つだけ変な値があるように見える。つまり、図を描いたことからわかることは、図1のデータ、つまり X_1 と Y_1 の関係以外は直線で説明すること自体がデタラメな分析であり、その結果出てきた数値はどんなに「確実」で「正確」であってもデタラメである。

実際のデータ分析ではここまで極端なことは滅多にないが、やはりデータの関係を散布図として図示してから難しい統計的手法を使う習慣はとても大事である。散布図を手作業で描くのは結構メンドウであるが、Excel を使えば瞬時に散布図は作成可能であるから、データ分析を行う時には「まず散布図」(本当はクロス集計表も)を合い言葉にしていきたいものである。

参考文献

Anscombe, F.J., “Graphs in Statistical Analysis,” *The American Statistician*, 1973.

Amemiya, T., *Advanced Econometrics*, Harvard University Press, 1985.

図1 比較的まともな推定

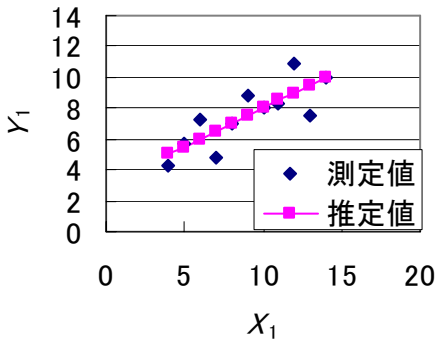

図2 曲線を直線で推定

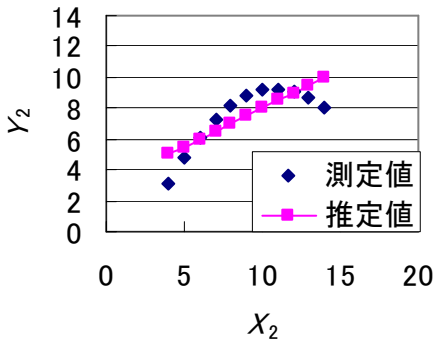

図3 異常値で引っ張られる直線

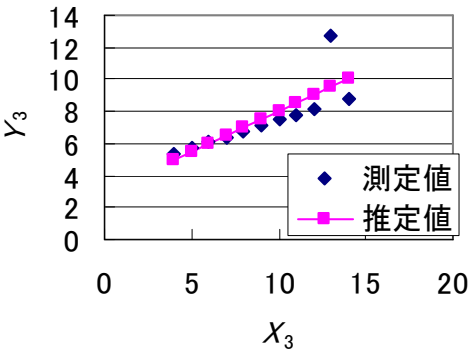

図4 異常値で全く違う直線を推定