実験や調査をして得られたデータの解析手法としてすぐに思い浮かぶのは、平均値や相関係数を求めるなどといった統計的な解析であろう。しかし、統計的な分析だけではなく、人間が見える形に変換してその傾向をつかむことも重要である。例えば、アンスコムの 4 つ組と呼ばれる有名な例がある。この例では、4 つのデータが提示される。k 番目のデータは次の書式で与えられる。

$$(a_{k,1}, b_{k,1}), (a_{k,2}, b_{k,2}), \ldots, (a_{k,11}, b_{k,11}) \tag{1}$$

ここで、$a_{k,n}, b_{k,n}$ は実数である。これら 4 つのデータは、統計的に分析して得られる平均値や相関係数といった値は全て（ほぼ）等しい。統計的な解析だけを頼りにすると、これら 4 つのデータは同質のデータであると言ってしまうかもしれない。しかし、これらのデータを散布図に直してみると、驚くほどその分布は異なっていることがわかる。

図 1 は、アンスコムの 4 つ組の第一のデータを散布図にしたものである。いかにも 2 つの変量の間に正の相関がありそうな散布図をしていることがわかる。我々が「2 つのデータに正の相関がある」と聞いたとき、真っ先に連想するであろう散布図は、図 1 のようなものであろう。

図 2 は、アンスコムの 4 つ組の第二のデータを散布図にしたものである。確かに正の相関はあるかもしれないが、そのデータの分布は直線よりかは放物線の方が近い形状をしている。このデータだけを見ると、直線での近似をためらう人もいるかもしれない。

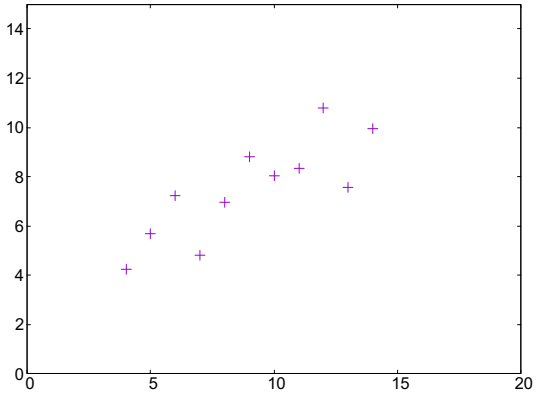

図 1　アンスコムの 4 つ組の第一のデータを散布図にしたもの

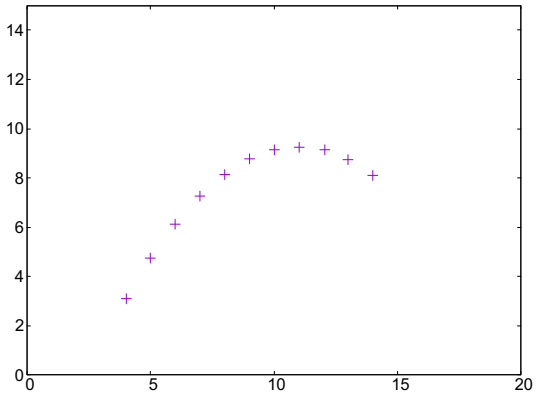

図 2　アンスコムの 4 つ組の第二のデータを散布図にしたもの

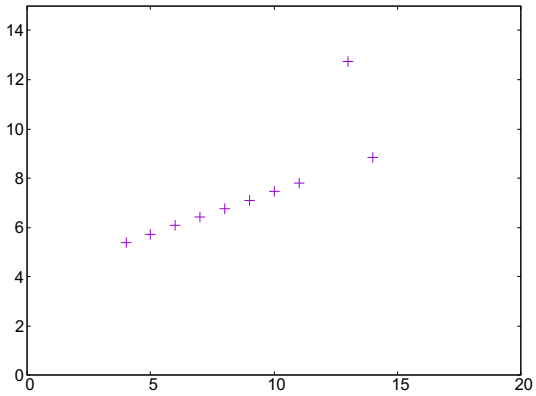

図 3　アンスコムの 4 つ組の第三のデータを散布図にしたもの

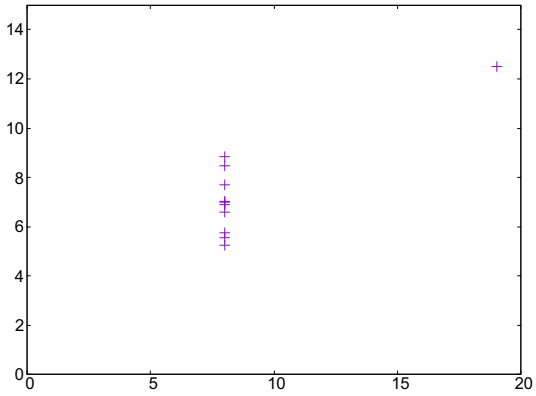

図 4　アンスコムの 4 つ組の第四のデータを散布図にしたもの

図 3 は、アンスコムの 4 つ組の第三のデータを散布図にしたものである。確かに正の相関があると言えるデータだろうが、1 つだけ外れ値があることが読み取れる。このような場合は、このままデータの解析を進めるのではなく、外れ値がなぜ得られたのかの考察をすべきだろう。

図 4 は、アンスコムの 4 つ組の第四のデータを散布図にしたものである。第三のデータと同じく 1 つだけ外れ値があるが、そのデータの分布は第三のデータとは大きく異なっている。このデータの場合、正の相関があると言うことは最早不適切であろう。