相関係数と順位相関係数について

相関係数と順位相関係数の使い分けについて考えてみる。今、図 1 と図 2 のような 2 種類のデータを考える。(相関と順位相関.txt)

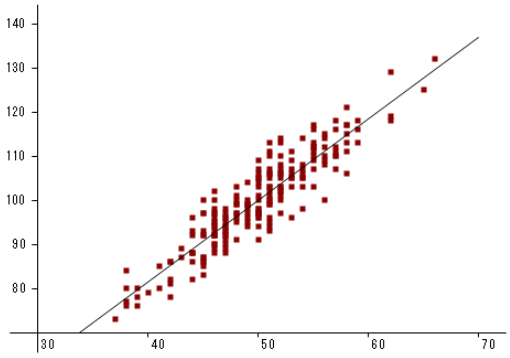


図 1　データ 1

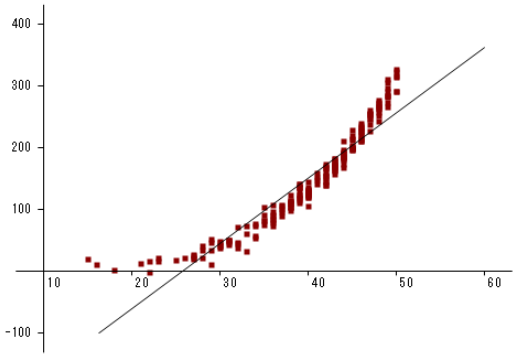


図 2　データ 2

この 2 つのデータには、x と y に関係がありそうである。左のデータ 1 は 2 変量正規分布と呼ばれるデータで、回帰直線がよく関係を表している。それに対して右側のデータ 2 の関係は曲線を描いており、直線で表すには無理がある。

この 2 種類のデータについて 2 つの変数の正規性を調べてみると図 3 と図 4 のようになる。

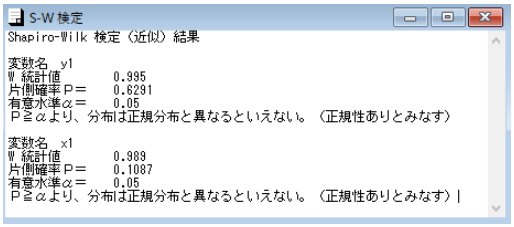


図 3　データ 1 の正規性

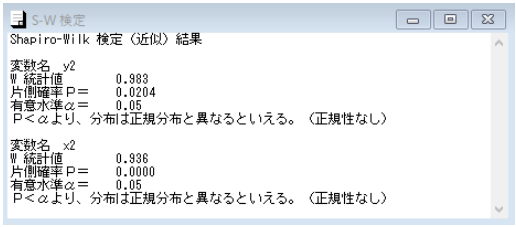
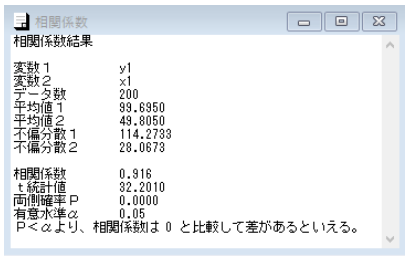


図 4　データ 2 の正規性

データ 1 は 2 変数ともほぼ正規分布であり、データ 2 は 2 変数とも正規分布といえない（もちろん一方が正規分布する場合もある）。

次に 2 つのデータの相関係数と順位相関係数を求めてみよう。結果は図 5 と図 6 の通りである。



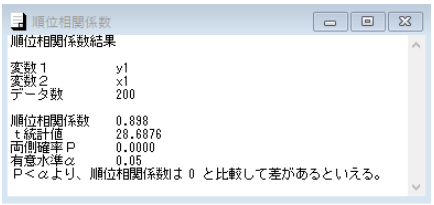


図 5　データ 1 の相関係数と順位相関係数

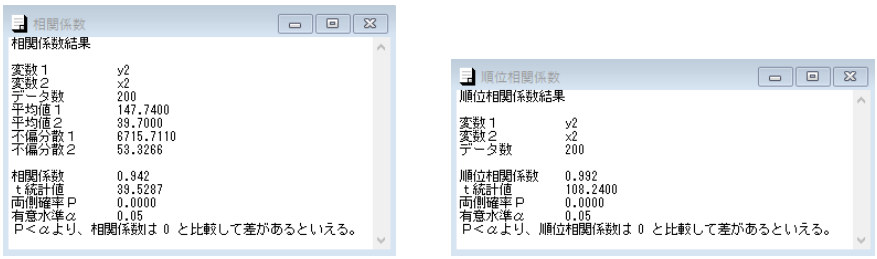


図６　データ２の相関係数と順位相関係数

図 5 のデータ１の場合、相関係数（0.916）＞順位相関係数（0.898）となり、どちらかと言うと相関係数が関係を表すのにふさわしいようである。これに対して図 6 のデータ２の場合、相関係数（0.942）＜順位相関係数（0.992）となり、順位相関係数がより強く関係を表している。散布図のデータ点の並びが曲線に沿うような場合は、2 つの変数に正規性が見られないことがあり、順位相関係数を利用する方が関係を強調できる場合がある。

特に図 7 のデータ３のように散布点が単調増加関数上に並ぶ場合、図 8 のように順位相関係数は誤差の範囲で１になる。

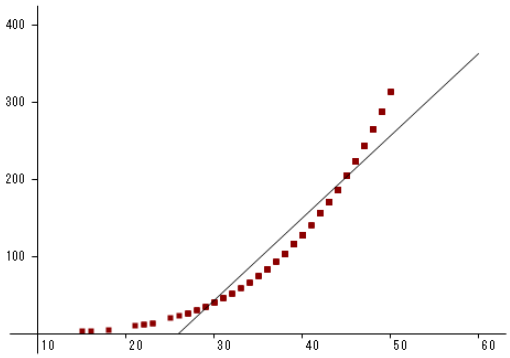


図７　データ３

図８　データ３の順位相関係数

最後に一つ注意を与えておく。ここでは 2 つの変数を別々に分けて（周辺分布の）正規性を調べていることである。本来相関係数は 2 変量正規分布を基にしているので、直接 2 変量正規分布に従うかどうか調べるのが正確な方法である。そのため、ここでの議論は簡易的方法と認識しておく必要がある。