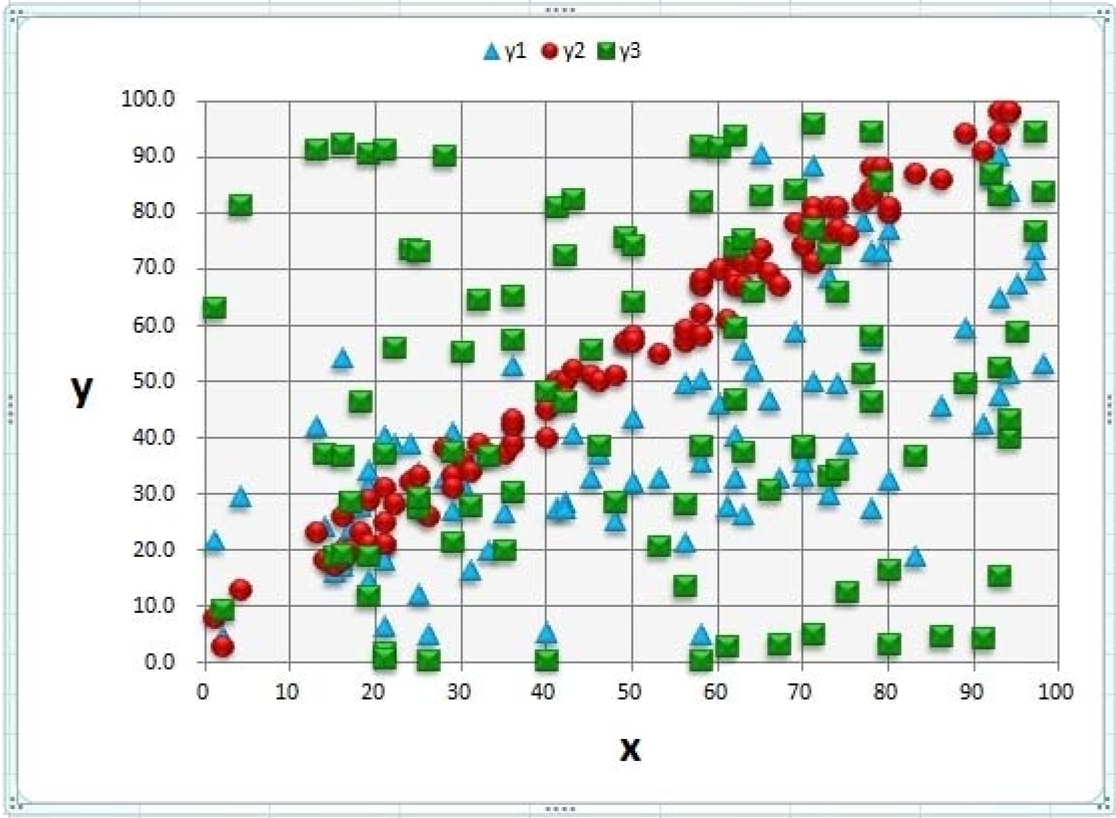
相関関係と相関係数

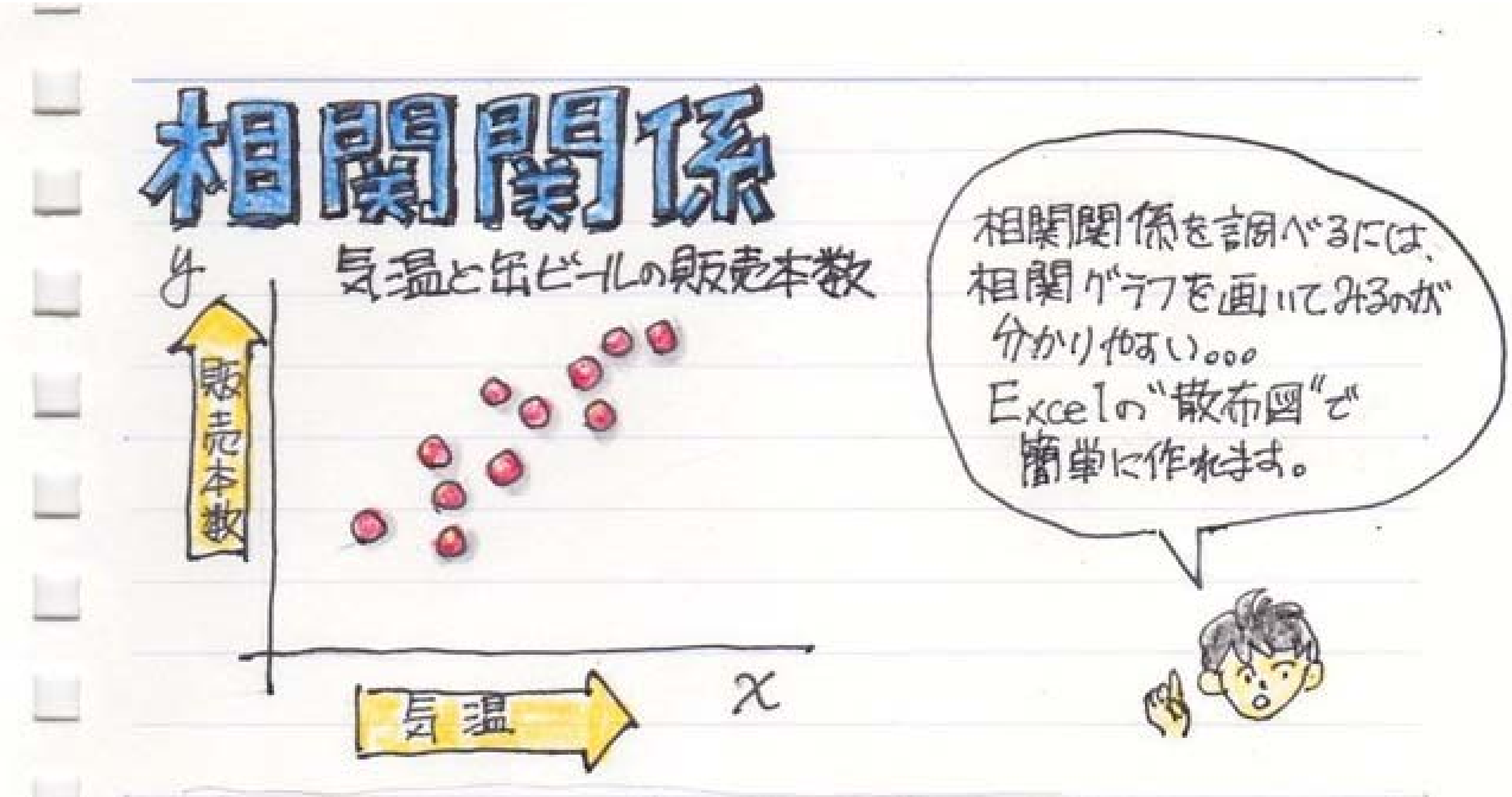
v0.5

舘野

相関グラフ（散布図）の例

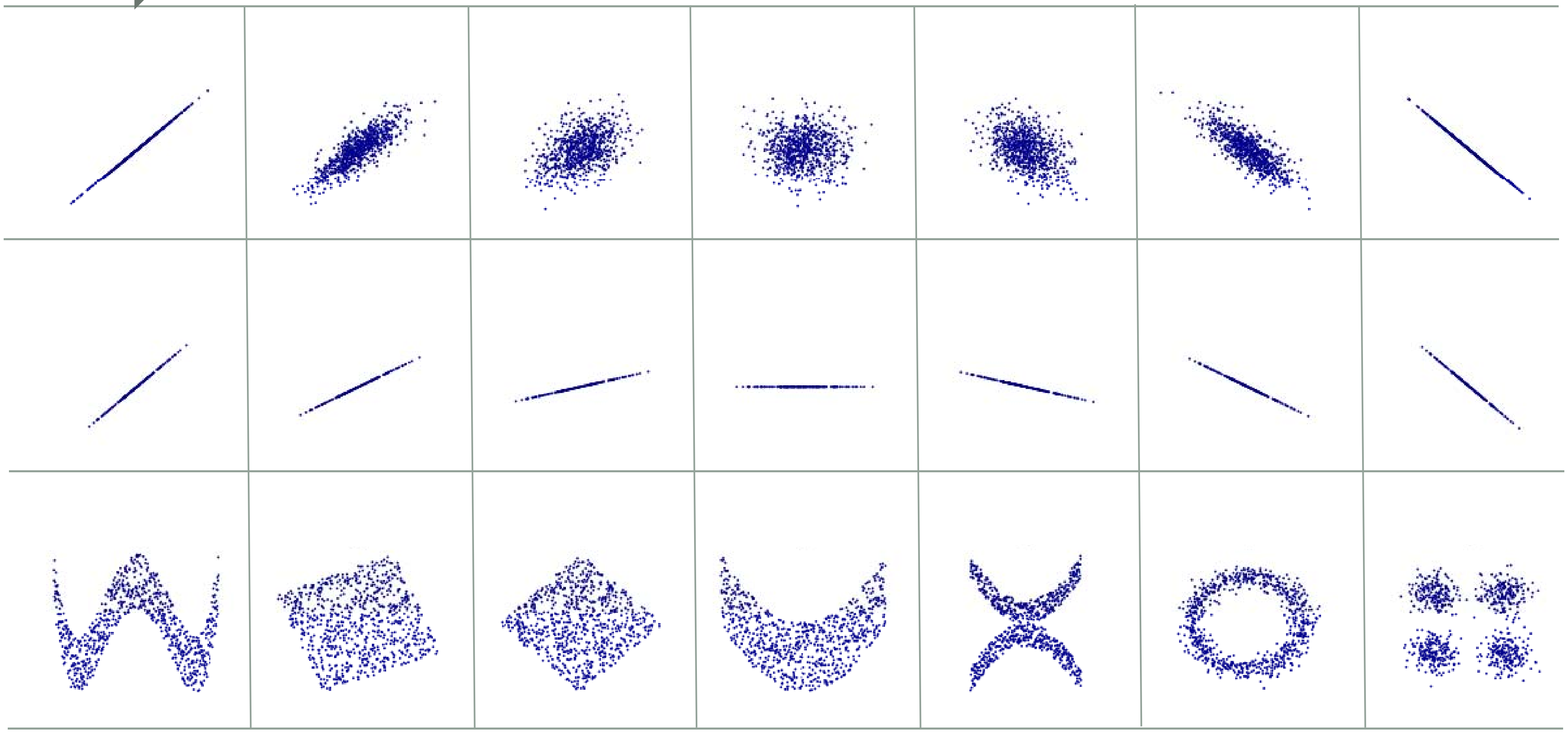


http://haku1569.seesaa.net/archives/201410-1.html



http://haku1569.seesaa.net/archives/201410-1.html

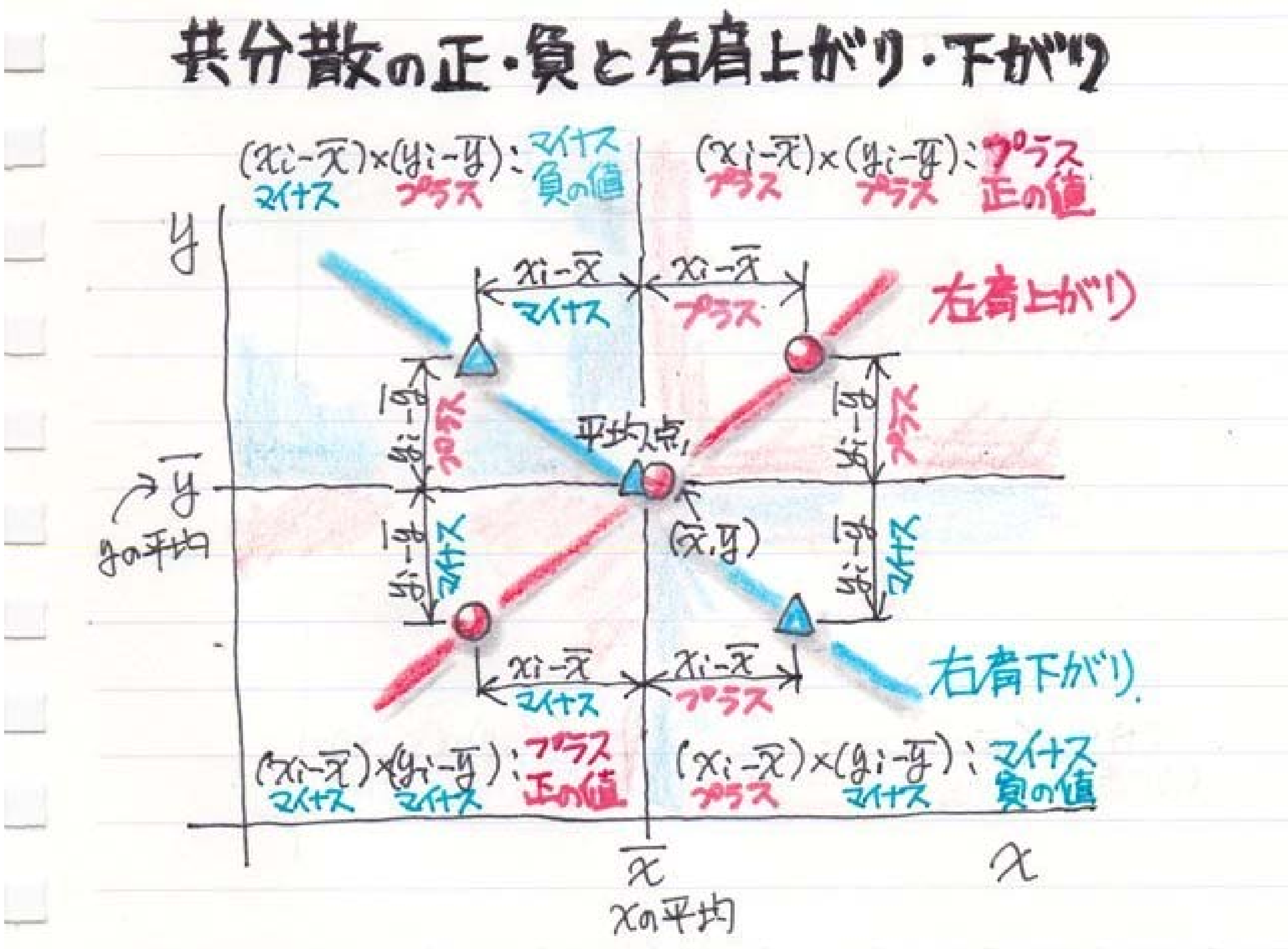
このグラフの相関係数はいくつでしょう？



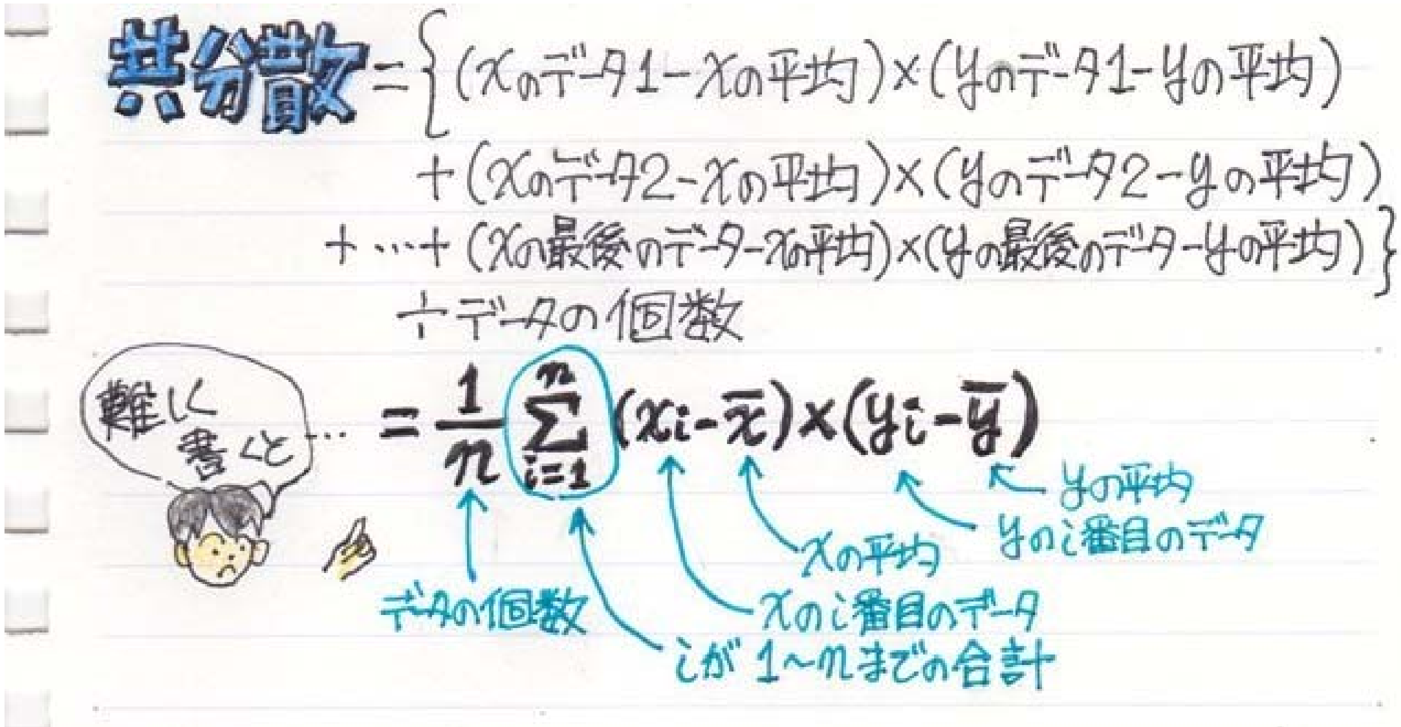
https://en.wikiversity.org/wiki/Statistics

共分散（きょうぶんさん、covariance）は、2 組の対応するデータ間での、平均からの偏差の積の平均値である。

https://ja.wikipedia.org/wiki共分散



http://haku1569.seesaa.net/archives/201410-1.html



http://haku1569.seesaa.net/article/407065774.html

$$\text{相関係数} = \frac{\text{共分散}}{x\text{の標準偏差} \times y\text{の標準偏差}}$$

ここで、
共分散 ≦ xの標準偏差 とyの標準偏差 の積

∑: i = 1 から i = n までの総和
xm: x の平均、xm = Σxi/n
ym: y の平均、ym = Σyi/n
sqrt: 平方根

共分散
((xiとxmとの差)と
((yiとymとの差)の積
の合計の平均

yの標準偏差 $\Sigma(yi\text{-}ym)^2$
((yiとymとの差)の2乗の合計)の平均の平方根

xの標準偏差 sqrt[$\Sigma(xi\text{-}xm)^2$
((xiとxmとの差)の2乗の合計)の平均の平方根

$$\text{相関係数} = \frac{\text{共分散}}{x\text{の標準偏差} \times y\text{の標準偏差}}$$

相関係数は2つの方法で求められますが、その結果は同じになります。

偏差積和

$$S_{xy} = \Sigma(x_i-\bar{x})(y_i-\bar{y})$$

相関係数 $$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

$$S_{xx} = \Sigma(x_i-\bar{x})^2$$

偏差平方和

共分散

$$s_{xy}^2 = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

$$r_{xy} = \frac{s_{xy}^2}{\sqrt{s_x^2 s_y^2}}$$

$$s_x^2 = \frac{\Sigma(x_i-\bar{x})^2}{n-1}$$

分散

http://www.kogures.com/hitoshi/webtext/stat-soukan/index.html

相関係数の値はなぜ [-1 ～ +1] の範囲か（証明）

相関係数：r（correlation coefficient）は2つの変数間の相関すなわち類似性を示す指標で、次式で定義されます。

実数データ：(xi, yi), i = 1, 2, .., n に対して
r = Σ(xi-xm)(yi-ym)/sqrt[Σ(xi-xm)^2・Σ(yi-ym)^2]
ここで、
∑: i = 1 から i = n までの総和
xm: x の平均、xm = Σxi/n
ym: y の平均、ym = Σyi/n
sqrt: 平方根

相関係数の値の範囲の証明には、**Schwartz（シュワルツ）の不等式**（下記注）や多次元ベクトルの内積を利用した方法などが知られていますが、ここでは2次関数の判別式を利用する簡単な方法を紹介します。

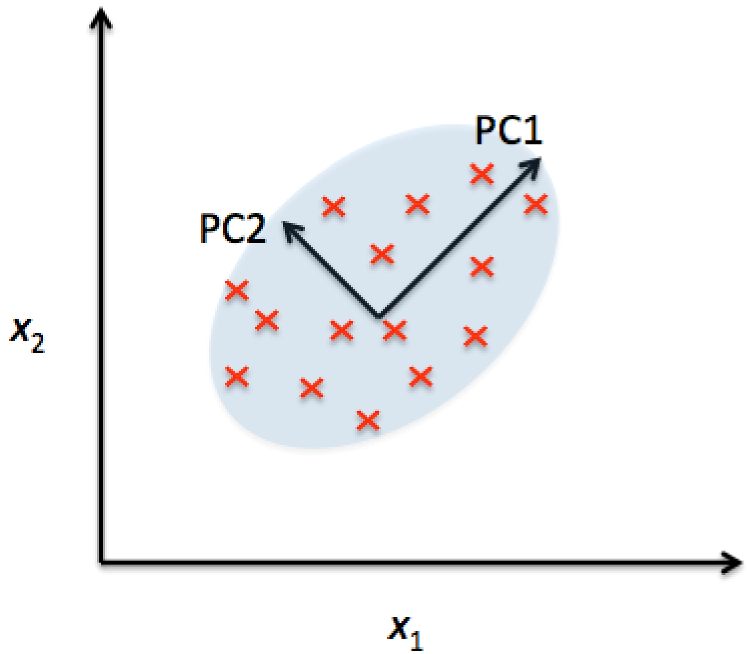
|r| ≦ 1 の証明
変数tを含む次の式：
Q = Σ[(xi-xm) + t(yi-ym)]^2
を展開すると、変数tに関する2次式が得られます。
Q = Σ(xi-xm)^2 + 2tΣ(xi-xm)(yi-ym) + t^{2}Σ(yi-ym)^2
Qの値は実数値の2乗和ゆえ、正または0です。
従って、上記2次式の判別式Dは負または0でなければなりません。
D = [Σ(xi-xm)(yi-ym)]^2 - Σ(xi-xm)^{2}Σ(yi-ym)^2 ≦ 0
[Σ(xi-xm)(yi-ym)]^2 ≦ Σ(xi-xm)^{2}Σ(yi-ym)^2
|Σ(xi-xm)(yi-ym)| ≦ sqrt[Σ(xi-xm)^2・Σ(yi-ym)^2]
|Σ(xi-xm)(yi-ym)|/sqrt[Σ(xi-xm)^2・Σ(yi-ym)^2] ≦ 1
よって、
-1 ≦ Σ(xi-xm)(yi-ym)/sqrt[Σ(xi-xm)^2・Σ(yi-ym)^2] ≦ 1
となり、相関係数の値が -1～+1の範囲であることが証明されました。

http://www.enjoy.ne.jp/~k-ichikawa/correlation2.html

主成分分析(principal component analysis:PCA)とは？

データの分散(ばらつき)が大きいところ(主成分)を見つける操作。つまり分散が大きいところが大事で、小さいところは気にしないようにする。

分散: ((xiとxmとの差)の2乗の合計の平均



http://yusuke-ujitoko.hatenablog.com/entry/2017/03/04/193628