【統計一口メモ　第 40 話】
＜マハラノビス距離とは？＞

名古屋市立大学大学院医学研究科 非常勤講師　薬学博士　松本一彦

マハラノビス距離って初めて聞いたという方も多いと思います。マハラノビスは 20 世紀前半で活躍したインドの統計数理学者で、インド統計研究所の設立者です。

マハラノビス距離とは、統計学において、データの相関関係を考慮した距離の概念です。通常の距離の概念であるユークリッド距離は、データのばらつきを一切考慮しませんが、マハラノビス距離は、データの相関関係を考慮した上で距離を算出します。実践では主成分分析や判別分析に使われている手法です。それでは、具体的に算出法をみていきましょう。例題は芳賀敏郎「多変量解析実務講座　テキストⅡ　実務教育研究所[1]」から抜粋しました。

§1. 基準化＝偏差値　*基準化は標準化ともいいます。

＜身長と体重から成人男子 A, B の体形を知りたい＞

身長と体重表

	身長cm x1	体重Kg x2
A	180	68
B	166	67
母平均	170	62
母標準偏差	8	5

手順１．身長と体重では単位が違うので、そのまま比較することができません。そこで、観測値と母平均の差＝偏差が母標準偏差の何倍あるかを求めて基準化することで比較することができます。それを u で表わすことにします。

$$u=\frac{x-\bar{x}}{s}$$

基準化＝偏差値

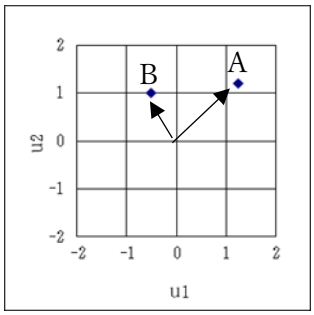
	身長				体重			
A	$u_1A=$	$\frac{180-170}{8}$	=	1.25	$u_2A=$	$\frac{68-62}{5}$	=	1.20
B	$u_1B=$	$\frac{166-170}{8}$	=	-0.50	$u_2B=$	$\frac{67-62}{5}$	=	1.00

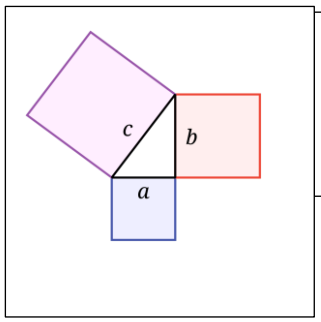
A は身長が 1.25、体重が 1.20 で両方とも1を超えていて体形としては大きい方であることがわかります。一方、B は身長が-0.50、体重が 1.00 でやや太めの体形であることがわかります。

§2. ユークリッド距離

そこで、この 2 人のどちらが成人男子の平均に近いかを調べることにします。

偏差値を図示する方法の一つ、ユークリッド距離は数値を平均（0）からの距離で求める方法です。





ピタゴラスの定理

$c^2=a^2+b^2$

中心から数値までの直線の距離をピタゴラスの定理を使って求めます。

$U^2=u_1^2+u_2^2$

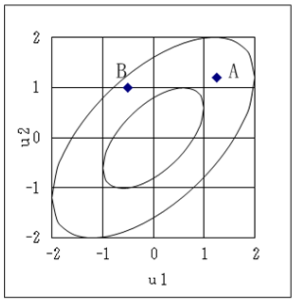
$U^2_A=1.25^2+1.20^2=3.00$

$U^2_B=(-0.50)^2+1.00^2=1.25$

ユークリッド距離では A の方が平均的成人男子の中心から離れているように見えます。

§3. マハラノビス距離

ユークリッド距離では中心に近いのは B であるという結果になりました。しかし、身長と体重には相関があります。そのような場合は、相関を考慮した距離であるマハラノビス距離を用いなければなりません。ちなみに、身長と体重の相関係数は 0.6 でした。



マハラノビスの距離は2つの要因に相関がある場合はデータが楕円に分布することを考慮して中心からの距離を求めるものです。言い換えるとマハラノビスの距離の中心から等しい点を結ぶと図のような等高線になります。

マハラノビス距離は次のような式から求めます。rは相関係数です

$$D^2=\frac{u_1^2+u_2^2-2ru_1u_2}{1-r^2}$$

$$D^2_A=\frac{1.25^2+1.20^2-2x0.6x1.25x1.20}{1-0.60^2}=1.88$$

$$D^2_B=\frac{(-0.50)^2+1.00^2-2x0.6x(0.50)x1.00}{1-0.60^2}=2.89$$

結果はユークリッド距離とは反対に、マハラノビス距離では B の方が中心から離れています。

上の式は2変量のみに使えるもので、変量が3以上になると下記のようにベクトルと行列で表わさなければなりません。

$$D^2=(u_1,\ u_2)\begin{pmatrix}1 & r\\ r & 1\end{pmatrix}^{-1}\begin{pmatrix}u_1\\ u_2\end{pmatrix}=\mathbf{u}^T\mathbf{R}^{-1}\mathbf{u}$$

式の中の $u_1,\ u_2$ を縦に並べたものは u のベクトルと呼びます。u^T はベクトル u を横にしたもので Transpose（転置）と呼びます。青の括弧でくくったものを行列（マトリックス）と呼びます。R^{-1} は R の逆行列と呼びます。この行列については、統計一口メモ第 39 話の「最小二乗平均」を求めるときに使いました。詳しくはそちらを参考にしてください。

§4.　2 変量マハラノビス距離

手順1. R（相関係数行列）に相関係数値を入力する

	x1	x2
R	1.000	0.600
	0.600	1.000

手順2. R^{-1}（逆行列）を求める。MINVERSE（R）

R^{-1}	1.5625	−0.9375
	−0.9375	1.5625

手順3. 二人の偏差値を縦ベクトル u として入力する

	A	B
u_1	1.250	−0.500
u_2	1.200	1.000

手順4. 二人の偏差値 u の転置行列 u^T を（TRANSPOSE）として求める

	x1	x2
A	1.250	1.200
B	−0.500	1.000

手順5.　u^T x R^{-1} を MMULT（u^T, R^{-1}）で求める

	x1	x2
A	0.828125	0.703125
B	−1.71875	2.03125

手順 6. マハラノビス距離 D^2 （$=u^TR^{-1}u$）を求める。（MMULT（u^T.R^{-1}),u）

	D^2				D^2
A	**1.879**	0.289		A	1.8789
B	0.289	**2.891**		B	2.8906

結果：相関を考慮したマハラノビス距離では B の方が中心から離れていることがわかる。

この計算方法は、変量の個数が 3 個以上の場合にもそのまま拡張できる。

1)　芳賀敏郎「エクセルによる多変量解析実務講座　テキスト II （財）実務教育研究所　2001