2007 年度後期　統計データ解析Ｂ　第１４回

データを分類する **(2)** － 判別分析

例えば，いま「健康とわかっている人たち」と「病気とわかっている人たち」のそれぞれの検査データがあるとします．さて，「健康か病気かわからない人」がやって来て検査をしたとき，そのデータからこの人が病気か健康かをどう判断すればよいでしょうか？ 判別分析は，この問題にひとつの答えを与えてくれます．今回は，１変量データ，２変量データの場合をあげて判別分析の原理を説明します．

判別分析の原理－１変量データの場合

判別分析は，２つのグループにあらかじめ分けられたデータがあるとき，新たに入ってきたデータがどちらのグループに属するデータかを判定する方法です．２つのグループに分けることはすなわち何かの行動を起こすかどうか，という意思決定に直結しますので，このような問題の例はさまざまな分野で見ることができます．例えば，病院での検査の結果からその人が病気かどうかを判別する，ある試料をしらべてそれがある物質かどうかを判定する，などいろいろ考えられます．

いま，１つの変量 X（例えば血圧）についてのたくさんのデータがあり，それが A，B の２つのグループ（例えば「健康」と「病気」，よく「健康群」「病気群」という表現をします)）に分かれているとします．このとき，新しい，どちらのグループに入るのかわからないデータ x が，どちらのグループに入るかをどう判断すればよいかを考えてみましょう．

ちょっと考えると，新しいデータと，両グループのデータの平均とのへだたりを調べ，近いほうのグループに分類すればよいと思われます．しかし，これでは不十分です．図１の数直線の例を見てみましょう．「新しいデータ（◎)」は，平均とのへだたりでいえば「健康（☆)」のグループに分類されることになります．しかし，これはどう見てもおかしいでしょう．◎は★の並んでいるところに入っていますから，「病気（★)」のグループに分類されるべきです．

これは，両グループの分布のしかたが大きく異なり，「病気」のグループのほうが分散がずっと大きいことが原因です．そこで，分散をとりいれた距離を定義して，これを使って「平均とのへだたり」を定めてみましょう．つまり，分散が大きい（小さい）グループの平均との隔たりは，実際よりも小さく（大きく）するわけです．

世の中の人を全て検査したわけではありませんから，★や☆が病気の人や健康な人の全てではありません．ですから，「病気」「健康」それぞれのグループには母集団があって，その分布は図２のようになっていると考えます．したがって，図１の★や☆のデータは，母集団分布と同じ確率分布にしたがって得られる標本と考えます．そこで，「病気」グループの母平均を μ_A，母分散を σ_A^2 とし，「健康」グループの

図 1: 「新しいデータ」は「健康」「病気」のどちらに分類すべきか？

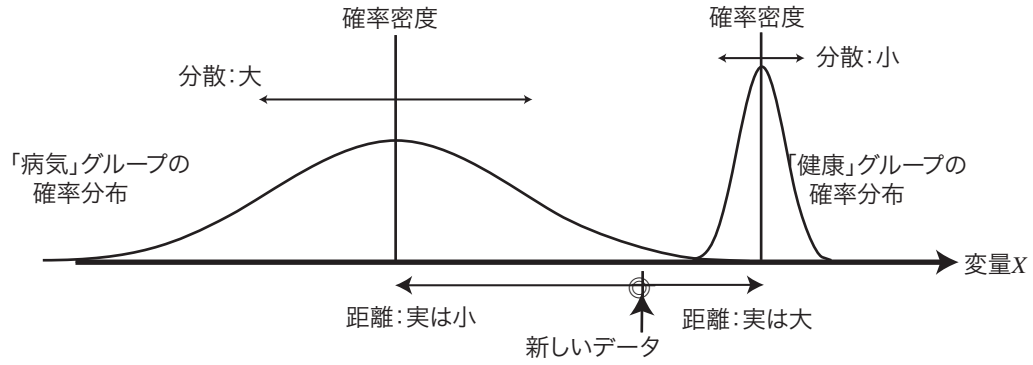

図 2: 両グループの確率分布と「距離」

母平均を μ_B，母分散を σ_B^2 としましょう．このとき，「新しいデータ」x と μ_A, μ_B との隔たり D_A^2, D_B^2 を，次のように定義します．

$$D_A^2 = \frac{(X-\mu_A)^2}{\sigma_A^2},\ D_B^2 = \frac{(X-\mu_B)^2}{\sigma_B^2} \tag{1}$$

上の式を見てわかるように，この隔たりは，グループの分散が大きい（小さい）ほど小さく（大きく）なることがわかります．この距離を用いて「新しいデータ」と両グループのそれぞれの平均との隔たりを求め，隔たりが小さいほうのグループに分類すればよいことになります．実際には母平均や母分散はわかりませんから，標本から求められる量で代用します．すなわち，母平均は標本平均で，母分散は不偏分散で代用します．

２変量データの場合は？ －多次元確率分布

では，「血圧」というひとつの変量だけではなく，２つの項目がある，つまり２変量の判別分析を考えてみましょう．例によって散布図上で考えます．２つの変量 X_1, X_2 の組によるデータがあり，それが A, B の２つのグループに別れているとします．このとき，ある値の組からなるデータは，２変量の確率分布にしたがって得られる標本と考えます．

さて，A，B の２つのグループの母集団分布が図 3 のように表されるとしましょう．図 2 の場合と同様に，それぞれのグループに属するデータは，それぞれの母集団の母集団分布と同じ確率分布にしたがって得られた標本と考えます．このとき，新しい（どちらのグループに入るかわからない）データ $x = (x_1, x_2)$ が，どちらのグループに入るべきかを考えてみましょう．

この場合も，新しいデータと，両グループの分布の中心との距離を求め，近いほうのグループに分類するのでは不十分です．図 4 のように，A グループと B グループとで分散が大きく違う場合を考えてみましょう．データ x から各グループの分布の中心への（ユークリッド）距離は同じです．しかし，B グループの確率分布は分布の中心の周りに集まっているため，データ x が B グループに属している可能性はほとんどゼロです．これに対して，A グループの確率分布は広くひろがっているので，このデータ x が A グループから得られる可能性はいくらかはあります．したがって，データ x は A グループに分類されるべきです．そうするために，データ x は B グループよりも A グループに「近い」と評価されるような「距離」を定義する必要があります．このような距離は，１次元の場合と同様，ユークリッド距離を分散で標準化することで定義できます．

ところが，２次元（以上）の場合は，さらに考慮すべき問題があります．図 5 の分布では b-d 方向の分散が大きく，a-c 方向の分散は小さくなっています．したがって，a, b, c, d の各点にあるデータは，分

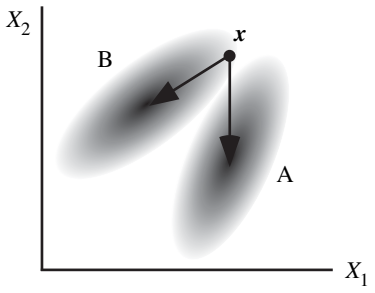

図 3: x は A, B どちらのグループに近いか？

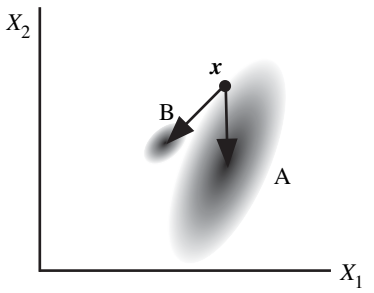

図 4: x は A に「近い」

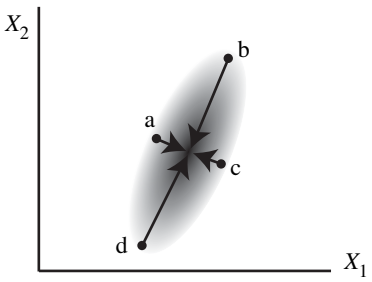

図 5: 分布の中心から「等距離」の点

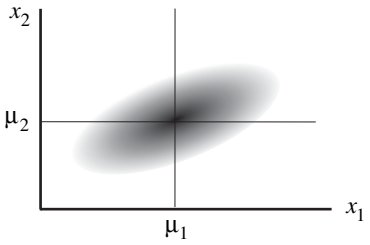

図 6: ２次元の確率分布

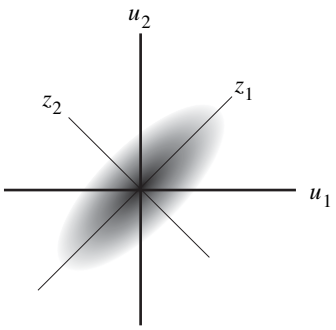

図 7: 主成分に変換

布の中心からの「距離」はいずれも同じになるように定義する必要があります．このように定義するためには，分布の形，すなわち変量 X_1, X_2 の相関係数（あるいは共分散）をも考慮する必要があります．

上で述べたような要求を満たす，「散布図上の点と分布の中心との『距離』」を定義してみましょう．あるグループの確率分布が，図 6 のように表されているとします．変量 X_1, X_2 の平均をそれぞれ μ_1, μ_2，分散をそれぞれ σ_1^2, σ_2^2 とし，X_1, X_2 の相関係数を ρ とします．ここで，簡単のために，変量 X_1, X_2 のデータ x_1, x_2 を

$$u_1 = \frac{x_1 - \mu_1}{\sigma_1},\quad u_2 = \frac{X_2 - \mu_2}{\sigma_2} \tag{2}$$

と標準化すると，平均は u_1, u_2 とも 0，分散はともに 1，相関係数は同様に ρ となります．

さらに，u_1, u_2 を互いに相関のない変量 $z_{(1)}, z_{(2)}$ に変換します．「互いに相関のない変量」とは，すなわち第１２回で取り上げた主成分です．標準化された２変量の主成分は相関係数によらず常に同じ[1]で，

$$z_{(1)} = \frac{u_1 + u_2}{\sqrt{2}},\quad z_{(2)} = \frac{u_1 - u_2}{\sqrt{2}} \tag{3}$$

となります（図 7）．

このように変換してしまうと，２つの変量の相関，すなわち「分布が散布図上でどちら向きに傾いているか」はもう考える必要がありません．そこで，散布図上のある 1 点 $(z_{(1)}, z_{(2)})$ と分布の中心（x_1，x_2 軸上では (μ_1, μ_2)，$z_{(1)}, z_{(2)}$ 軸上では $(0, 0)$）との「分散で標準化した」ユークリッド距離（平方距離）を，三平方の定理により，

$$D^2 = \frac{z_{(1)}^2}{V(z_{(1)})} + \frac{z_{(2)}^2}{V(z_{(2)})} \tag{4}$$

[1]付録参照．「相関係数によらず常に同じ」になるのは２次元の時だけで，３次元以上の場合はこうはなりません．

のように，$z_{(1)}, z_{(2)}$ 軸上でのそれぞれの分散 $V(z_{(1)})$，$V(z_{(2)})$ で標準化した平方距離の和で表します．この D^2 をマハラノビスの汎距離といいます．

$V(z_{(1)})$，$V(z_{(2)})$ は，標準化した場合の主成分分析における $z_{(1)}, z_{(2)}$ の固有値，すなわち $z_{(1)}, z_{(2)}$ の各軸上の分散で，付録にあるように

$$V(z_{(1)}) = 1 + \rho, \quad V(z_{(2)}) = 1 - \rho, \tag{5}$$

です．これと (2) 式を使って $z_{(1)}, z_{(2)}$ を u_1，u_2 に戻すと，

$$\begin{aligned} D^2 &= \frac{z_{(1)}^2}{V(z_{(1)})} + \frac{z_{(2)}^2}{V(z_{(2)})} = \frac{\left(\frac{u_1+u_2}{\sqrt{2}}\right)^2}{1+\rho} + \frac{\left(\frac{u_1-u_2}{\sqrt{2}}\right)^2}{1-\rho} \\ &= \frac{(1-\rho)(u_1+u_2)^2 + (1+\rho)(u_1-u_2)^2}{2(1+\rho)(1-\rho)} \\ &= \frac{\{(1-\rho)+(1+\rho)\}(u_1^2+u_2^2) + \{(1-\rho)-(1+\rho)\}2u_1u_2}{2(1+\rho)(1-\rho)} \\ &= \frac{2(u_1^2+u_2^2) - 2\rho\cdot 2u_1u_2}{2(1-\rho^2)} = \frac{u_1^2+u_2^2-2\rho u_1u_2}{1-\rho^2} \end{aligned} \tag{6}$$

が得られます．

判別の方法

マハラノビスの汎距離を使うと，あるデータがグループ A，B のどちらに属するかは「A，B それぞれに対応する確率分布の中心とのマハラノビスの汎距離が短いほうのグループに属する」という基準で判別されます．ただし実際の判別では，グループ A，B の確率分布は通常わかりません．そこで，グループ A，B の母平均・母分散を，すでにわかっているグループ A，B に属するデータから得られる推定量で代用します．すなわち，n をデータ数（標本サイズ）として（データ数はグループ A，B で違っていてもかまいません），i 番目のデータの変量 x_1 の値を $x_{1,i}$，変量 x_2 の値を $x_{2,i}$ とするとき，母平均 μ_1, μ_2 を

$$\mu_1 \leftarrow \bar{x_1} = \frac{1}{n}\sum_{i=1}^{n} x_{1,i}, \quad \mu_2 \leftarrow \bar{x_2} = \frac{1}{n}\sum_{i=1}^{n} x_{2,i} \tag{7}$$

のように標本平均で代用します．また，分散 σ_1^2, σ_2^2 や共分散 σ_{12}，相関係数 ρ を

$$\begin{aligned} \sigma_1^2 &\leftarrow s_{11} = \frac{1}{n-1}\sum_{1} n(x_{1,i}-\mu_1)^2, \quad \sigma_2^2 \leftarrow s_{22} = \frac{1}{n-1}\sum_{1} n(x_{2,i}-\mu_1)^2 \\ \sigma_{12} &\leftarrow s_{12} = \frac{1}{n-1}\sum_{1} n(x_{1,i}-\mu_1)(x_{2,i}-\mu_1), \quad \rho = \frac{\sigma_{12}}{\sigma_1\sigma_2} \leftarrow r_{12} = \frac{s_1 2}{\sqrt{s_{11}}\sqrt{s_{22}}} \end{aligned} \tag{8}$$

のように，標本から求められる不偏分散・共分散を使って代用します（不偏分散を用いるので $n-1$ で割ることに注意して下さい）．

グループ A, B それぞれについて，すでにわかっているデータからこれらの値を事前に計算しておけば，A, B どちらに属するかわからない「新しいデータ」が A, B どちらのグループに判別されるかは次のようにして求められます．すなわち，A, B それぞれのグループについて，(7)(8) 式で計算される μ_1, μ_2，σ_1, σ_2 を用いて，(2) 式で新しいデータの x_1, x_2 の値を使って u_1, u_2 を求めます．さらに，(6) 式によってマハラノビスの汎距離を求めます．グループ A，グループ B それぞれについて求めた汎距離をそれぞれ

D_A^2, D_B^2 とすると，「新しいデータ」は

$$D_A^2 - D_B^2 < 0 \Rightarrow \text{グループ } A \text{ に判別}$$
$$D_A^2 - D_B^2 > 0 \Rightarrow \text{グループ } B \text{ に判別} \tag{9}$$

のように，A，B のうち汎距離が小さいほうのグループに属すると判定します．

≡ ∧∧ ≡
(・ ・)〜

「判別の方法」で，「母平均・母分散を，標本平均・不偏分散で代用する」と書いてありますけど，そんなにうまくいくんですか？

もっともな疑問やね．区間推定という方法があるように，母平均や母分散を，標本からそう正確には推定できるものやない．それを何とかするために，今でも判別分析はいろいろな方法が研究されてるんや．

Ξ ∧◆∧ Ξ
(o- o)〜

今日の演習

「血圧」と「コレステロール」の２つの検査項目で，ある病気の診断をするとします．「健康群」「病気群」のそれぞれの検査項目の平均，不偏分散，相関係数は次の通りでした．

健康群：血圧の平均 100・分散 100，コレステロールの平均 150・分散 144，相関係数 0.7
病気群：血圧の平均 150・分散 169，コレステロールの平均 200・分散 225，相関係数 0.8

さて，ある人を検査すると血圧 120，コレステロール 160 でした．この人はどちらの群に判別すべきかを答えてください．［「健康群」「病気群」が上のグループ A, B に，「血圧」「コレステロール」が上の変量 1, 2 に相当します］

付録：標準化された２変量の主成分

２変量のとき，標準化された変量 u_1, u_2 については，分散はどちらも 1 であり，共分散はすなわち相関係数 ρ となります．このとき，固有値問題

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \lambda \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \tag{A1}$$

を解くと，特性方程式は

$$(1-\lambda)^2 - \rho = 0 \tag{A2}$$

となり，固有値は

$$\begin{aligned} \lambda &= 1 \pm \sqrt{1-(1-\rho^2)} \\ &= 1 \pm \rho \end{aligned} \tag{A3}$$

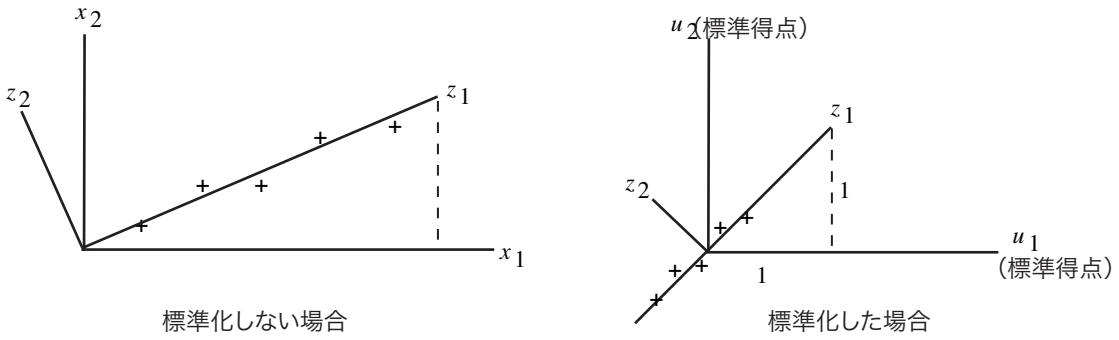

図 A1: 標準化されたデータの主成分分析

となります．すなわち両主成分 $z_{(1)}, z_{(2)}$ の分散 $V(z_{(1)}), V(z_{(2)})$ は $1 \pm \rho$ となります．また，(A1) から導かれる方程式

$$(1 - \lambda)u_1 + \rho u_2 = 0 \tag{A4}$$

に (A3) 式の結果を代入すると，

$$\begin{aligned} (1 - \lambda)^2 - \rho &= 0 \\ (1 - (1 \pm \rho))u_1 + \rho u_2 &= 0 \\ \rho(\mp u_1 + u_2) &= 0 \ \text{ゆえに } u_2 = \pm u_1 \end{aligned} \tag{A5}$$

となります．$u_1^2 + u_2^2 = 1$ の関係を考慮すると，求められる主成分 z_1, z_2 は常に

$$\frac{u_1 + u_2}{\sqrt{2}}, \ \frac{u_1 - u_2}{\sqrt{2}} \tag{A6}$$

の２つになります．どちらが第 1 主成分か，および寄与率は，相関係数の正負・値によります．

これは当然といえば当然で，元の変量が標準化された結果どちらも同じ分散（すなわち 1）に圧縮（伸長）されていますから，図 A1 のように主成分は常に右上 45 度・左上 45 度の２つの直線になります．ただし，このようになるのは２変量の場合だけで，変量の個数が３つ以上の場合は，それぞれの変量間の相関が問題になるため，こんなに簡単にはなりません．