判別分析演習問題解答

作成:来島 愛子

1　対象データ

表 1 の 2 つの群 (汚染地区と非汚染地区) からなるデータが得られている．このデータに対して，新しいデータ (表 2) が汚染地区，非汚染地区どちらに属するか判別分析によって決定する．

表 1: 対象データ

I. 汚染地区			II. 非汚染地区		
SO2	D-D	NOX	SO2	D-D	NOX
0.010	0.078	0.210	0.017	0.078	0.000
0.015	0.073	0.110	0.013	0.047	0.050
0.024	0.092	0.200	0.008	0.082	0.040
0.014	0.070	0.080	0.030	0.021	0.010
0.025	0.113	0.230	0.050	0.040	0.020
0.015	0.064	0.090	0.060	0.035	0.030
0.007	0.063	0.100	0.008	0.045	0.140
0.016	0.075	0.080	0.009	0.074	0.040
0.011	0.072	0.090	0.007	0.039	0.050
0.025	0.074	0.140	0.007	0.059	0.050
			0.010	0.042	0.060
			0.120	0.060	0.110

表 2: 新しいデータ

SO2	D-D	NOX
0.09	0.069	0.12

2　解析結果

まず，各地区と全体の平均値，分散共分散行列とその逆行列を求める．

I. 汚染地区			II. 非汚染地区		
SO2	D-D	NOX	SO2	D-D	NOX
0.016	0.077	0.133	0.028	0.052	0.050

SO2	D-D	NOX
0.023	0.063	0.088

	I. 汚染地区の分散共分散行列				I. 汚染地区の分散共分散行列の逆行列		
	SO2	D-D	NOX		SO2	D-D	NOX
SO2	0.00004	0.00006	0.00017	SO2	51699.469	-18545.794	984.591
D-D	0.00006	0.00020	0.00064	D-D	-18545.794	21437.981	-3404.288
NOX	0.00017	0.00064	0.00308	NOX	984.591	-3404.288	972.949
	II. 非汚染地区の分散共分散行列				II. 非汚染地区の分散共分散行列の逆行列		
	SO2	D-D	NOX		SO2	D-D	NOX
SO2	0.00106	-0.00008	0.00026	SO2	1005.718	229.859	-177.462
D-D	-0.00008	0.00033	-0.00001	D-D	229.859	3122.613	-17.622
NOX	0.00026	-0.00001	0.00145	NOX	-177.462	-17.622	721.140
	プールした分散共分散行列				プールした分散共分散行列の逆行列		
	SO2	D-D	NOX		SO2	D-D	NOX
SO2	0.00060	-0.00002	0.00022	SO2	1759.155	327.758	-217.894
D-D	-0.00002	0.00027	0.00028	D-D	327.758	4357.819	-591.774
NOX	0.00022	0.00028	0.00218	NOX	-217.894	-591.774	555.578

2.1 正規性の検定

各地区のデータについて正規性の仮定が成り立つか P-P プロットによって確かめる．

n 個の p 次元データ $x_i (i = 1, 2, \ldots, n)$ が p 次元正規分布をもつ母集団から得られる無作為標本であるという仮説 H を検定することを考える．

このとき，平均値 $\overline{x}$，標本分散共分散行列を S とすると，点 $\overline{x}$ から各点 x_i までの距離

$$d_i = (x_i - \overline{x})' S^{-1} (x_1 - \overline{x})$$

は仮説 H のもとで，近似的に自由度 p の χ^2 分布からの標本値とみなしてよいことが知られている．

P-P プロットを描く手順は次のとおりである．

1. $d_1, d_2, \ldots, d_n$ を大きさの順に並び替えたものを $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(n)}$ とする．
2. $F(d_{(i)})$ を求める．ただし，$F(\cdot)$ は自由度 p のカイ 2 乗分布の累積分布関数である．
3. n 個の点

$$\left(\left(i - \frac{1}{2} \right) / n, F(d_{(i)}) \right)$$

をプロットする．

以上の手順によって，得られた P-P プロットは汚染地区，非汚染地区ともに正規分布を表す直線の近くに並んでおり，正規分布に従うとみなしてよいと考えられる．

2.2 等分散の検定

次に各地区の分散が等しいかどうか検定する．各群, 全体の分散共分散行列をそれぞれ, $\Sigma^{(1)}, \Sigma^{(2)}, \Sigma$ として，等分散の仮定が成り立つか次の検定を行う．

$$\text{帰無仮説 } H_0 : \Sigma^{(1)} = \Sigma^{(2)},$$

$$\text{対立仮説 } H_1 : \Sigma^{(1)} \neq \Sigma^{(2)}.$$

この検定で用いるのはそれぞれ n_1，n_2 個のデータからの標本値であり，それらを $\hat{\Sigma}^{(1)}$，$\Sigma^{(2)}$ とする．このとき，仮説 H_0 のもとで,

$$W = \frac{|\hat{\Sigma}^{(1)}|^{\frac{n_1}{2}}|\hat{\Sigma}^{(2)}|^{\frac{n_2}{2}}}{|\hat{\Sigma}|^{\frac{n}{2}}}, \quad (n_1 + n_2 = n)$$

ただし,

$$\hat{\Sigma} = \frac{1}{n-2}\{(n_1-1)\hat{\Sigma}^{(1)}(n_2-1)\hat{\Sigma}^{(2)}\}$$

とおくと,

$$\chi^2_W = -2\log_e W$$

が近似的に自由度 $p(p+1)/2$ の χ^2 分布に従うことが知られている．今回のデータから計算される値は

$$\chi^2_W = -2\log_e W = 36.89$$

であり，自由度は 6 であるので，有意水準 5 %として上側 5 %点は $\chi^2_0 = 12.59$ で $\chi^2_W > \chi^2_0$ となり帰無仮説 H_0 は棄却される．

従って，$\Sigma^{(1)} \neq \Sigma^{(2)}$ とみなして，マハラノビスの距離を用いた判別方式を用いる．

2.3 マハラノビスの距離を用いた判別

新しいデータ x がどちらの地区に属するかマハラノビスの距離で判別する．新しいデータの汚染地区に対するマハラノビスの距離 D_1 は

$$D_1^2 = (x - \overline{x_1})'\hat{\Sigma}_1^{-1}(x - \overline{x_1})$$

で表される．非汚染地区に対するマハラノビスの距離 D_2 も同様に求められる．そして，$D_1 > D_2$ ならば I に，$D_1 < D_2$ ならば II に判別されることになる．データから計算すると,

$$D_2^2 - D_1^2 = -269.794$$

となり，非汚染地区に判別される．