【統計一口メモ　第41話】

＜判別分析とは？＞

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

「判別分析」は多変量解析として臨床試験ではよく見かけますが、毒性試験や薬理試験ではめったに目にすることはありません。それでも、ある値が2つの群のうちどちらに属するのかを調べなければならないこともあります。今回は臨床でのガン診断の例で「判別分析」を実施してみましょう。

例題　癌バイオマーカーの ILX 値と IL-Y 値には相関があることが知られている（仮想例）。非癌患者 10 名と癌患者 10 名の IL-XとILY測定値から、新規患者の癌診断を行いたい。

涌井良幸、涌井貞美：実習多変量解析入門　技術評論社　データ利用および一部改変

表1

行/列　　C　　D

	非癌			癌	
番号	IL-X	IL-Y	番号	IL-X	IL-Y
1	80	26	11	98	45
2	82	27	12	96	47
3	84	29	13	90	39
4	82	33	14	89	36
5	87	30	15	86	33
6	84	38	16	95	48
7	91	35	17	95	39
8	85	31	18	92	36
9	81	34	19	88	36
10	85	28	20	94	46
平均	84.1	31.1	平均	92.3	40.5
分散	10.32	14.77	分散	15.34	30.06
標準偏差	3.21	3.84	標準偏差	3.92	5.48

5

新規患者	
IL-X	IL-Y
90	30

§1. 判別分析の種類

判別分析は目的変数がカテゴリカルデータ（癌か非癌）で説明変数が数量値（検査データ）のときに適用される手法です。癌患者と非癌患者から集められたデータから、判別関係式を作成して、患者の癌の有無を判別（予測）します。

判別分析手法には＜線形判別分析＞と＜マハラノビス距離分析＞があります。さらに後者のマハラノビス距離分析手法には①相関係数を使用する手法と②分散・共分散を使用する手法があります。JMPソフトを含め、一般的には線形判別分析が使われているので、その手法を知ることと、さらにマハラノビス距離で②の分散・共分散を使用する手法にチャレンジしてみましょう。

§2. 線形判別分析

癌と非癌の重なったデータを2分するために使われるのが**線形判別関数**という1次式です。

$z = ax + by + c$　　(a, b, c は定数で a, b を**判別係数**といいます)

2群ができるだけ離れていることが理想です。そのために、「**相関比 η^2**(イータ)」を用います。

相関比とは2つの群の離れ具合を示す指標です。相関比が大きいほど 2 群は離れていることを表します。

各群のデータにはバラツキ(変動)があります。その内容は次のように分類されます。

① 各群の中にある変動＝群内変動

② 2 群間の変動＝群間変動

③ ①と②を合わせた全変動

全変動 S_T	
群間変動 S_B	群内変動 S_w

相関比は全変動 S_T の中に占める群間変動 S_B の割合です。

$$\eta^2 = S_B / S_T$$

相関比は、線形判別関数　$z = ax + by + c$ を決定する道具でもあります。言い換えると相関比が最大になるように、定数 a と　b を 決めることになります。それにはエクセルの**ソルバー**機能を用います。求めた z は**判別得点**として表します。それでは判別得点を求める手順を見ていきましょう。

手順1. 判別係数、定数項を 1.00 として仮設定する。

行/列	C	D	E
	a	b	c
18	1.000	1.000	1.000

手順2. 判別得点を求める。

エクセル関数 SUMPRODUCT(配列 1、配列2)を用いる。

表2

行/列	C	D	E
	非癌		癌	
	番号	判別得点	番号	判別得点
23	1	107	11	144
	2	110	12	144
	3	114	13	130
	4	116	14	126
	5	118	15	120
	6	123	16	144
	7	127	17	135
	8	117	18	129
	9	116	19	125
32	10	114	20	141

例：患者番号1の 107 は SUMPRODUCT(C18:D18,C5:D5)+E18

手順3．仮の判別得点から群間変動と全変動を求める。

① 非癌変動：DEVSQ(C23:C32)＝299.6

② 癌変動： DEVSQ(E23:E32)＝731.6

③ 全変動(S_T)： DEVSQ(C23:C32, E23:E32)＝2580.0

④ 群内変動(S_W)：非癌変動＋癌変動＝299.6＋731.6＝1031.2

⑤ 群間変動(S_B)：全変動－群内変動＝2580.0－1031.2＝1548.8

表にする。

行/列	G	H	I	J	K
	非癌	癌	全変動	群内変動	群間変動
23	299.6	731.6	2580	1031.2	1548.8

手順4．相関比を求める

相関比 $\eta^2 = S_B / S_T = 1548.8/2580.0 = 0.60$

行/列		H
25	相関比	0.60

手順5．各群の判別得点の平均とその和を求める。

行/列		C	D
34	平均	非癌	116.2
35		癌	133.8
36		合計	250.0

手順6．ソルバーを使うためのパラメータを設定する。

① 目的セル：H25　　相関比が最大になるように設定

② 変数セル：C18:E18　定数 a, b, c　を設定。

③ 制約条件の対象： D36＝0　　平均の合計

G23＝20　全変動＝個体数20

目標値＝最大値

□「制約のない変数を非負数にする」はチェックを入れない。

手順7．ソルバーを「データ」「分析」から立ち上げる。

① 判別係数と定数項は次のように算出される。

a	b	c
0.133	0.050	-13.542

② 判別得点平均は次のように算出され「非癌」はマイナス(－)表示となる。

平均	非癌	-0.8
	癌	0.8
	合計	0.0

手順8．線形判別関数を作成する。

z＝0.13x+0.05y-13.54

手順9. 新規患者の分類

IL-X＝90， IL－Y＝30 を線形判別関数zにあてはめるとz＝-0.34 となり、「非癌」に分類される。

※線形判別分析では「**判別的中率**」（＝正しく判定された個体数/全個体数）が表示されることも多い。それは患者ごとに z が計算されるような場合に限られ、今回のような単独例では計算されない。

§3 分散・共分散を用いるマハラノビス距離判別分析

Excel でマハラノビス距離を計算する

手順1. それぞれの群で「共分散」を求める。

共分散 Covar(x,y)=S_{XY} は次のような式で表される

$$Sxy=\frac{(x_1-\bar{x})(y_1-\bar{y}) + (x_2-\bar{x})(y_2-\bar{y}) + \text{-------} + (x_n-\bar{x})(y_n-\bar{y})}{n}$$

エクセル関数「COVAR（配列1，配列2）」で非癌と癌それぞれで求める。

非癌＝3.69　　癌＝16.15

手順2. 分散・共分散行列を作成する

表 1 の分散と上記共分散値から次のような行列を作成する。

分散・共分散行列			
非癌		癌	
10.32	3.69	15.34	16.15
3.69	14.77	16.15	30.06

手順3. 分散・共分散行列の逆行列を求める。

逆行列は MINVERSE（行列範囲）で求める。

分散・共分散逆行列			
非癌		癌	
0.106	-0.027	0.150	-0.081
-0.027	0.074	-0.081	0.077

手順4. 偏差データ（観測値―平均値）を算出する。

新規患者の観測値

IL-X＝	90
IL-Y＝	30

	IL-X	IL-Y
非癌	84.1	31.1
癌	92.3	40.5

非癌偏差		癌偏差	
IL-X	IL-Y	IL-X	IL-Y
5.9	-1.1	-2.3	-10.5

手順5. 新規患者の判別分析

＝MMULT(MMULT(偏差の範囲、逆行列の範囲)、TRANSPOSE(偏差の範囲))

手順5の偏差の範囲、手順4の逆行列の範囲を使用する。

マハラノビス距離	
非癌	癌
4.14	5.34

マハラノビス距離で小さい方の値の群を判定群とする。

結果：新規患者は非癌群に判別する。この結果は§2. 線形判別分析の結果と一致する。

※ボクのつぶやき：マハラノビス距離を用いて判別分析をする方法に①相関係数を用いる方法もあるが、今回のように 2 群それぞれのn数が 10 例程度では相関係数を求めることは適切ではない。過去のデータからの類推では正確さに乏しいケースもあり、使用は限定される気がする。

1)涌井良幸、涌井貞美：実習多変量解析入門　技術評論社 2018 年