３．判別分析　【動画】

重回帰分析では量的データの目的変数を複数の説明変数で直接予測するものでしたが、判別分析は質的な目的変数の分類を複数の説明変数で予測する手法です。質的データにはその値自身に意味はありませんので、直接値を予測するのではなく、判別関数という 1 次式を求め、その値によって分類を予測します。その際、予測の精度やどの変数の影響が強いかなどが検討されます。

今、試験の合否をそれまでの平均的な勉強時間と模擬試験の成績で判定するとします。合格、不合格という 2 群の場合の判別関数は、以下のように与えられます。

判別関数　$z=b_1$ 勉強時間＋b_2 平均点＋b_0

ここに、b_1, b_2, b_0 は分析結果として決まるパラメータです。この判別関数の値が 0 以上か 0 未満かで合否を判別します。以下の図を見て下さい。

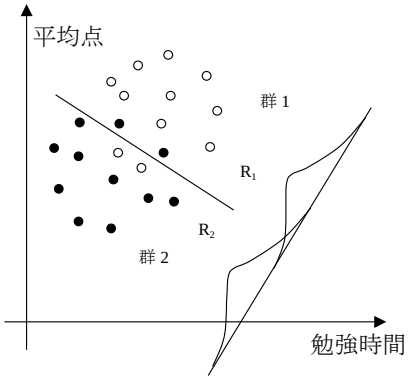


図 1　判別分析での群分け

白丸を合格群、黒丸を不合格群として、その間に線を引き、その線より上を合格、下を不合格と判定することを考えます。その線が判別関数 $z=b_1$ 勉強時間＋b_2 平均点＋$b_0=0$ を表しています。これを理論的にうまく引いてやることが大切です。実際の分析を、例を使ってパソコンで見てみましょう。

例

入学試験の合否と勉強時間・模擬試験の平均点のデータを求めたところ以下のような結果を得た（判別分析.txt (p1)）。合否を判定するための勉強時間と平均点の 1 次関数を求めよ。またこの関数によってデータを判別し、誤判別の確率を求めよ。

合否	勉強時間	平均点	合否	勉強時間	平均点
1	5.6	70.2	2	3.8	67.4
1	5.9	74.2	2	3.8	61.3
1	4.1	72.7	2	1.7	60.6
1	5.1	84.9	2	2.7	77.2
1	5.0	93.0	2	4.3	65.9
:	:	:	:	:	:
1	3.6	85.5	2	2.5	64.4
2	3.8	47.9	2	5.2	50.7
2	3.9	70.8	2	2.2	65.7

C.Analysis のメニュー［分析－多変量解析他－判別手法－判別分析］を選択すると以下の分析実行メニューが表示されます。

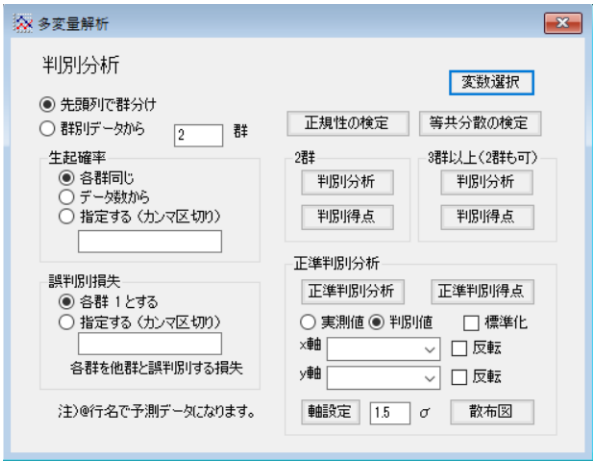
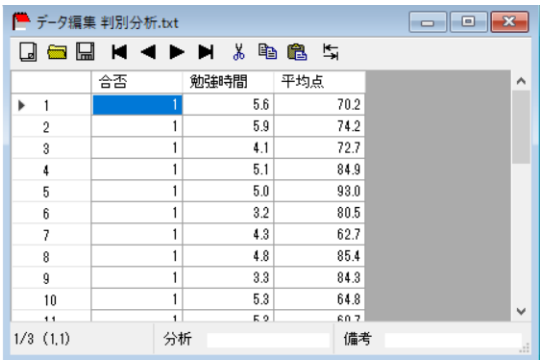


図 2　判別分析実行画面

判別分析.txt (p1) は以下の通りです。ここで、「合否」の合格は 1、不合格は 2 にしています。



	合否	勉強時間	平均点
1	1	5.6	70.2
2	1	5.9	74.2
3	1	4.1	72.7
4	1	5.1	84.9
5	1	5.0	93.0
6	1	3.2	80.5
7	1	4.3	62.7
8	1	4.8	85.4
9	1	3.3	84.3
10	1	5.3	64.8

図 3　判別分析.txt (p1)

「変数選択」は先頭に「合否」を指定して、「先頭列で群分け」にすればよいので、「All」ですべてを選びます。群は合否の 2 群なので、2 群のところの「判別分析」ボタンをクリックすると、以下の結果が表示されます。

	判別関数	標準化係数	F検定値	自由度	確率値
勉強時間	2.246	2.621	19.882	1,27	0.0001
平均点	0.201	2.279	15.027	1,27	0.0006
定数項	-23.019	-0.379			
マハラノビスの距離	5.682				
誤判別確率	1群を2群と	2群を1群と			
理論から	0.117	0.117			
実測から	0.077	0.059			
判別数・確率	1群予測	2群予測	1群予測	2群予測	
1群実測	12	1	0.923	0.077	
2群実測	1	16	0.059	0.941	

図 4　分析結果

判別関数の係数は、分析結果の「判別関数」のところです。結果への影響の強さは、隣の

「標準化係数」で分かります。重回帰分析のところでも述べたように、標準化係数は絶対値が大きいほど重要です。ここではほぼ同じですが、少し勉強時間の影響が大きいようです。また、これらの変数の係数（影響）が 0 かどうかも判定できます。結果の右端の確率の値が有意水準 0.05 より小さい場合、その変数の係数は 0 と異なると判断され、判定に影響がある変数になります。ここでは、共に有意水準（0.05）より小さいので必要な変数です。

次に判定の正確性ですが、これは「実測値から」の「誤判別確率」がよく使われます。ただ判別関数の係数は、その上にある「理論から」の値が小さくなるようにして決められています。「マハラノビスの距離」というのが 2 つの群の分布の中心間の距離を表しており、これが大きくなるほど、誤判別確率の理論値が小さくなります。その概略は以下の表で与えられます。

表 1　マハラノビス距離と誤判別確率

マハラノビス距離	1	4	9	16	25
誤判別確率	0. 309	0. 159	0. 067	0. 023	0. 006

理論は少し横において、我々は実測値からの値に注目して行きましょう。例えば、合格（1 群）を不合格（2 群）と誤判別する確率は 1/13=0.077、その逆の確率は 1/17=0.059 となります。これは判別得点の結果を見ても分かります。分析実行画面の 2 群の「判別得点」をクリックすると以下の結果になります。

	所属群	判別得点	判別群
1	1	3.651	1
2	1	5.128	1
3	1	0.784	1
4	1	5.479	1
5	1	6.880	1
6	1	0.328	1
7	1	-0.774	2
8	1	4.905	1

図 5　判別得点

元々の所属群と判別得点と判別群が表示されています。この中で誤判別は、所属群が 1 群なのに判別得点が負の値となり、2 群と判別された 7 番です。

これは少し進んだ話になりますが、判別分析では、予め 2 つの群に差を付けて判別する方法もあります。実行画面の「判別の生起確率」は元々の判別群のデータ数が大きく異なる場合に利用します。例えば、非常に難しい試験の場合、合格群は不合格群に比べて小さくなります。つまり、合格の確率は不合格の確率に比べて小さくなります。このような場合、データ数に応じて、判別に補正を加えます。また、誤判別した場合の損失（ダメージ）が大きく異なるような場合も、判別に補正を加えることが考えられます。これは実行画面の「誤判別損失」のところで設定することができます。これらについてはここまでにしておきます。

今までは 2 群の場合の話をしましたが、3 群以上（2 群も含む）の場合には、各群に対応した判別関数を作ります。即ち、3 群の場合は 3 種類の判別関数を作ります。群の判別は、これらの判別関数の中にデータを代入して、最大の値となった群に属するものと判断しま

す。特に 2 群の場合は、2 つの判別関数の大きい方の群に属するものとしますが、これまでの 2 群の場合の判定は、2 つの判別関数の差を取って、値が正になるか負になるかで大きい方の群を判別していたわけです。3 群の判別を、昔フィッシャーが使ったあやめの分類の話を例に示しておきましょう。判別分析.txt (p3) を見て下さい。

図 6　判別分析.txt (p3)

このデータの群は 3 つです。分析実行画面の結果は以下となります。

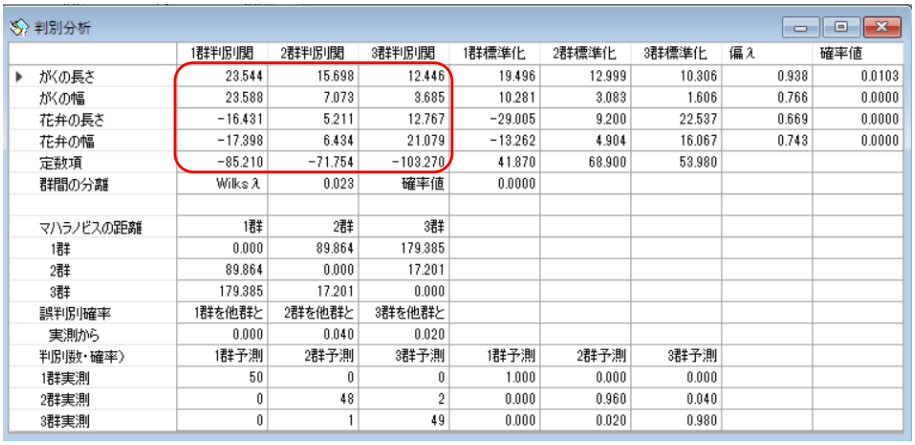


図 7　3 群の分析結果

「判別得点」をクリックすると結果は以下の通りです。

図 8　3 群の判別得点

間違った 134 番の判別関数では 2 群が 83.177 で一番大きくなっています。

最後に判別分析についてまとめておきましょう。

判別分析まとめ

判別分析の目的

2 群（多群）を判別する最適な 1 次式を求める。

$z = b_1$ 勉強時間 $+ b_2$ 平均点 $+ b_0$　　判別関数

判別関数の係数は？　　→　判別関数の欄

判別関数で群を分けるのは？　　→　判別の分点 0 （多群の場合値が最大の群）

判定に影響を与える変数は？　　→　標準化係数の絶対値の大きい変数

各係数の有効性は？（要正規性・等共分散性）→　確率の欄（係数が 0 と異なるかの検定）

誤判別の程度は？　　→　誤判別確率（実測と理論）（理論値は要正規性・等共分散性）

マハラノビス距離とは　　→　どの程度 2 群が離れているかを表わす指標

マハラノビス距離	1	4	9	16	25
誤判別確率	0.309	0.159	0.067	0.023	0.006

データ毎の判別関数の値と判別状況　　→　判別得点

問題 1

判別分析.txt (p2) は、適性の有無の判定（有：1，無：2）と適性検査の結果とＳＰＩの結果を与えたデータである。判定を適性検査とＳＰＩで予測する判別分析を行い、以下の問いに答えよ。但し、事象の生起確率は各群同じ、誤判別損失は２群とも１とすること。

１）判別関数を求めよ。

判別得点＝［　　］適性検査＋［　　］ＳＰＩ＋［　　］

２）どちらの変数が判定に影響があると思われるか。[適性検査・ＳＰＩ]

３）実測値から求めた誤判別の確率は？

適性有りを無しと［　　］　適性無しを有りと［　　］

４）先頭（1 番）の人の判別得点はいくらか。［　　］

５）適性検査 50 点，ＳＰＩ 55 点の人の判別得点はいくらか、またその人の適性の有無を判定せよ。　　判別得点［　　］　適性［有り・無し］

問題２

判別分析.txt (p3) はあやめの種類をがくの長さと幅、花弁の長さと幅で３群に分類したデータである。あやめの群を他の変数の１次式で判別する３群以上の判別分析を行い、以下の問題に答えよ。但し、設定は前問と同じとする。

１）３つの判別得点の式を求めよ。

判別得点１＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］

判別得点２＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］

判別得点３＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］

２）実測値から求めた誤判別確率はいくらか。

群１を他と［　　］群２を他と［　　］群３を他と［　　］

３）先頭のデータの３つの判別得点を求めよ。

判別得点１［　　］判別得点２［　　］判別得点３［　　］

これは何群に予想されたか。　［1群・2群・３群］

問題１解答（判別分析.txt (p2)）

１）判別関数を求めよ。

判別得点＝［ -0.190 ］適性検査＋［ 0.645 ］ＳＰＩ＋［ -20.467 ］

２）どちらの変数が判定に影響があると思われるか。［適性検査・ＳＰＩ］

３）実測値から求めた誤判別の確率は？

適性有りを無しと［ 0.053 ］　適性無しを有りと［ 0.095 ］

４）先頭（1番）の人の判別得点はいくらか。［ -5.118 ］

５）適性検査 50 点，ＳＰＩ 55 点の人の判別得点はいくらか、またその人の適性の有無を判定せよ。　判別得点［ 5.508 ］　適性［有り・無し］

問題２解答（判別分析.txt (p3)）

１）３つの判別得点の式を求めよ。

判別得点１＝［ 23.544 ］がくの長さ＋［ 23.588 ］がくの幅
＋［ -16.431 ］花弁の長さ＋［ -17.398 ］花弁の幅＋［ -85.210 ］

判別得点２＝［ 15.698 ］がくの長さ＋［ 7.073 ］がくの幅
＋［ 5.211 ］花弁の長さ＋［ 6.434 ］花弁の幅＋［ -71.754 ］

判別得点３＝［ 12.446 ］がくの長さ＋［ 3.685 ］がくの幅
＋［ 12.767 ］花弁の長さ＋［ 21.079 ］花弁の幅＋［--103.270 ］

２）実測値から求めた誤判別確率はいくらか。

群１を他と［ 0 ］群２を他と［ 0.040 ］群３を他と［ 0.020 ］

３）先頭のデータの３つの判別得点を求めよ。

判別得点１［ 90.940 ］判別得点２［ 41.644 ］判別得点３［-4.808 ］

これは何群に予想されたか。　［1群・2群・３群］