データ解析

http://coconut.sys.eng.shizuoka.ac.jp/data/06/

静岡大学工学部
安藤和敏

2006.11.09

第5章 Excelで学ぶ判別分析

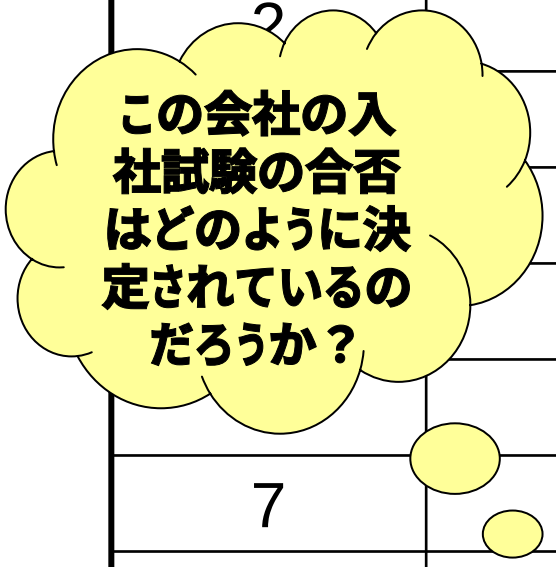
5-1 **相関図で判別分析**
5-2 **線形判別関数を利用した判別分析**
5-3 **マハラノビスの距離を利用した判別分析**

第5章 Excelで学ぶ判別分析

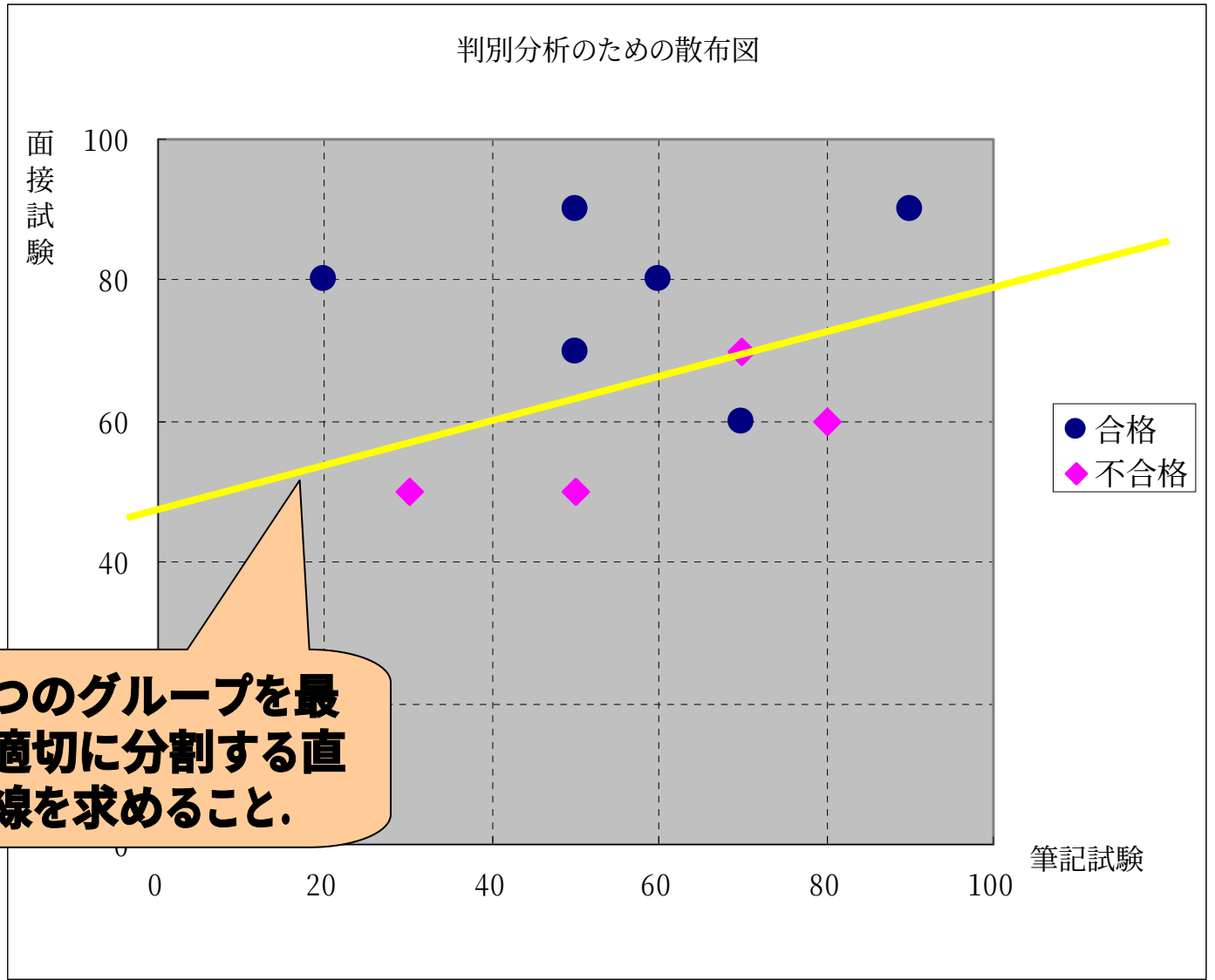
5-1 相関図で判別分析
5-2 線形判別関数を利用した判別分析
5-3 マハラノビスの距離を利用した判別分析

判別分析のデータの例
(ある会社の入社試験の結果)

学生No	筆記試験	面接試験	試験合否
1	20	80	合格
2	50	50	不合格
3	50	70	合格
4	70	60	合格
5	90	90	合格
6	50	90	合格
7	80	60	不合格
8	70	70	不合格
9	60	80	合格
10	30	50	合格



判別分析の目的


判別分析のための散布図
面接試験
100
80
60
40
0
20
40
60
80
100
筆記試験
合格
不合格
2つのグループを最も適切に分割する直線を求めること.

Excelで学ぼう

ファイル：第5章/5_1

第5章 Excelで学ぶ判別分析

5-1 相関図で判別分析
5-2 線形判別関数を利用した判別分析
5-3 マハラノビスの距離を利用した判別分析

判別分析のデータ(2変数の場合)

No	変数 x	変数 y	グループ
1	x_1	y_1	P
⋮	⋮	⋮	⋮
i	x_i	y_i	P
⋮	⋮	⋮	⋮
m	x_m	y_m	P
$m+1$	x_{m+1}	y_{m+1}	Q
⋮	⋮	⋮	⋮
j	x_j	y_j	Q
⋮	⋮	⋮	⋮
n	x_n	y_n	Q

n個の個体からなるデータは，2つのグループPとQに分けられる．

1番からm番のデータはPに，m+1番からn番のデータはQに分類されていると仮定する．

線形判別関数 z

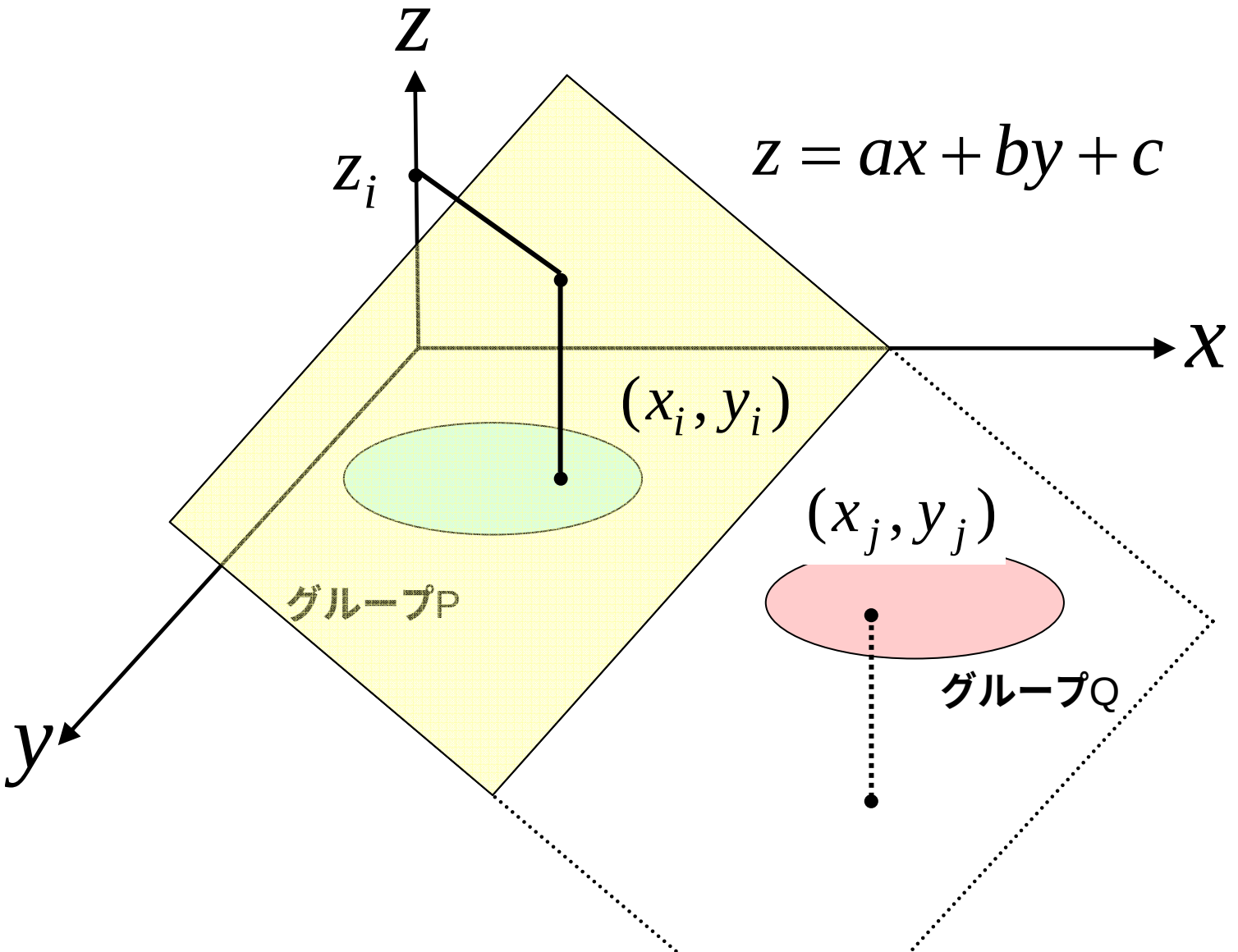
与えられたデータの2つのグループに属する個体がなるべく遠ざかって見えるように，合成変数

$$z = ax + by + c \qquad \cdots\cdots(1)$$

を求めたい．

(1) 式の z は，x，y を変数とする関数とみることも出来るので，z を**線形判別関数**と呼ぶ．

線形判別関数 z



判別得点 $z_i = ax_i + by_i + c$

No	変数 x	変数 y	z
1	x_1	y_1	$ax_1 + by_1 + c$
⋮	⋮	⋮	⋮
i	x_i	y_i	$ax_i + by_i + c$
⋮	⋮	⋮	⋮
m	x_m	y_m	$ax_m + by_m + c$
$m+1$	x_{m+1}	y_{m+1}	$ax_{m+1} + by_{m+1} + c$
⋮	⋮	⋮	⋮
j	x_j	y_j	$ax_j + by_j + c$
⋮	⋮	⋮	⋮
n	x_n	y_n	$ax_n + by_n + c$

変動の分離 (分散の分離)

$$s_z^2 = \frac{1}{n}\sum_{i=1}^{n}\left(z_i - \bar{z}\right)^2$$

$$S_T = \sum_{i=1}^{n}\left(z_i - \bar{z}\right)^2$$

変動の分離(1)

$$S_T = \sum_{i=1}^{m}\left(z_i - \bar{z}\right)^2 + \sum_{j=m+1}^{n}\left(z_j - \bar{z}\right)^2$$

$$= \sum_{i=1}^{m}\left(z_i - \bar{z}_P + \bar{z}_P - \bar{z}\right)^2 + \sum_{j=m+1}^{n}\left(z_j - \bar{z}_Q + \bar{z}_Q - \bar{z}\right)^2$$

$$= \sum_{i=1}^{m}\left\{\left(z_i - \bar{z}_P\right) + \left(\bar{z}_P - \bar{z}\right)\right\}^2$$

$$+ \sum_{j=m+1}^{n}\left\{\left(z_j - \bar{z}_Q\right) + \left(\bar{z}_Q - \bar{z}\right)\right\}^2$$

変動の分離(2)

$$= \sum_{i=1}^{m} \left\{ \left(z_i - \bar{z}_P \right)^2 + 2\left(z_i - \bar{z}_P \right)\left(\bar{z}_P - \bar{z} \right) + \left(\bar{z}_P - \bar{z} \right)^2 \right\}$$

$$+ \sum_{j=m+1}^{n} \left\{ \left(z_j - \bar{z}_Q \right)^2 + 2\left(z_j - \bar{z}_Q \right)\left(\bar{z}_Q - \bar{z} \right) + \left(\bar{z}_Q - \bar{z} \right)^2 \right\}$$

$$= \sum_{i=1}^{m} \left(z_i - \bar{z}_P \right)^2 + \sum_{j=m+1}^{n} \left(z_j - \bar{z}_Q \right)^2$$

$$+ n_P \left(\bar{z}_P - \bar{z} \right)^2 + n_Q \left(\bar{z}_Q - \bar{z} \right)^2$$

$$+ 2\sum_{i=1}^{m} \left(z_i - \bar{z}_P \right)\left(\bar{z}_P - \bar{z} \right) + 2 \sum_{j=m+1}^{n} \left(z_j - \bar{z}_Q \right)\left(\bar{z}_Q - \bar{z} \right)$$

変動の分離(3)

したがって，次式を得る．

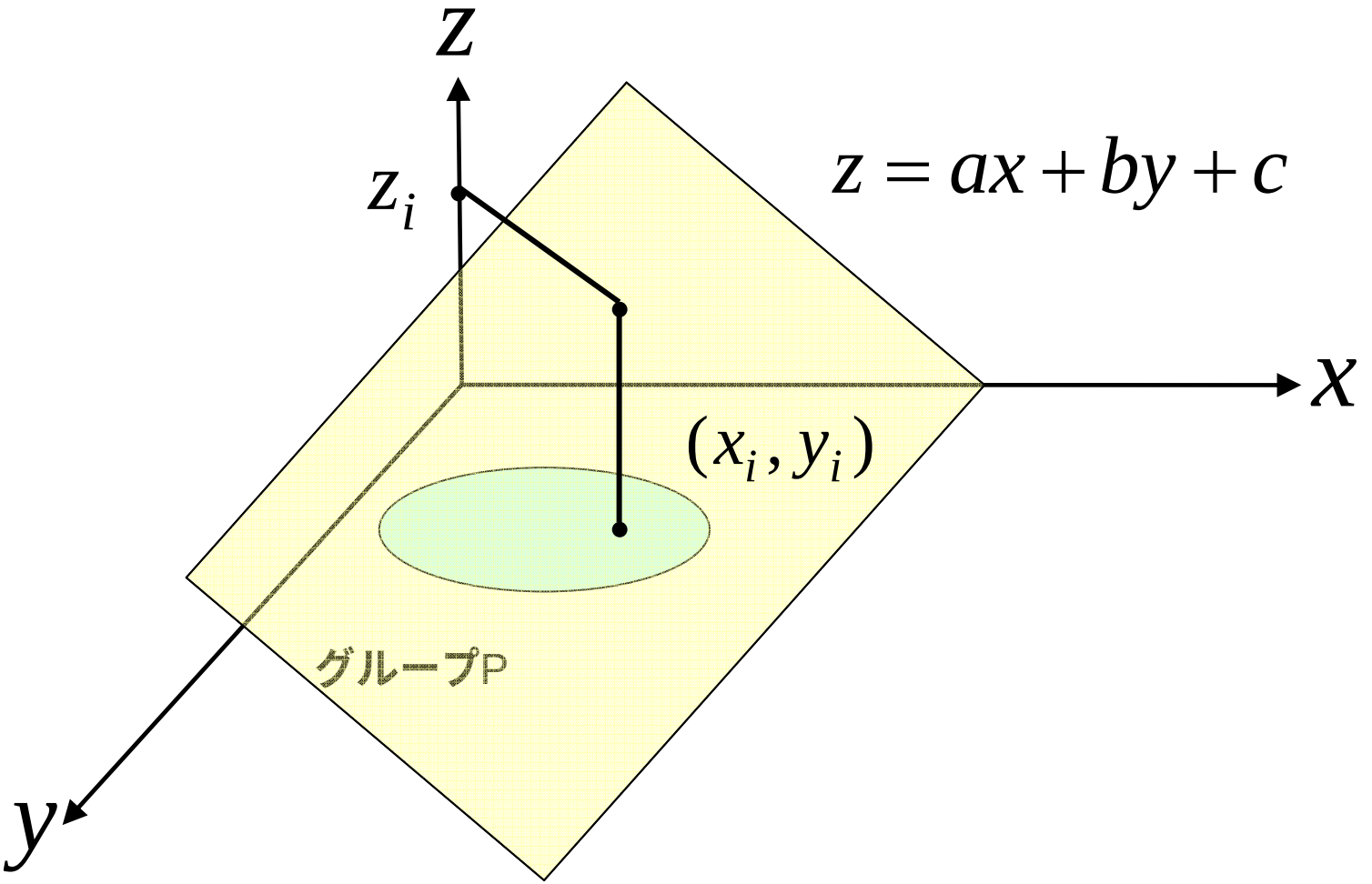
$$S_T = S_B + S_W \quad \ldots(4)$$

ここで，S_BとS_Wは以下で定義される．

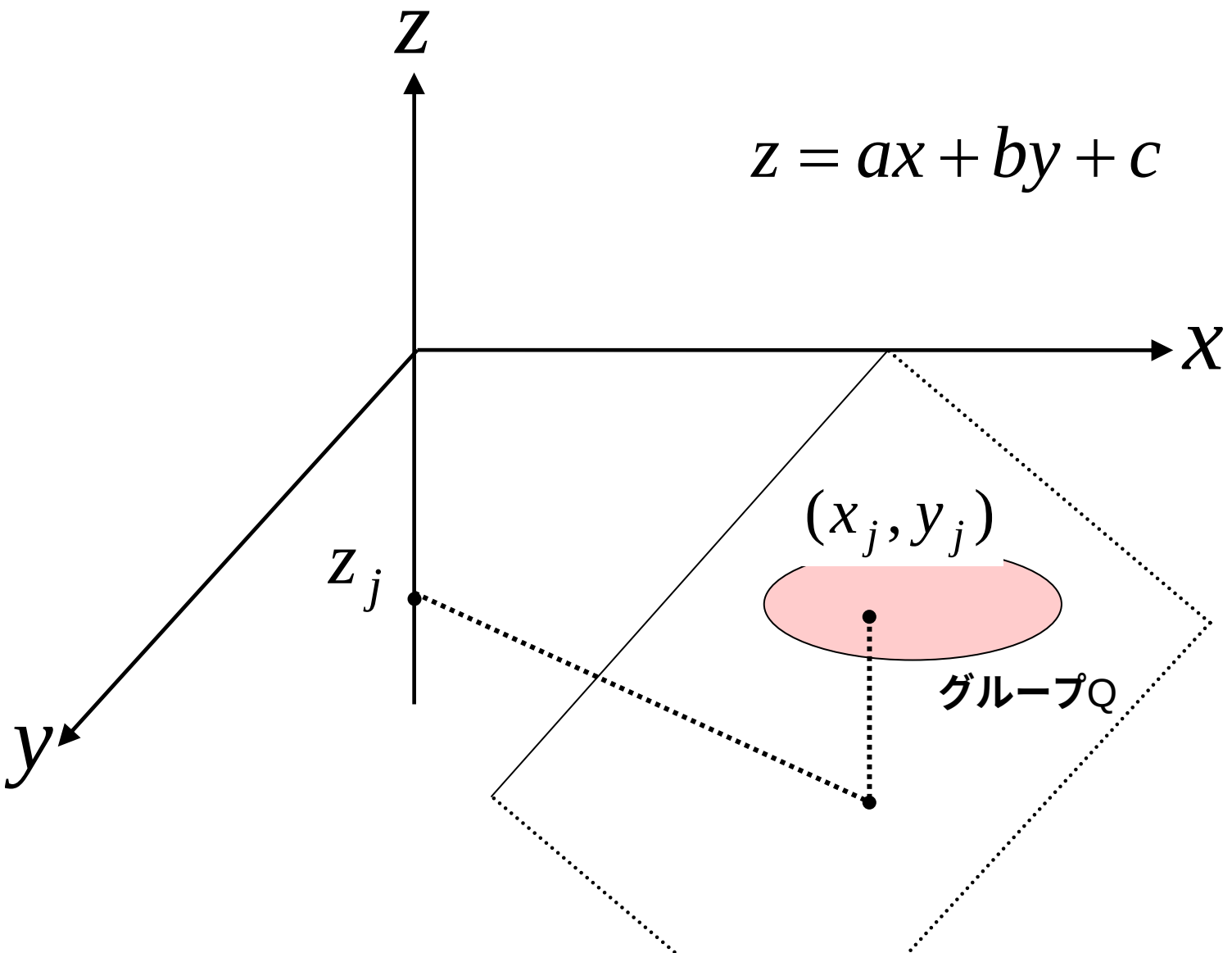
$$S_B = n_P\left(\bar{z}_P - \bar{z}\right)^2 + n_Q\left(\bar{z}_Q - \bar{z}\right)^2 \quad \ldots(5)$$

$$S_W = \sum_{i=1}^{m}\left(z_i - \bar{z}_P\right)^2 + \sum_{j=m+1}^{n}\left(z_j - \bar{z}_Q\right)^2 \quad \ldots(6)$$

線形判別関数 z



線形判別関数 z



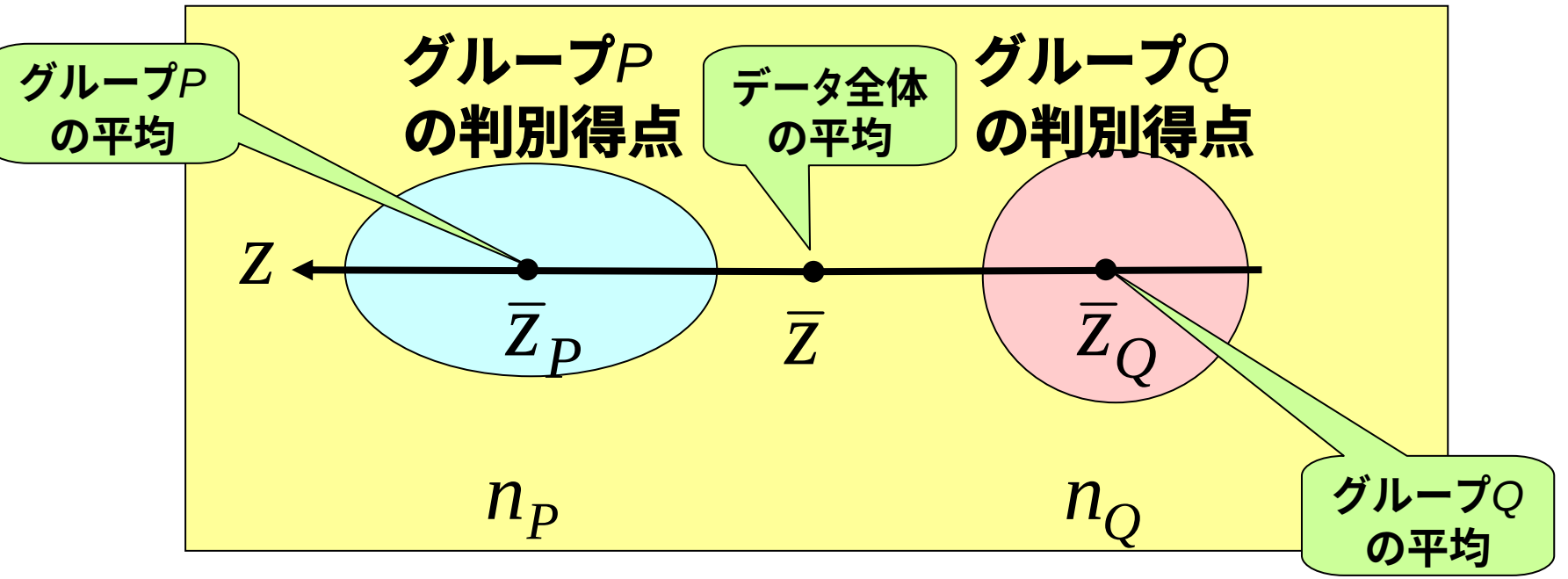
群間変動

グループP全体がどれだけ資料の中心から離れているかを表す

グループQ全体がどれだけ資料の中心から離れているかを表す

$$S_B = n_P\left(\bar{z}_P - \bar{z}\right)^2 + n_Q\left(\bar{z}_Q - \bar{z}\right)^2 \quad \ldots(5)$$

グループPとグループQがどれだけ離れているかを表す



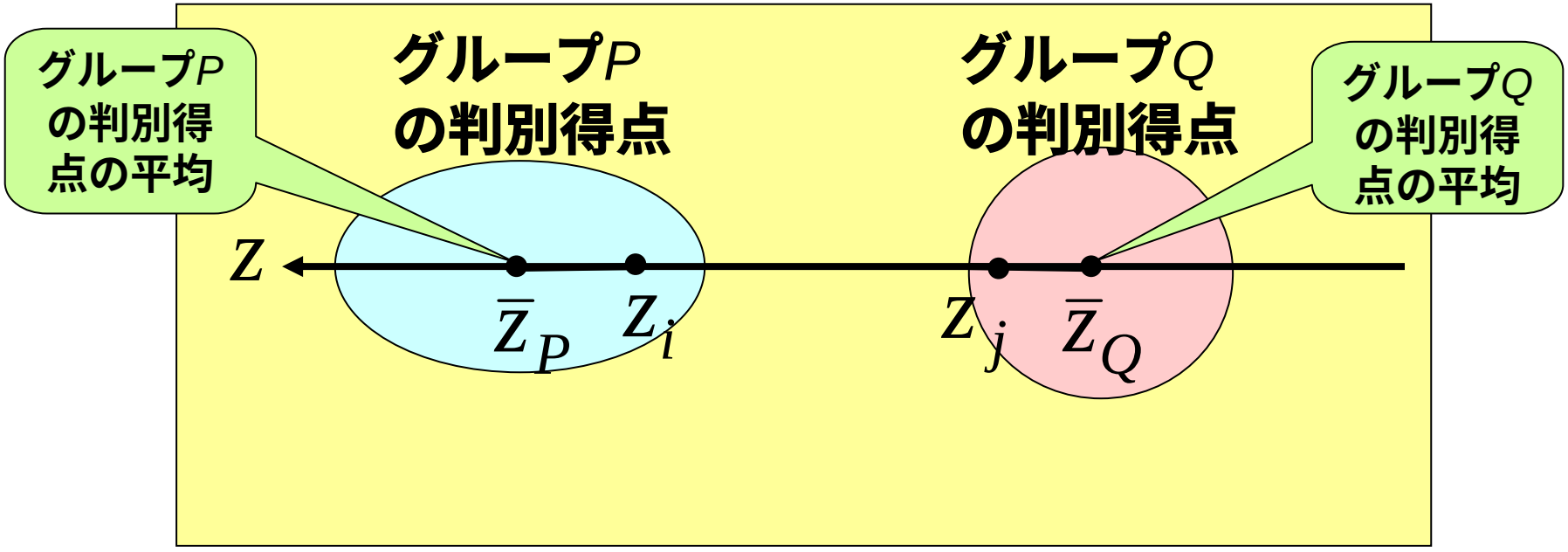
群内変動

グループPに属する個体の判別得点の変動

グループQに属する個体の判別得点の変動

$$S_W = \sum_{i=1}^{m} (z_i - \bar{z}_P)^2 + \sum_{j=m+1}^{n} (z_j - \bar{z}_Q)^2 \quad \cdots(6)$$

2つのグループ内での変動の大きさを表す



相関比 η^2

$$\eta^2 = \frac{S_B}{S_T} \quad \ldots(7)$$

全変動に対する群間変動の割合

$$= \frac{n_P\left(\bar{z}_P - \bar{z}\right)^2 + n_Q\left(\bar{z}_Q - \bar{z}\right)^2}{\sum_{i=1}^{n}\left(z_i - \bar{z}\right)^2}$$

相関比を最大にするようにa,b,cを決定すれば，2つのグループが最も離れて見える，合成変数

$$z = ax + by + c$$

が得られる．

本日のまとめ

- Excelによって散布図を描く方法を理解した.
- 判別分析の目的を理解した.
- 群間変動, 群内変動, 相関比の意味を理解した.

データ解析

http://coconut.sys.eng.shizuoka.ac.jp/data/06/

静岡大学工学部
安藤和敏

2006.11.13

第5章 Excelで学ぶ判別分析

5-1 相関図で判別分析
5-2 線形判別関数を利用した判別分析
5-3 マハラノビスの距離を利用した判別分析

Excelで学ぼう

ファイル：第5章/5_1

データ解析

http://coconut.sys.eng.shizuoka.ac.jp/data/06/

静岡大学工学部
安藤和敏

2006.11.16

第5章 Excelで学ぶ判別分析

5-1 相関図で判別分析
5-2 線形判別関数を利用した判別分析
5-3 マハラノビスの距離を利用した判別分析