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

S_Tの書き直し

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

$$= \sum_{i=1}^{n} \left(a x_i + b y_i + c - a\bar{x} - b\bar{y} - c \right)^2$$

$$= \sum_{i=1}^{n} \left\{ a(x_i - \bar{x}) + b(y_i - \bar{y}) \right\}^2$$

S_Tの書き直し

$$= \sum_{i=1}^{n} \left\{ a^2 (x_i - \bar{x})^2 + b^2 (y_i - \bar{y})^2 \right.$$

$$\left. + 2ab(x_i - \bar{x})(y_i - \bar{y}) \right\}$$

$$= a^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + b^2 \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$+ 2ab \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= n(a^2 s_x^2 + b^2 s_y^2 + 2ab s_{xy})$$

S_Bの書き直し

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

$$= n_P\left(a\bar{x}_P + b\bar{y}_P + c - a\bar{x} - b\bar{y} - c\right)^2$$

$$+ n_Q\left(a\bar{x}_Q + b\bar{y}_Q + c - a\bar{x} - b\bar{y} - c\right)^2$$

$$= n_P\left\{a(\bar{x}_P - \bar{x}) + b(\bar{y}_P - \bar{y})\right\}^2$$

$$+ n_Q\left\{a(\bar{x}_Q - \bar{x}) + b(\bar{y}_Q - \bar{y})\right\}^2$$

$\frac{a}{b}$ の決定

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

$$= \frac{n_P\left\{a(\bar{x}_P - \bar{x}) + b(\bar{y}_P - \bar{y})\right\}^2 + n_Q\left\{a(\bar{x}_Q - \bar{x}) + b(\bar{y}_Q - \bar{y})\right\}^2}{n(a^2 s_x^2 + b^2 s_y^2 + 2abs_{xy})}$$

$\frac{a}{b}$ の決定

$$t = \frac{a}{b} \quad \cdots\cdots(6)$$

$$\eta^2 = \frac{\begin{array}{r} n_P\left\{t(\bar{x}_P - \bar{x}) + (\bar{y}_P - \bar{y})\right\}^2 \\ + n_Q\left\{t(\bar{x}_Q - \bar{x}) + (\bar{y}_Q - \bar{y})\right\}^2 \end{array}}{n(t^2 s_x^2 + s_y^2 + 2ts_{xy})}$$

η^2 を最大化する t は $\frac{d\eta^2}{dt} = 0$ の解である．

$\frac{a}{b}$ の決定

$$A = n_P(\bar{x}_P - \bar{x})^2 + n_Q(\bar{x}_Q - \bar{x})^2,$$

$$B = 2\left\{n_P(\bar{x}_P - \bar{x})(\bar{y}_P - \bar{y}) + n_Q(\bar{x}_Q - \bar{x})(\bar{y}_Q - \bar{y})\right\},$$

$$C = n_P(\bar{y}_P - \bar{y})^2 + n_Q(\bar{y}_Q - \bar{y})^2,$$

$$D = ns_x^2,\ E = 2ns_{xy},\ F = ns_y^2$$

と置くと

$$\eta^2 = \frac{At^2 + Bt + C}{Dt^2 + Et + F}$$

と書ける.

$\frac{a}{b}$ の決定

$$\frac{d\eta^2}{dt} = \frac{(2At+B)(Dt^2+Et+F)-(At^2+Bt+C)(2Dt+E)}{(Dt^2+Et+F)^2}$$

$$= \frac{(AE-BD)t^2+(2AF-2CD)t+(BF-CE)}{(Dt^2+Et+F)^2}$$

であるから，

$$(AE-BD)t^2+(2AF-2CD)t+(BF-CE)=0$$

の2つの解のうちの一方が η^2 を最大化する．

$\frac{a}{b}$ の決定

$$(AE-BD)t^2+(2AF-2CD)t+(BF-CE)=0$$

の2つの解をt_1, t_2 ($t_1 < t_2$)とする.

$$t_1=\frac{-(2AF-2CD)-\sqrt{(2AF-2CD)^2-4(AE-BD)(BF-CE)}}{2(AE-BD)}$$

$$t_2=\frac{-(2AF-2CD)+\sqrt{(2AF-2CD)^2-4(AE-BD)(BF-CE)}}{2(AE-BD)}$$

高校数学より, $AE-BD \geqq 0$のときは, 求めるtはt_1であり, $AE-BD<0$のときは, 求めるtはt_2である.

a,b の決定

$s_z^2 = 1$ を仮定すると，

$$s_z^2 = \frac{S_T}{n} = a^2 s_x^2 + 2abs_{xy} + b^2 s_y^2 = 1.$$

上式に $a = bt$ を代入して，

$$b^2(t^2 s_x^2 + 2ts_{xy} + s_y^2) = 1.$$

a, b の決定

したがって,

$$b = \frac{1}{\sqrt{t^2 s_x^2 + 2ts_{xy} + s_y^2}},$$

$$a = bt = \frac{t}{\sqrt{t^2 s_x^2 + 2ts_{xy} + s_y^2}}.$$

c の決定

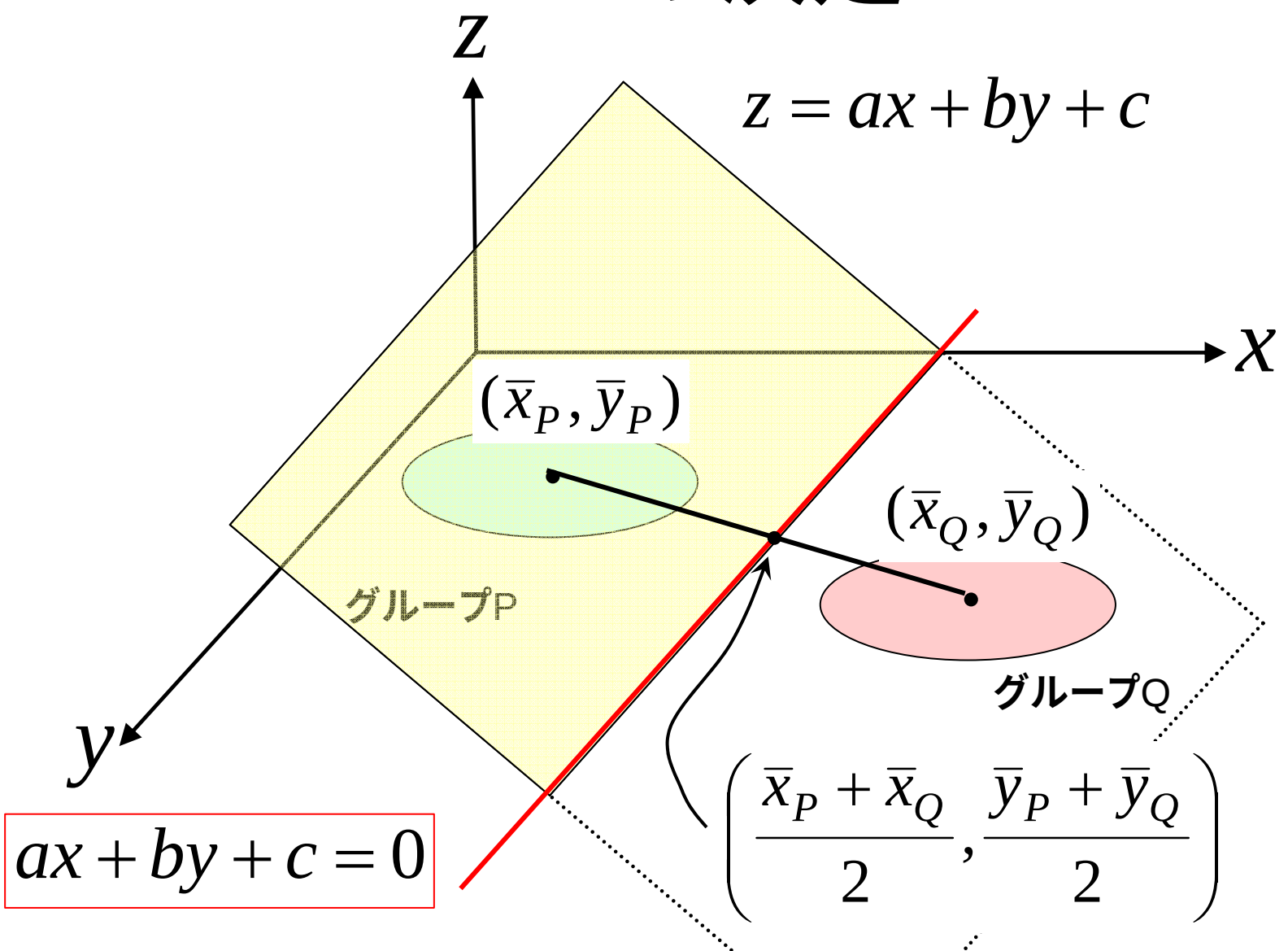
c を, 直線 $0 = ax + by + c$ が $(\overline{x}_P, \overline{y}_P)$ と $(\overline{x}_Q, \overline{y}_Q)$ の中点

$$\left(\frac{\overline{x}_P + \overline{x}_Q}{2}, \frac{\overline{y}_P + \overline{y}_Q}{2} \right)$$

を通るように決定する. すなわち,

$$c = -a \frac{\overline{x}_P + \overline{x}_Q}{2} - b \frac{\overline{y}_P + \overline{y}_Q}{2}.$$

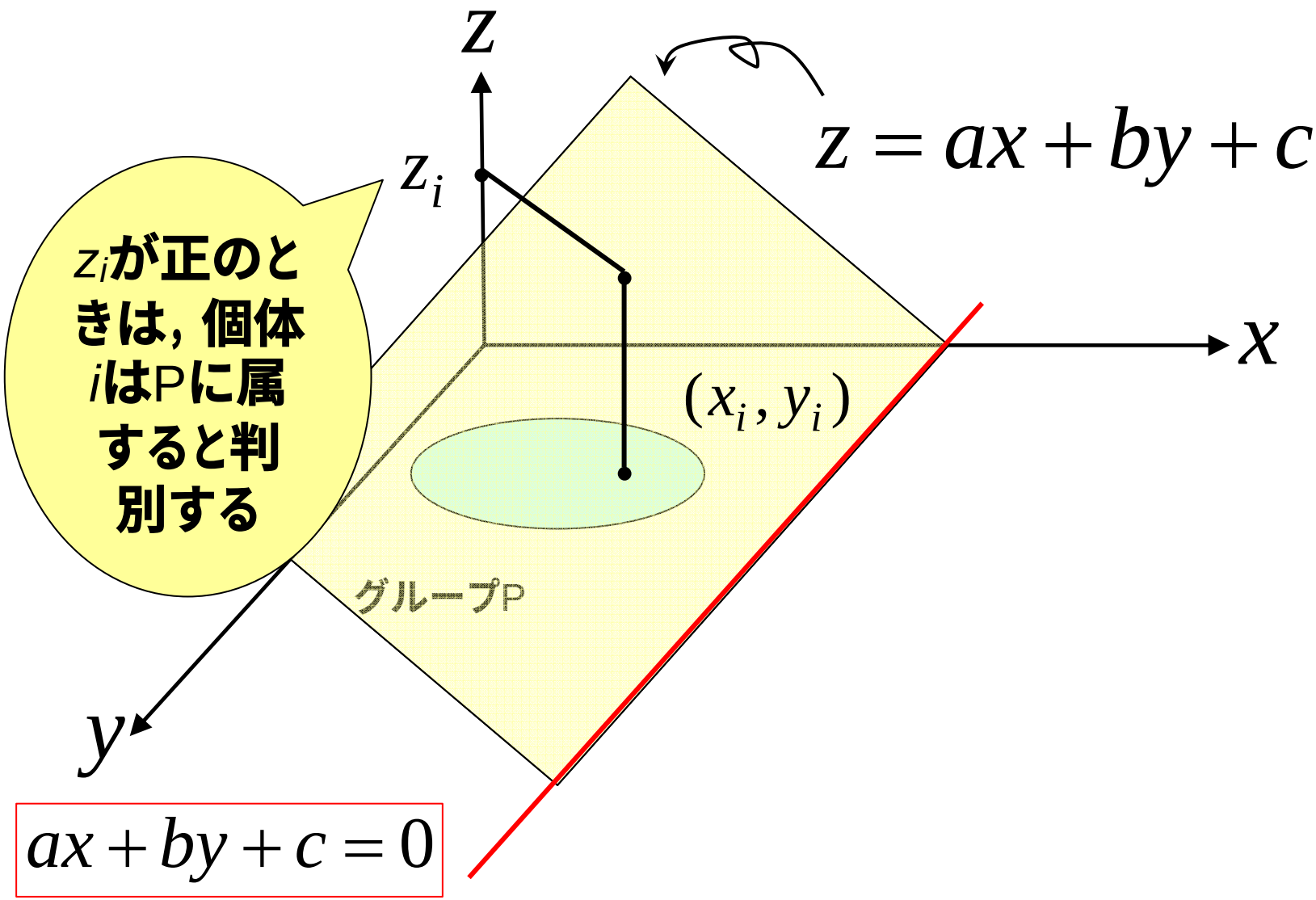
c の決定


z
x
y
$z = ax + by + c$
$(\bar{x}_P, \bar{y}_P)$
$(\bar{x}_Q, \bar{y}_Q)$
グループP
グループQ
$\left(\frac{\bar{x}_P + \bar{x}_Q}{2}, \frac{\bar{y}_P + \bar{y}_Q}{2} \right)$
$ax + by + c = 0$

2グループを判別得点で判別

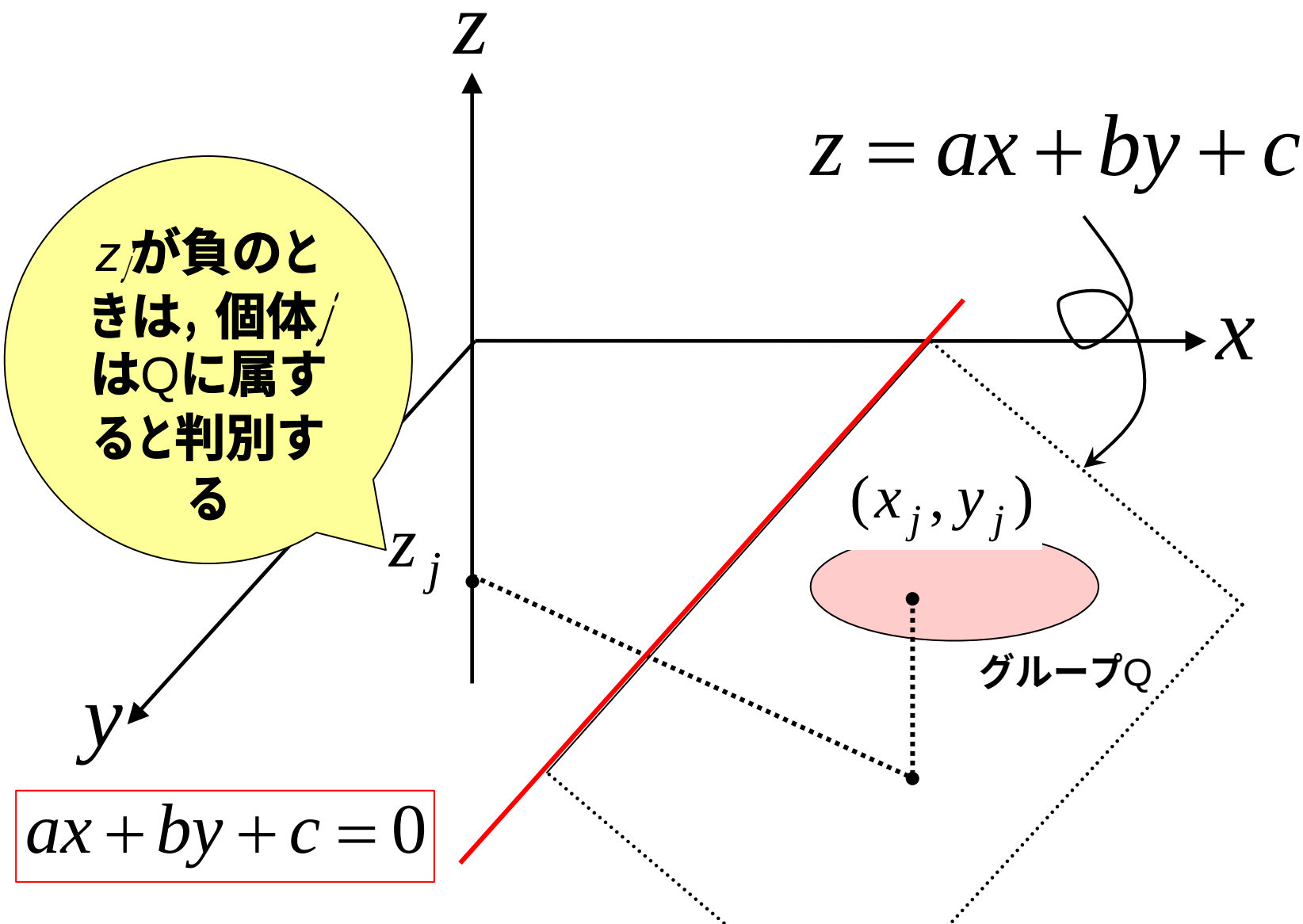
以上で線形判別関数 $z=ax+by+c$ が決定された.

2グループを判別得点で判別

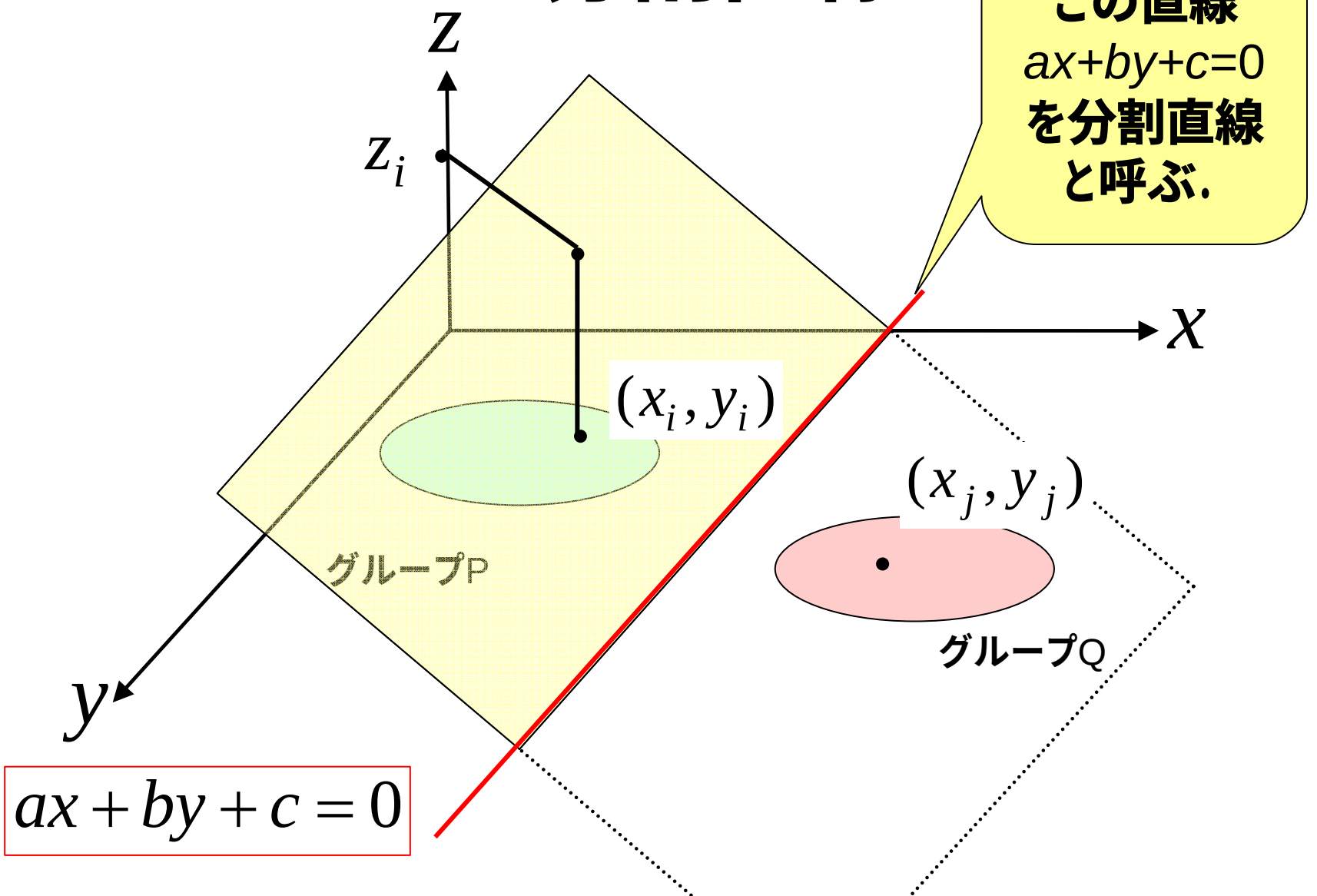


z_iが正のときは，個体iはPに属すると判別する

2グループを判別得点で判別



分割直線



判別的中率

求められた判別関数によって，全ての個体が正しく分類されるわけではない．判別的中率は線形判別関数がどれだけ正しく分類を行っているかを表す．

$$\text{判別的中率} = \frac{\text{正しく判別された個体数}}{\text{全個体数}}$$

ただし，判別的中立は必ず0.5以上になるということはすぐに分かる．

判別的中率

ただし，判別的中立は必ず0.5以上になるということはすぐに分かる．

線形判別関数の良さを表す基準としては，以下の表がよく利用される．

判別的中率	評価
0.9以上	良い
0.8〜0.9	やや良い
0.5〜0.8	良くない

Excelで学ぼう

ファイル：第5章/5_2

本日のまとめ

- 線形判別関数 $z=ax+by+c$ のa,b,cの求め方を学んだ.
- 求められた線形判別関数を用いて，各個体を2つのグループに分類する方法を学んだ.
- 分割直線，判別的中率などの用語の意味を理解した.