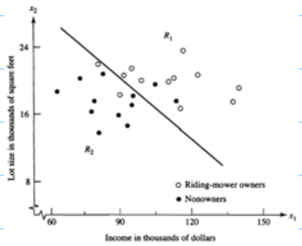
Discrimination Analysis

- Data and Problem:



observed data/ training sample:

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \qquad \begin{array}{c} \textbf{category} \\ 1 \\ 1 \\ \vdots \\ k \end{array}$$

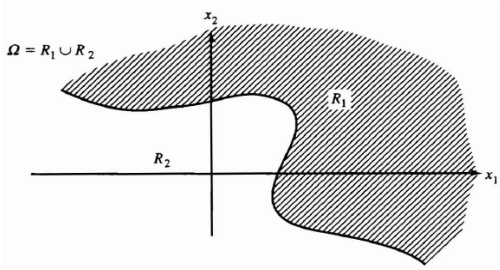
future observation/test sample: $(x_{f1} \quad x_{f2} \quad \ldots \quad x_{fp})$ **?**

- objective: use certain *observed* measurements **X** of some objects whose categories or grouping are *known*, to *determine a rule* that can be used to assign a new object (whose category is *unknown*) to one of the pre-specified categories
- discrimination analysis also known as *pattern recognition*, (*statistical*) *classification*, or *numerical taxonomy*
- examples for k=2

Populations π_1 and π_2	Measured variables **X**
5. Purchasers of a new product and laggards (those "slow" to purchase).	Education, income, family size, amount of previous brand switching.
6. Successful or unsuccessful (fail to graduate) college students.	Entrance examination scores, high school grade-point average, number of high school activities.
9. Alcoholics and nonalcoholics.	Activity of monoamine oxidase enzyme, activity of adenylate cyclase enzyme.

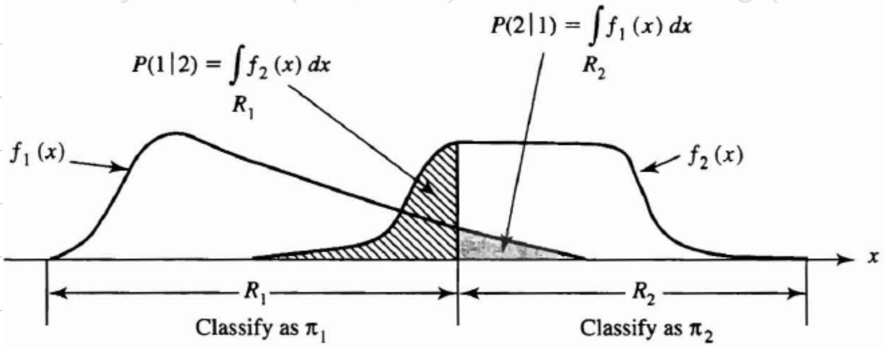
- Q: Why the categories of some objects are known, some unknown? Here are some possible conditions:
  - incomplete knowledge of "future" performance
  - "perfect" information required destroying the object
  - unavailable or expensive information
- a good classification procedure should
  - result in few misclassifications
  - take "prior probabilities of occurrence" into account
    - example. There tend to be more financially sound firms than bankrupt firm. If we really believe that the prior probability of a financially distressed and ultimately bankrupted firm is very small, then one should classify a randomly selected firm as non-bankrupt unless the data overwhelmingly favor bankruptcy.
  - consider the cost
    - Suppose that classifying a π_1 object as belonging to π_2 represents a more serious error than classifying a π_2 object as belonging to π_1. Then, one should be cautious about making the former assignment
    - example. Diagnosis of a potentially fatal illness

- Separation and Classification for Two Populations
  - Modeling of the data and problem
    - Let $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ be the probability density functions associated with the $p \times 1$ vector random variable $\mathbf{X}$ for the populations π_1 and π_2, respectively.
    - Let Ω be the sample space—that is, the collection of all possible observations $\mathbf{x}$.
    - Let R_1 be that set of $\mathbf{x}$ values for which we classify objects as π_1 and $R_2 = \Omega - R_1$ the sets R_1 and R_2 are mutually exclusive and exhaustive.



    - Let p_1 be the *prior* probability of π_1 and p_2 be the *prior* probability of π_2, where $p_1 + p_2 = 1$.
    - The costs of misclassification can be defined by a cost matrix:

		Classify as: π_1	Classify as: π_2
True population:	π_1	0	$c(2\mid 1)$
	π_2	$c(1\mid 2)$	0



  - Calculation of some probabilities
    - The conditional probability, $P(2\mid 1)$, of classifying an object as π_2 when, in fact, it is from π_1 is

$$P(2\mid 1) = P(\mathbf{X} \in R_2 \mid \pi_1) = \int_{R_2 = \Omega - R_1} f_1(\mathbf{x})\,d\mathbf{x}$$

    - the conditional probability, $P(1\mid 2)$, of classifying an object as π_1 when it is really from π_2 is

$$P(1\mid 2) = P(\mathbf{X} \in R_1 \mid \pi_2) = \int_{R_1} f_2(\mathbf{x})\,d\mathbf{x}$$

    - the overall probabilities of correctly or incorrectly classifying objects are
      - $P(\text{observation is correctly classified as } \pi_1) = P(\text{observation comes from } \pi_1 \text{ and is correctly classified as } \pi_1)$

$$= P(\mathbf{X} \in R_1 \mid \pi_1) P(\pi_1) = P(1\mid 1) p_1$$

- P(observation is misclassified as π_1) $= P$(observation comes from π_2 and is misclassified as π_1)
$$= P(\mathbf{X} \in R_1 | \pi_2) P(\pi_2) = P(1|2) p_2$$
- P(observation is correctly classified as π_2) $= P$(observation comes from π_2 and is correctly classified as π_2)
$$= P(\mathbf{X} \in R_2 | \pi_2) P(\pi_2) = P(2|2) p_2$$
- P(observation is misclassified as π_2) $= P$(observation comes from π_1 and is misclassified as π_2)
$$= P(\mathbf{X} \in R_2 | \pi_1) P(\pi_1) = P(2|1) p_1$$

➢ *expected cost of misclassification* (ECM) criterion

$$\text{ECM} = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$$

⇒ a reasonable classification rule should have ECM as small as possible

➢ **Result 11.1.** The regions R_1 and R_2 that minimize the ECM are defined by the values $\mathbf{x}$ for which the following inequalities hold:

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} \geq \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix}$$

$$R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} < \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix}$$

- Note. Implementation of the minimum ECM rule requires only three *ratios*
- special cases of minimum expected cost regions
  - $p_2/p_1 = 1$ (equal prior probabilities)
$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \qquad R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)}$$
  - $c(1|2)/c(2|1) = 1$ (equal misclassification costs)
$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \qquad R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1}$$
  - $p_2/p_1 = c(1|2)/c(2|1) = 1$ or $p_2/p_1 = 1/(c(1|2)/c(2|1))$ (equal prior probabilities and equal misclassification costs)
$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \qquad R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1$$

- other criteria
  - total probability of misclassification (TPM)

$$\begin{aligned}\text{TPM} &= P(\text{misclassifying a } \pi_1 \text{ observation } \textit{or} \text{ misclassifying a } \pi_2 \text{ observation}) \\ &= P(\text{observation comes from } \pi_1 \text{ and is misclassified}) \\ &\quad + P(\text{observation comes from } \pi_2 \text{ and is misclassified}) \\ &= p_1 \int_{R_2} f_1(\mathbf{x})\,d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})\,d\mathbf{x}\end{aligned}$$

$\Rightarrow$ equivalent to minimizing ECM when costs of misclassification are equal

  - "posterior" probability approach

$$\begin{aligned}P(\pi_1 | \mathbf{x}_0) &= \frac{P(\pi_1 \text{ occurs and we observe } \mathbf{x}_0)}{P(\text{we observe } \mathbf{x}_0)} \\ &= \frac{P(\text{we observe } \mathbf{x}_0 | \pi_1)P(\pi_1)}{P(\text{we observe } \mathbf{x}_0 | \pi_1)P(\pi_1) + P(\text{we observe } \mathbf{x}_0 | \pi_2)P(\pi_2)} \\ &= \frac{p_1 f_1(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)}\end{aligned}$$

$$P(\pi_2 | \mathbf{x}_0) = 1 - P(\pi_1 | \mathbf{x}_0) = \frac{p_2 f_2(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)}$$

$\Rightarrow$ Classifying an observation $\mathbf{x}_0$ as π_1 when $P(\pi_1 | \mathbf{x}_0) > P(\pi_2 | \mathbf{x}_0)$

$\Rightarrow$ equivalent to minimizing ECM when costs of misclassification are equal

- Classification with Two Multivariate Normal Populations
  - now, further assume that $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are multivariate normal densities,
    - $f_1(\mathbf{x})$ with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_1$
    - $f_2(\mathbf{x})$ with mean vector $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$.
  - Classification of Normal Populations When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$
    - Suppose that the joint densities of $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ for populations π_1 and π_2 are given by

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right] \quad \text{for } i = 1, 2$$

    - minimum ECM regions become

$$R_1: \quad \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R_2: \quad \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

    - **Result 11.2.** Let the populations π_1 and π_2 be described by multivariate normal densities of the form (11-10). Then the allocation rule that minimizes the ECM is as follows: Allocate $\mathbf{x}_0$ to π_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Allocate $\mathbf{x}_0$ to π_2 otherwise.

<u>proof</u>.

- in practical situation, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are usually unknown $\Rightarrow$ replacing the population parameters by their counterparts (Q: how?)

- Suppose, then, that we have n_1 observations of the multivariate random variable $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ from π_1 and n_2 measurements of this quantity from π_2, with $n_1 + n_2 - 2 \geq p$. Then the respective data matrices are

$$\underset{(n_1 \times p)}{\mathbf{X}_1} = \begin{bmatrix} \mathbf{x}'_{11} \\ \mathbf{x}'_{12} \\ \vdots \\ \mathbf{x}'_{1n_1} \end{bmatrix} \qquad \underset{(n_2 \times p)}{\mathbf{X}_2} = \begin{bmatrix} \mathbf{x}'_{21} \\ \mathbf{x}'_{22} \\ \vdots \\ \mathbf{x}'_{2n_2} \end{bmatrix}$$

the sample mean vectors and covariance matrices are

$$\underset{(p\times 1)}{\bar{\mathbf{x}}_1} = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{x}_{1j}, \quad \underset{(p\times p)}{\mathbf{S}_1} = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)'$$

$$\underset{(p\times 1)}{\bar{\mathbf{x}}_2} = \frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{x}_{2j}, \quad \underset{(p\times p)}{\mathbf{S}_2} = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)'$$

Since it is assumed that the parent populations have the same covariance matrix $\boldsymbol{\Sigma}$, the sample covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ are combined (pooled) to derive a single, unbiased estimate of $\boldsymbol{\Sigma}$

$$\mathbf{S}_{\text{pooled}} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_2$$

- **The Estimated Minimum ECM Rule for Two Normal Populations**

Allocate $\mathbf{x}_0$ to π_1 if

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Allocate $\mathbf{x}_0$ to π_2 otherwise.

- If $\left(\dfrac{c(1|2)}{c(2|1)}\right)\left(\dfrac{p_2}{p_1}\right) = 1$

then $\ln(1) = 0$, and the estimated minimum ECM rule for two normal populations amounts to comparing the scalar variable

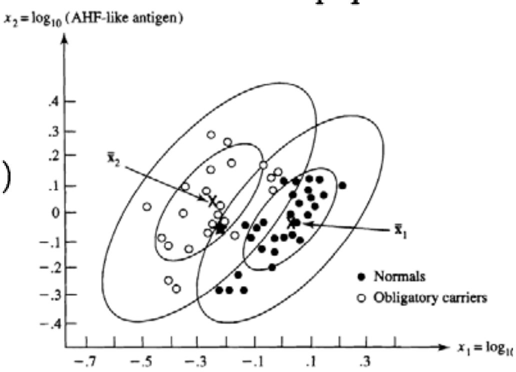
$$\hat{y} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} = \hat{\mathbf{a}}'\mathbf{x}$$

evaluated at $\mathbf{x}_0$, with the number

$$\hat{m} = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) = \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$$

where $\bar{y}_1 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_1 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_1$

$\bar{y}_2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_2 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_2$



- the estimated minimum ECM rule is to

(1) creating two univariate populations for the y values by taking an appropriate linear combination of the observations from two populations

(2) assign a new observation $\mathbf{x}_0$ to π_1 or π_2 depending upon whether $\hat{y}_0 = \hat{\mathbf{a}}'\mathbf{x}_0$ falls to the right or left of the midpoint $\hat{m}$ between the two univariate means $\bar{y}_1$ and $\bar{y}_2$

$\Rightarrow$ it is called *linear discriminant* analysis