- other criteria
  - total probability of misclassification (TPM)

$$\begin{aligned}\text{TPM} &= P(\text{misclassifying a } \pi_1 \text{ observation } \textit{or} \text{ misclassifying a } \pi_2 \text{ observation})\\ &= P(\text{observation comes from } \pi_1 \text{ and is misclassified}) = P(X\in\pi_1)P(X\in R_2 \mid X\in\pi_1)\\ &\quad + P(\text{observation comes from } \pi_2 \text{ and is misclassified})\\ &= p_1 \int_{R_2} f_1(\mathbf{x})\,d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})\,d\mathbf{x} \xleftrightarrow{\text{c.f.}} \text{ECM in LN p.8-5}\end{aligned}$$

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

  - "posterior" probability approach

Bayesian rule

$$\begin{aligned}P(\pi_1 \mid \mathbf{x}_0) &= \frac{P(\pi_1 \text{ occurs and we observe } \mathbf{x}_0)}{P(\text{we observe } \mathbf{x}_0) = P(\text{observe } x_0 \,\&\, x_0\in\pi_1) + P(\text{observe } x_0 \,\&\, x_0\in\pi_2)}\\ &= \frac{P(\text{we observe } \mathbf{x}_0 \mid \pi_1)P(\pi_1)}{P(\text{we observe } \mathbf{x}_0 \mid \pi_1)P(\pi_1) + P(\text{we observe } \mathbf{x}_0 \mid \pi_2)P(\pi_2)}\\ &= \frac{p_1 f_1(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)}\end{aligned}$$

R_1 region in Result 11.1, LN p. 8-5 c.f. $f_1(x_0)/f_2(x_0) > \frac{p_2}{p_1}$ $\Updownarrow$ $p_1 f_1(x_0) > p_2 f_2(x_0)$ $\Updownarrow$

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

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

$\Rightarrow$ 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 $\boxed{\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}}$
    - 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

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

    - minimum ECM regions become

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

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

$\frac{f_1(x)}{f_2(x)} \equiv (*)$

- **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

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

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

proof. $\ln[(*)] = -\frac{1}{2}\left[x'\Sigma^{-1}x - \mu_1'\Sigma^{-1}x - x'\Sigma^{-1}\mu_1 + \mu_1'\Sigma^{-1}\mu_1\right] + \frac{1}{2}\left[x'\Sigma^{-1}x - \mu_2'\Sigma^{-1}x - x'\Sigma^{-1}\mu_2 + \mu_2'\Sigma^{-1}\mu_2\right]$

- 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 $n_1 + n_2 - 2 \ge p$. Then the respective data matrices are

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

dataset $= \begin{bmatrix} X_1 \\ \hline X_2 \end{bmatrix}$ popu 1 ⋮ 2 ⋮ 2

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 $\boldsymbol{\Sigma}$ If $n_1 >> n_2$, $n_1 \approx n_2$.

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

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

Allocate $\mathbf{x}_0$ to π_1 if   future observation

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

$\bar{\mathbf{x}}_1 \to \mu_1$, $\bar{\mathbf{x}}_2 \to \mu_2$, $\mathbf{S}_{\text{pooled}} \to \Sigma$

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

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

classification rule.

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

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

→ x projected on the direction $\hat{a}$.

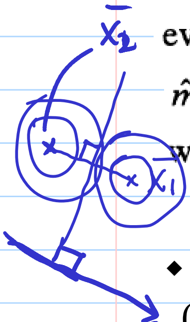
defined by Spooled.

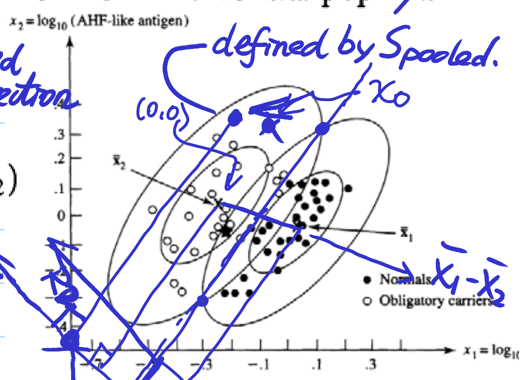
evaluated at $\mathbf{x}_0$, with the number

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

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 $\bar{y}_2$ → the sample ~~set~~ space is

$\Rightarrow$ it is called *linear discriminant* analysis Seperated into R_1, R_2 by a plane.

➢ Fisher's approach to classification with two populations

- Fisher's idea was to transform the *multivariate* variables $X_1, \ldots, X_p$ to a *univariate* variable Y, which is a linear function of the X variables, i.e.,

$Var(Y) = a^T \Sigma a$ ← $cov(X)$

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p, = \underbrace{a}_{=(a_1,\ldots,a_p)}{}^T X \leftarrow (X_1,\ldots,X_p)^T$$

such that the Y observations derived from the two populations were separated as much as possible

Recall: two sample problem.

- A fixed linear combination of the $\mathbf{x}$'s takes the values $y_{11}, y_{12}, \ldots, y_{1n_1}$ for the observations from the first population and the values $y_{21}, y_{22}, \ldots, y_{2n_2}$ for the observations from the second population. The separation of these two sets of univariate y's is assessed in terms of the difference between $\bar{y}_1$ and $\bar{y}_2$ expressed in standard deviation units. That is, → c.f. t-test.

$$\text{separation} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}, \quad \text{where } s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j}-\bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j}-\bar{y}_2)^2}{n_1+n_2-2}$$

$$= \frac{n_1-1}{n_1+n_2-2} s_{y_1}^2 + \frac{n_2-1}{n_1+n_2-2} s_{y_2}^2$$

μ_{y_1} μ_{y_2}

is the pooled estimate of the variance.

- **Result 11.3.** The linear combination $\hat{y} = \hat{\mathbf{a}}'\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}$ maximizes the ratio

$$\frac{\left(\begin{array}{c}\text{squared distance}\\ \text{between sample means of } y\end{array}\right)}{(\text{sample variance of } y)} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{\mathbf{a}}'\bar{\mathbf{x}}_1 - \hat{\mathbf{a}}'\bar{\mathbf{x}}_2)^2}{\hat{\mathbf{a}}'\mathbf{S}_{\text{pooled}}\hat{\mathbf{a}}} = \frac{(\hat{\mathbf{a}}'\mathbf{d})^2}{\hat{\mathbf{a}}'\mathbf{S}_{\text{pooled}}\hat{\mathbf{a}}}$$

between group variation; within group variation; $\mathbf{d} = (\bar{x}_1 - \bar{x}_2)$

μ_{y_1} μ_{y_2}

over all possible coefficient vectors $\hat{\mathbf{a}}$ where $\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. The maximum of the ratio is $D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$.

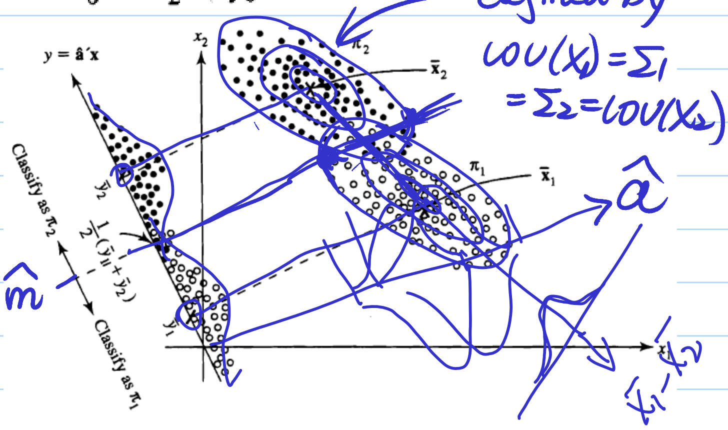
proof. By maximization Lemma in LNp.2-17.
$\hat{a} = c \cdot S_{pooled}^{-1} d$, maximum value $= d' S_{pooled}^{-1} d$.

- allocation rule based on Fisher's discriminant function

Allocate $\mathbf{x}_0$ to π_1 if $\hat{y}_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0$

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

Allocate $\mathbf{x}_0$ to π_2 if $\hat{y}_0 < \hat{m}$

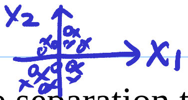


Note: 1st LDF minimize E(ECM). Fisher's LDF maximize "separation".

→ do not need normality assumption.

- Note. Fisher's linear discriminant function was developed under the assumption that the two population, whatever their form, have a common covariance matrix.

➢ Is classification a good idea for your data?



- Fro two populations, the maximum relative separation that can be obtained by considering linear combinations of the multivariate observations is equal to the distance D^2
- Note. D^2 can be used to test whether the population means differ significantly (Hotelling's T^2 test)

  ⇒ a test for differences in mean vectors can be viewed as a test for the "significance" of the separation that can be achieved
- Note. Significant separation does not necessarily imply good classification. By contrast, if the separation is not significant, the search for a useful classification rule will probably prove fruitless

$$\boxed{f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}}$$

• Classification of Normal Populations When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

➢ **Result 11.4.** Let the populations π_1 and π_2 be described by multivariate normal densities with mean vectors and covariance matrices $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2$, respectively. The allocation rule that minimizes the expected cost of misclassification is given by

$f_1(x;\mu_1,\Sigma_1)$
$f_2(x;\mu_2,\Sigma_2)$

quadratic form.

$$R_1:\quad -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + \underbrace{(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x}}_{\text{linear function of } x} - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

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

where

$$k = \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) + \frac{1}{2}(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2) \leftarrow \text{function of parameter.}$$

- it is called *quadratic* classification because of the quadratic term

  c.f. LD linear.

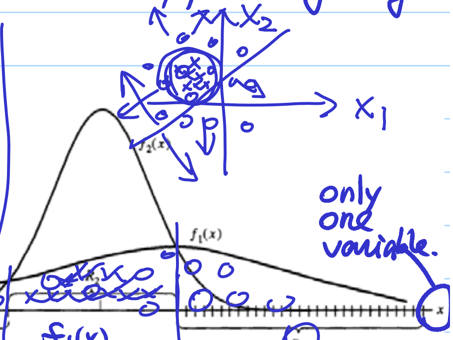
$$-\tfrac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} \rightarrow t x_i^2 + \cdots + t x_p^2 + \sum t_{ij} x_i x_j.$$

➢ Quadratic classification rule

Allocate $\mathbf{x}_0$ to π_1 if

$$-\frac{1}{2}\mathbf{x}_0'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x}_0 + (\bar{\mathbf{x}}_1'\mathbf{S}_1^{-1} - \bar{\mathbf{x}}_2'\mathbf{S}_2^{-1})\mathbf{x}_0 - k(\bar{x}_1, \bar{x}_2, S_1, S_2)$$
$$\geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

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



• Evaluating classification functions

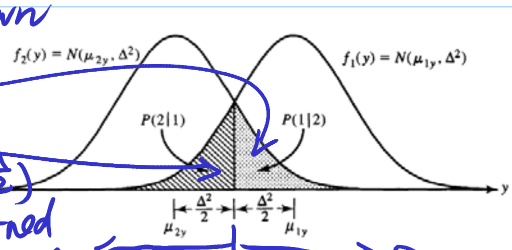
➢ one important way of judging the performance of any classification procedure is to calculate its "error rate," or misclassification probabilities

Type I error
Type II error.

➢ total probability of misclassification (population)

unknown

$$\text{TPM} = p_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x}$$



➢ actual error rate (sample)

$f_1(x,\hat{\mu},\hat{\Sigma})$ $f_2(x,\hat{\mu},\hat{\Sigma})$

$$\text{AER} = p_1 \int_{\hat{R}_2} f_1(\mathbf{x})\, d\mathbf{x} + p_2 \int_{\hat{R}_1} f_2(\mathbf{x})\, d\mathbf{x}$$

determined by training sample

➢ apparent error rate (do not dependent on population densities) → no need to know f_1, f_2

APER = proportion of items in the training set that are misclassified

Q: does $n_1 : n_2$ reflect $P_1 : P_2$?

		Predicted membership		
		π_1	π_2	
Actual membership	π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
	π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

$$\text{APER} = \frac{n_{1M} + n_{2M}}{n_1 + n_2} = \underbrace{\frac{n_2}{n_1+n_2}}_{\hat{P}_2} \cdot \overbrace{\frac{n_{2M}}{n_2}}^{\hat{f}(1|2)} + \underbrace{\frac{n_1}{n_1+n_2}}_{\hat{P}_1} \cdot \underbrace{\frac{n_{1M}}{n_1}}_{\hat{f}(2|1)}$$

- it is easy to calculate and can be calculated for *any* classification procedure
- it tends to *underestimate* the AER because the data used to build the classification function are also used to evaluate it
- one procedure is to split the total sample into a training sample and a validation sample, but it required large sample and the information in the validation sample is not used to construct the classification function

➢ cross-validation method (leave-one-out method)

1. Start with the π_1 group of observations. Omit one observation from this group, and develop a classification function based on the remaining $n_1 - 1$, n_2 observations.
2. Classify the "holdout" observation, using the function constructed in Step 1.
3. Repeat Steps 1 and 2 until all of the π_1 observations are classified. Let $n_{1M}^{(H)}$ be the number of holdout (H) observations misclassified in this group.

4. Repeat Steps 1 through 3 for the π_2 observations. Let $n_{2M}^{(H)}$ be the number of holdout observations misclassified in this group.

- $$\hat{P}(2|1) = \frac{n_{1M}^{(H)}}{n_1} \qquad \hat{P}(1|2) = \frac{n_{2M}^{(H)}}{n_2}$$

$$\hat{E}(\text{AER}) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2}$$

$$= \underbrace{\frac{n_1}{n_1+n_2}}_{\hat{P}_1} \cdot \frac{n_{1M}^{(H)}}{n_1} + \overbrace{\frac{n_2}{n_1+n_2}}^{\hat{P}_2} \cdot \frac{n_{2M}^{(H)}}{n_2}$$

❖ **Reading**: textbook, 11.1, 11.2, 11.3, 11.4