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.

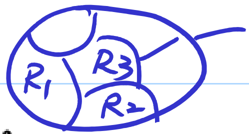
- $$\hat{P}(2|1) = \frac{n_{1M}^{(H)}}{n_1} \qquad \hat{P}(1|2) = \frac{n_{2M}^{(H)}}{n_2}$$

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

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

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

- Classification with Several Populations
  - ➢ generalization of classification procedure from 2 to $g \geq 2$ groups — generalization of the theory is straight forward.
  - ➢ minimum expected cost of misclassification method
    - Let $f_i(\mathbf{x})$ be the density associated with population π_i, $i = 1, 2, \ldots, g$.
    - Let p_i = the prior probability of population π_i, $i = 1, 2, \ldots, g$
    - Let $c(k|i)$ = the cost of allocating an item to π_k when, in fact, it belongs to π_i, for $k, i = 1, 2, \ldots, g$
      $c(i|i) = 0.$
    - Let R_k be the set of $\mathbf{x}$'s classified as π_k



whole sample space

- Then,

$$P(k|i) = P(\text{classifying item as } \pi_k | \pi_i) = \int_{R_k} f_i(\mathbf{x})\, d\mathbf{x} \quad \text{for } k, i = 1, 2, \ldots, g$$

- The conditional expected cost of misclassifying an $\mathbf{x}$ from π_1 into π_2, or $\pi_3, \ldots,$ or π_g is

$$\text{ECM}(1) = P(2|1)c(2|1) + P(3|1)c(3|1) + \cdots + P(g|1)c(g|1) = \sum_{k=2}^{g} P(k|1)c(k|1)$$

- In a similar manner, we can obtain the conditional expected costs of misclassification $\text{ECM}(2), \ldots, \text{ECM}(g)$.
- The overall ECM is:

$$\text{ECM} = p_1\text{ECM}(1) + p_2\text{ECM}(2) + \cdots + p_g\text{ECM}(g)$$

$$= p_1\left(\sum_{k=2}^{g} P(k|1)c(k|1)\right) + p_2\left(\sum_{\substack{k=1\\k\neq 2}}^{g} P(k|2)c(k|2)\right) + \cdots + p_g\left(\sum_{k=1}^{g-1} P(k|g)c(k|g)\right)$$

$$= \sum_{i=1}^{g} p_i\left(\sum_{\substack{k=1\\k\neq i}}^{g} P(k|i)c(k|i)\right) = \sum_{i=1}^{g} p_i\left(\sum_{k=1}^{g}\int_{R_k} f_i(x)dx\; c(k|i)\right) = \sum_{k=1}^{g}\int_{R_k}\underbrace{\left[\sum_{i=1}^{g} p_i f_i(x) c(k|i)\right]}_{= R_k(x)} dx$$

$\because c(i|i) = 0$

- **Result 11.5.** The classification regions that minimize the ECM are defined by allocating $\mathbf{x}$ to that population π_k, $k = 1, 2, \ldots, g$, for which

$$\sum_{\substack{i=1\\i\neq k}}^{g} p_i f_i(\mathbf{x}) c(k|i)$$

is smallest. If a tie occurs, $\mathbf{x}$ can be assigned to any of the tied populations.

$R_k = \{x \mid R_k(x) = \min_{k'} R_{k'}(x)\}$

all $c(k|i)$ are equal $k \neq i$

- Suppose all the misclassification costs are equal. The minimum ECM is the minimum total probability of misclassification. In which case, we would allocate $\mathbf{x}$ to that population $\pi_k, k = 1, 2, \ldots, g$, for which

$$ECM = \sum_{i=1}^{g} p_i \left(\sum_{\substack{k=1 \\ k \neq i}}^{g} P(k|i) \right) = \sum P(x \in \pi_i, \text{claim } x \in \pi_k)$$

$$\sum_{\substack{i=1 \\ i \neq k}}^{g} p_i f_i(\mathbf{x}) = 1 - p_k f_k(x)$$

is smallest. It will be smallest when the omitted term, $p_k f_k(\mathbf{x})$, is *largest*.

- Minimum ECM Classification Rule with equal misclassification costs

$$\text{Allocate } \mathbf{x}_0 \text{ to } \pi_k \text{ if } p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x}) \quad \text{for all } i \neq k \Rightarrow \frac{f_k}{f_i} > \frac{p_i}{p_k}$$

- Notice that the classification rule is identical to the one that maximize the posterior probability

Bayes rule

$$P(\pi_k | \mathbf{x}) = P \text{ (}\mathbf{x} \text{ comes from } \pi_k \text{ given that } \mathbf{x} \text{ was observed)}$$

$$= \frac{p_k f_k(\mathbf{x})}{\sum_{i=1}^{g} p_i f_i(\mathbf{x})} = \frac{(\text{prior}) \times (\text{likelihood})}{\sum [(\text{prior}) \times (\text{likelihood})]} \quad \text{for } k = 1, 2, \ldots, g$$

➢ Classification with Normal Populations

- Under *normality assumption*,

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

- Allocate $\mathbf{x}$ to π_k if

$p_k f_k > p_i f_i$ $\Leftrightarrow$ $\ln p_k f_k > \ln p_i f_i$.

$$\ln p_k f_k(\mathbf{x}) = \ln p_k - \cancel{\left(\frac{p}{2}\right) \ln(2\pi)} - \frac{1}{2} \ln |\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)$$

$$= \max_i \ln p_i f_i(\mathbf{x})$$

- define quadratic discrimination score for ith population

large if data spread in a wide region.

$$d_i^Q(\mathbf{x}) = -\tfrac{1}{2} \ln |\boldsymbol{\Sigma}_i| - \tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \ln p_i \qquad i = 1, 2, \ldots, g$$

- Minimum Total Probability of Misclassification (TPM) rule for normal populations with unequal $\boldsymbol{\Sigma}_i$

Allocate $\mathbf{x}$ to π_k if the quadratic score $d_k^Q(\mathbf{x})$ = largest of $d_1^Q(\mathbf{x}), d_2^Q(\mathbf{x}), \ldots, d_g^Q(\mathbf{x})$

- In practice, the $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are unknown $\Rightarrow$ replaced by their sample quantities

$$\hat{d}_i^Q(\mathbf{x}) = -\tfrac{1}{2} \ln |\mathbf{S}_i| - \tfrac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_i)' \mathbf{S}_i^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i) + \ln p_i, \quad i = 1, 2, \ldots, g$$

- Estimated Minimum TPM rule for normal population with unequal $\boldsymbol{\Sigma}_i$

Allocate $\mathbf{x}$ to π_k if the quadratic score $\hat{d}_k^Q(\mathbf{x})$ = largest of $\hat{d}_1^Q(\mathbf{x}), \hat{d}_2^Q(\mathbf{x}), \ldots, \hat{d}_g^Q(\mathbf{x})$

- When $\boldsymbol{\Sigma}_i \stackrel{\wedge}{=} \boldsymbol{\Sigma}$ ← assumption. for $i = 1, 2, \ldots, g$,

$$d_i^Q(\mathbf{x}) = \cancel{-\tfrac{1}{2} \ln |\boldsymbol{\Sigma}|} - \cancel{\tfrac{1}{2} \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x}} + \boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} \mathbf{x} - \tfrac{1}{2} \boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \ln p_i$$

The first two terms are the same for $d_1^Q(\mathbf{x}), d_2^Q(\mathbf{x}), \ldots, d_g^Q(\mathbf{x})$.

- define the linear discriminant score → direction for projection

C.f. 2 groups ↔

$$d_i(\mathbf{x}) = \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i \quad \text{for } i = 1, 2, \ldots, g$$

An estimate $\hat{d}_i(\mathbf{x})$ of the linear discriminant score $d_i(\mathbf{x})$ is based on the pooled

$$\mathbf{S}_{\text{pooled}} = \frac{1}{n_1 + n_2 + \cdots + n_g - g}\left((n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \cdots + (n_g - 1)\mathbf{S}_g\right)$$

and is given by

$$\hat{d}_i(\mathbf{x}) = \bar{\mathbf{x}}_i'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} - \tfrac{1}{2}\bar{\mathbf{x}}_i'\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_i + \ln p_i \quad \text{for } i = 1, 2, \ldots, g$$

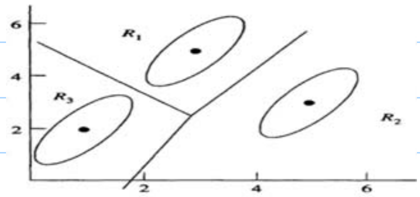
- Estimated Minimum TPM rule for normal populations with equal covariance

  Allocate $\mathbf{x}$ to π_k if

  the linear discriminant score $\hat{d}_k(\mathbf{x})$ = the largest of $\hat{d}_1(\mathbf{x}), \hat{d}_2(\mathbf{x}), \ldots, \hat{d}_g(\mathbf{x})$

  - An equivalent classifier for the equal-covariance case is to use

    $$D_i^2(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_i)'\mathbf{S}_{\text{pooled}}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i)$$

    It measure the squared distances from $\mathbf{x}$ to the sample mean vector $\bar{\mathbf{x}}_i$. The allocatory rule is then

    Assign $\mathbf{x}$ to the population π_i for which $-\frac{1}{2}D_i^2(\mathbf{x}) + \ln p_i$ is largest

  - If the prior probabilities are unknown, the usual procedure is to set $p_1 = p_2 = \cdots = p_g = 1/g$. An observation is then assigned to the closest population.