➢ 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.,

$$Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p,$$

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

- 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,

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

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

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

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 $D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$.

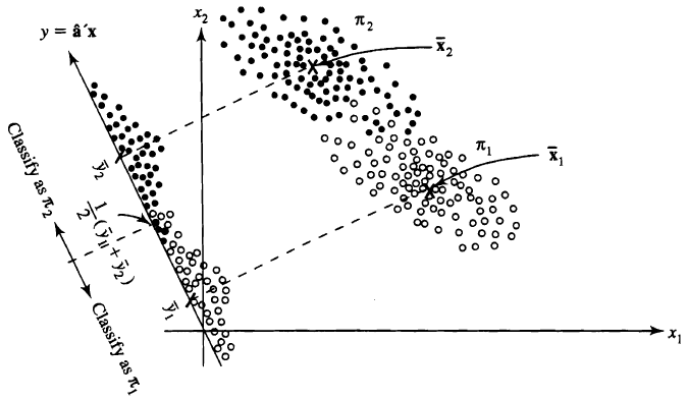
proof.

- 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. 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)

    $\Rightarrow$ 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

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

• 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

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

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

where $k = \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) + \frac{1}{2}(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)$

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

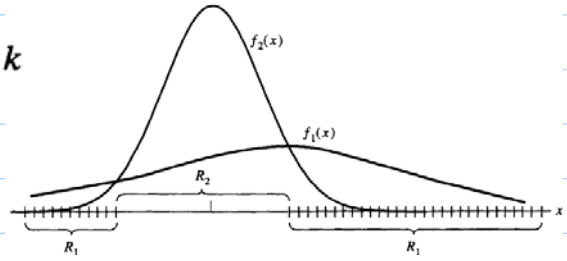
$$-\tfrac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x}$$

- Quadratic classification rule

Allocate $\mathbf{x}_0$ to π_1 if

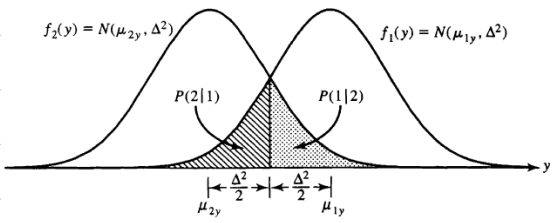
$$-\frac{1}{2}\mathbf{x}_0'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x}_0 + (\bar{\mathbf{x}}_1'\mathbf{S}_1^{-1} - \bar{\mathbf{x}}_2'\mathbf{S}_2^{-1})\mathbf{x}_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

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
  - total probability of misclassification (population)

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



  - actual error rate (sample)

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

  - apparent error rate (do not dependent on population densities)

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

		Predicted membership π_1	Predicted membership 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

$$\text{APER} = \frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

- 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.

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

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