

We can estimate the sensitivity and specificity by the following statistics:

$$\widehat{\eta}(c) = \frac{1}{n_1} \sum_{i=1}^{n_1} I_{\{y_{1i} \geq c\}}, \quad \widehat{\xi}(c) = \frac{1}{n_1} \sum_{j=1}^{n_1} I_{\{z_j \leq c\}}$$

Following Example 1 of 3.1.1, both estimates can be viewed as one-sample U-statistics.

Since  $\eta(c)$  and  $\xi(c)$  depend on the cut-off  $c$ , they vary as  $c$  changes. By varying  $c$ , we can bring either sensitivity or specificity arbitrarily close to 1. Thus, to compare performance between two diagnostic tests, we must examine sensitivity and specificity simultaneously. The Receiver Operating Characteristics (ROC) curve is widely used as an index of performance of test kit by studying  $\eta(c)$  and  $\xi(c)$  jointly along the continuum  $c$ .

An ROC curve is the plot of the bivariate  $(1 - \eta(c), \xi(c))$  as a function of  $c$  with a theoretical range  $(-\infty, \infty)$ . A good test kit should maintain high values of both  $\eta(c)$  and  $\xi(c)$  across all values of  $c$ . The area under the ROC curve (AUC)  $\theta$  can be expressed as (see exercise):

$$\begin{aligned} \theta &= \int_0^1 \eta(\xi) d\xi = \int_{-\infty}^{+\infty} [1 - F_1(t)] dF_2(t) = 1 - \int_{-\infty}^{+\infty} F_1(t) dF_2(t) \quad (3.24) \\ &= 1 - E[F_1(y_{2j})] = 1 - E\left[E\left(I_{\{y_{1i} \leq y_{2j}\}} \mid y_{2j}\right)\right] \\ &= 1 - E\left(I_{\{y_{1i} \leq y_{2j}\}}\right) = E\left(I_{\{y_{2j} \leq y_{1i}\}}\right) \end{aligned}$$

In the next example, we show how to construct a U-statistic for estimating this parameter  $\theta$ .

**Example 12 (U-statistic for single ROC curve).** Let  $y_{1i}$  and  $y_{2j}$  denote the continuous test outcomes from the diseased and non-disease samples as defined above ( $1 \leq i \leq n_1, 1 \leq j \leq n_2$ ). Define a symmetric kernel as:  $h(y_{1i}; y_{2j}) = I_{\{y_{2j} \leq y_{1i}\}}$ . Then, the following two-sample U-statistic is an unbiased estimate of the AUC parameter of interest  $\theta$ :

$$\widehat{\theta}_n = \left[ \binom{n_1}{1} \binom{n_2}{1} \right]^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I_{\{y_{2j} \leq y_{1i}\}} \quad (3.25)$$

We can readily extend the two-sample U-statistics to general  $K$ -sample U-statistics. Consider  $K$  i.i.d. samples of random vectors,  $\mathbf{y}_{ki}$  ( $1 \leq i \leq n_k, 1 \leq k \leq K$ ). Let

$$h_{1i_1, \dots, 1i_{m_1}; \dots; Ki_1, \dots, Ki_{m_K}} = h\left(\mathbf{y}_{1i_1}, \dots, \mathbf{y}_{1i_{m_1}}; \dots; \mathbf{y}_{Ki_1}, \dots, \mathbf{y}_{Ki_{m_K}}\right)$$