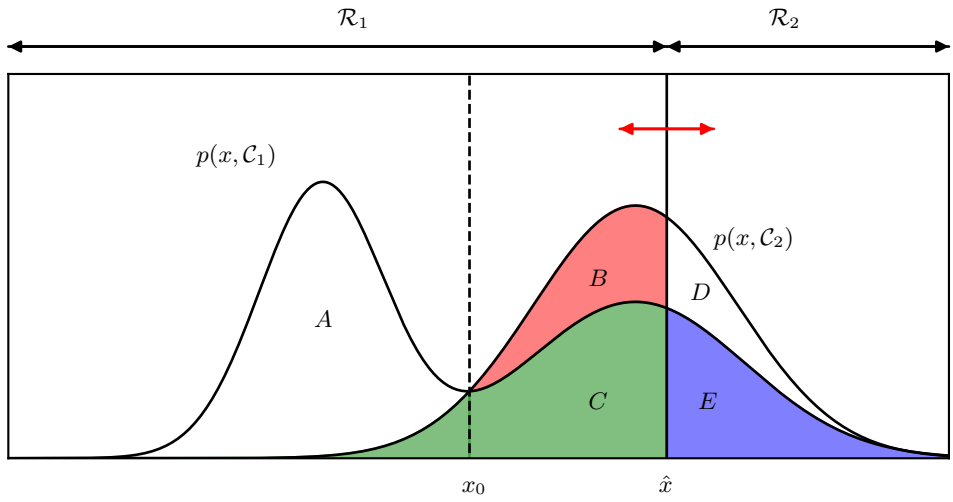

Figure 5.10 As in Figure 5.5, with the various regions labelled. In the cancer classification problem, region $\mathcal{R}_1$ is assigned to the normal class whereas region $\mathcal{R}_2$ is assigned to the cancer class.

correct detection of cancer on the y-axis versus the cumulative fraction of incorrect detection on the x-axis. Note that a specific confusion matrix represents one point along the ROC curve. The best possible classifier would be represented by a point at the top left corner of the ROC diagram. The bottom left corner represents a simple classifier that assigns every point to the normal class and therefore has no true positives but also no false positives. Similarly, the top right corner represents a classifier that assigns everything to the cancer class and therefore has no false negatives but also no true negatives. In Figure 5.11, the classifiers represented by the blue curve are better than those of the red curve for any choice of, say, false positive rate. It is also possible, however, for such curves to cross over, in which case the choice of which is better will depend on the choice of operating point.

As a baseline, we can consider a random classifier that simply assigns each data point to cancer with probability ρ and to normal with probability $1 - \rho$. As we vary the value of ρ it will trace out an ROC curve given by a diagonal straight line, as shown in Figure 5.11. Any classifier below the diagonal line performs worse than random guessing.

Sometimes it is useful to have a single number that characterises the whole ROC curve. One approach is to measure the area under the curve (AUC). A value of 0.5 for the AUC represents random guessing whereas a value of 1.0 represents a perfect classifier.

Another measure is the *F-score*, which is the geometric mean of precision and