Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

3 NonBayesian Hypothesis Testing

In the last section, we developed the so-called Bayesian approach to decision making, whereby we assign *a priori* probabilities to the possible hypotheses, and we select an (average) cost criterion with respect to which we evaluate the quality of a candidate decision rule. As we saw, in the case of our binary hypothesis testing scenario, the resulting optimum decision rule took the form of a likelihood ratio test (LRT)

$$L(\mathbf{y}) = \frac{p_{\mathsf{y}|H}(\mathbf{y}|H_1)}{p_{\mathsf{y}|H}(\mathbf{y}|H_0)} \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \eta, \tag{1}$$

where the threshold η is determined by the choice of prior probabilities and cost criterion.

However, in many problems of interest, it is not clear how to choose such prior probabilities, nor appropriate costs. Indeed, the threshold, and therefore the decision rule, can be quite sensitive to the choice of these quantities. This has led to a variety of alternative *nonBayesian* formulations of decision theory. Interestingly, for essentially any reasonable formulation of the binary hypothesis testing problem, the optimum (deterministic) decision rule turns out to be an LRT of the form (1) for some choice of the threshold η. In this section, we will see a second example of such formulation.

Given the apparently central role of the LRT in binary hypothesis testing, we first explore some of its properties in more detail.

3.1 The Operating Characteristic of the Likelihood Ratio Test

In this section, starting from a Bayesian viewpoint, we develop some additional perspectives on likelihood ratio tests. These perspectives will useful in our subsequent nonBayesian development.

We begin by observing that the performance of any decision rule $\hat{H}(\cdot)$ may be fully specified in terms of two quantities

$$\begin{aligned} P_{\mathrm{D}} &= \mathbb{P}\left(\hat{H}(\mathbf{y}) = H_1 \;\middle|\; H = H_1\right) = \int_{\mathcal{Y}_1} p_{\mathsf{y}|H}(\mathbf{y}|H_1)\,\mathrm{d}\mathbf{y} \\ P_{\mathrm{F}} &= \mathbb{P}\left(\hat{H}(\mathbf{y}) = H_1 \;\middle|\; H = H_0\right) = \int_{\mathcal{Y}_1} p_{\mathsf{y}|H}(\mathbf{y}|H_0)\,\mathrm{d}\mathbf{y}, \end{aligned} \tag{2}$$

where $\mathcal{Y}_0$ and $\mathcal{Y}_1$ are, as before, the sets of $\mathbf{y}$ for which we make decisions $\hat{H}(\mathbf{y}) = H_0$ and $\hat{H}(\mathbf{y}) = H_1$, respectively. Using terminology that originated in the radar community, where H_1 refers to the presence of a target and H_0 the absence, the quantities

P_{D} and P_{F} are generally referred to as the "detection" and "false-alarm" probabilities, respectively (and, hence, the choice of notation). In the statistics community, by contrast, P_{F} is referred to as the *size* of the test and P_{D} as the *power* of the test.

It is worth emphasizing that the characterization in terms of $(P_{\mathrm{D}}, P_{\mathrm{F}})$ is not unique, however. For example, any invertible linear or affine transformation of the pair $(P_{\mathrm{D}}, P_{\mathrm{F}})$ is also complete. For instance, the pair of "probabilities of error of the first and second kind" defined respectively via

$$\begin{aligned} P_{\mathrm{E}}^1 &= \mathbb{P}\left(\hat{H}(\mathbf{y}) = H_1 \,\middle|\, \mathsf{H} = H_0\right) = P_{\mathrm{F}} \\ P_{\mathrm{E}}^2 &= \mathbb{P}\left(\hat{H}(\mathbf{y}) = H_0 \,\middle|\, \mathsf{H} = H_1\right) = 1 - P_{\mathrm{D}} \triangleq P_{\mathrm{M}} \end{aligned} \tag{3}$$

constitute such a characterization, and are preferred in some communities that make use of decision theory. As (3) indicates, from the radar perspective the probability of error of the first kind is the probability of a false alarm, while probability of error of the second kind is the probability of a miss, which is denoted by P_{M}. Additionally, In medical and other applications where H_1 refers to the presence of a disease and H_0 to its absence, P_{F} and P_{M} are the probabilities of what are referred to as "false positive" and "false negative," respectively.

In general, a good decision rule (detector) is one with a large P_{D} and a small P_{F} (or equivalently, small P_{E}^1 and P_{E}^2). However, ultimately these are competing objectives. As an illustration of this behavior, let us examine the performance of a likelihood ratio test (1) when the threshold η is varied. Note that each choice of η completely specifies a decision rule, with which is associated a particular $(P_{\mathrm{D}}, P_{\mathrm{F}})$ operating point. Hence, each value of η is associated with a single point in the P_{D}–P_{F} plane. Moreover, as η is varied from 0 to ∞, a curve is traced out in this plane as illustrated in Fig. 1. This curve is referred to as the *operating characteristic* of the likelihood ratio test.

As Fig. 1 suggests, good P_{D} is generally obtained at the expense of high P_{F}, and so choosing a threshold η for a particular problem involves making an acceptable tradeoff. Indeed, as $\eta \to 0$ we have $(P_{\mathrm{D}}, P_{\mathrm{F}}) \to (1, 1)$, while as $\eta \to \infty$ we have $(P_{\mathrm{D}}, P_{\mathrm{F}}) \to (0, 0)$. From this perspective, the Bayesian test represents a particular tradeoff, and corresponds to a single point on this curve. To obtain this tradeoff, we effectively selected as our objective function a linear combination of P_{D} and P_{F}. More specifically, we minimize

$$\varphi(f) = \alpha P_{\mathrm{F}} - \beta P_{\mathrm{D}} + \gamma,$$

over all possible decision rules, where the choice of α and β is, in turn, determined by the cost assignment (C_{ij}'s) and the *a priori* probabilities (P_m's). In particular, rewriting the Bayes' Risk (expected cost) in the form

$$\varphi(f) = \sum_{i,j} C_{ij} \; \mathbb{P}\left(\hat{H}(\mathbf{y}) = H_i \,\middle|\, \mathsf{H} = H_j\right) \; P_j$$

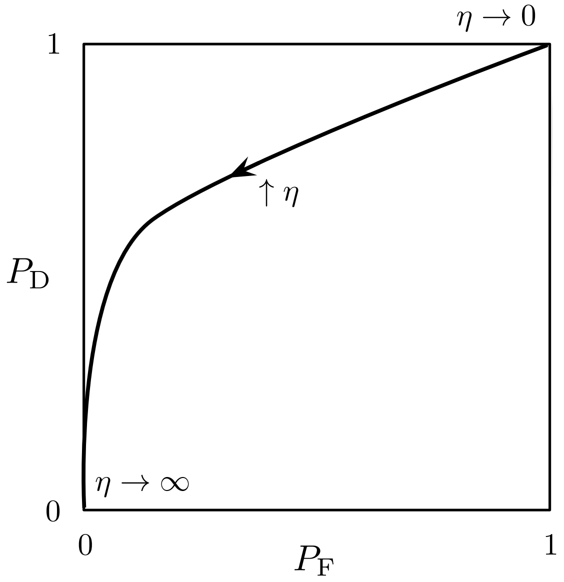

Figure 1: Operating characteristicoc associated with a likelihood ratio test.

we obtain

$$\alpha = (C_{10} - C_{00})P_0 \qquad \beta = (C_{01} - C_{11})P_1 \qquad \gamma = (C_{00}P_0 + C_{01}P_1).$$

Let us explore a specific example to gain further insight.

Example 1. Let us consider the following special case of our earlier simple scalar Gaussian detection problem:

$$\begin{aligned} &H_0 : y \sim \mathtt{N}(0, \sigma^2) \\ &H_1 : y \sim \mathtt{N}(\mu, \sigma^2), \qquad \mu \geq 0, \end{aligned} \tag{4}$$

which corresponds to choosing $s_0 = 0$ and $s_1 = \mu$. Following from our our earlier Gaussian example development, we obtain that the optimum decision rule takes the form

$$y \underset{\hat{H}(y)=H_0}{\overset{\hat{H}(y)=H_1}{\gtrless}} \frac{\mu}{2} + \frac{\sigma^2 \ln \eta}{\mu} \triangleq \gamma,$$

so that

$$P_\mathrm{D} = \int_\gamma^\infty p_{y|H}(y|H_1)\,\mathrm{d}y \tag{5a}$$

$$P_\mathrm{F} = \int_\gamma^\infty p_{y|H}(y|H_0)\,\mathrm{d}y. \tag{5b}$$

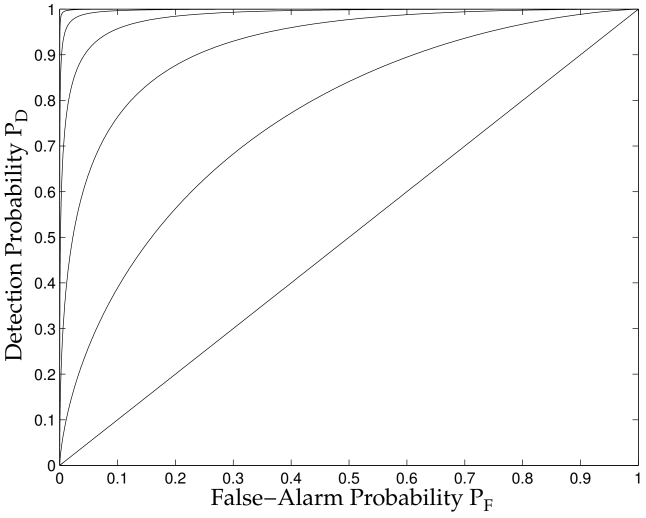

Figure 2: Operating characteristic of the likelihood ratio test for the scalar Gaussian detection problem. The successively higher curves correspond to $d = \mu/\sigma = 0, 1, \ldots, 5$.

The expressions (5a) and (5b) each correspond to tail probabilities in a Gaussian distribution, which are useful to express in "standard form." In particular, we have

$$P_{\mathrm{D}} = \mathbb{P}\left(\mathsf{y} > \gamma \mid \mathsf{H} = H_1\right) = \mathbb{P}\left(\frac{\mathsf{y}-\mu}{\sigma} > \frac{\gamma-\mu}{\sigma} \,\middle|\, \mathsf{H} = H_1\right) = Q\left(\frac{\gamma-\mu}{\sigma}\right) \tag{6a}$$

$$P_{\mathrm{F}} = \mathbb{P}\left(\mathsf{y} > \gamma \mid \mathsf{H} = H_0\right) = \mathbb{P}\left(\frac{\mathsf{y}}{\sigma} > \frac{\gamma}{\sigma} \,\middle|\, \mathsf{H} = H_0\right) = Q\left(\frac{\gamma}{\sigma}\right), \tag{6b}$$

where, following notation we introduced earlier,

$$Q(\alpha) \triangleq \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\infty} e^{-x^2/2}\,\mathrm{d}x \tag{7}$$

is the area under the tail of the unit Gaussian density. From (6a) and (6b) we see that as γ is varied, a curve is traced out in the P_{D}–P_{F} plane. Moreover, this curve is parameterized by $d = \mu/\sigma$. The quantity d^2 can be viewed as a "signal-to-noise ratio," so that d is a normalized measure of "distance" between the hypotheses. Several of these curves are plotted in Fig. 2. Note that when $m = 0$ ($d = 0$), the hypotheses (4) are indistinguishable, and the curve $P_{\mathrm{D}} = P_{\mathrm{F}}$ is obtained. As d increases, better performance is obtained—e.g., for a given P_{F}, a larger P_{D} is obtained. Finally, as $d \to \infty$, the P_{D}–P_{F} curve approaches the ideal operating characteristic: $P_{\mathrm{D}} = 1$ for all $P_{\mathrm{F}} > 0$.

The operating characteristic of the likelihood ratio test has several noteworthy properties. For now we mention one of immediate interest.

Fact 1. *The operating characteristic of the likelihood ratio test is monotonically non-decreasing.*

Proof. This fact follows rather immediately from the structure of these tests. In particular, let $P_{\mathrm{D}}(\eta)$ and $P_{\mathrm{F}}(\eta)$ be the detection and false-alarm probabilities, respectively, for the deterministic test associated with a generic threshold η. Then for any η_1 and η_2 such that $\eta_2 > \eta_1$ we have

$$P_{\mathrm{D}}(\eta_2) \leq P_{\mathrm{D}}(\eta_1) \tag{8a}$$

$$P_{\mathrm{F}}(\eta_2) \leq P_{\mathrm{F}}(\eta_1). \tag{8b}$$

Hence,

$$\frac{P_{\mathrm{D}}(\eta_1) - P_{\mathrm{D}}(\eta_2)}{P_{\mathrm{F}}(\eta_1) - P_{\mathrm{F}}(\eta_2)} \geq 0.$$

□

Later, we explore additional properties of the operating characteristic associated with the likelihood ratio test.

In the meantime, we turn to an alternative formulation of binary hypothesis testing for which the LRT is the optimum decision rule.

3.2 Neyman-Pearson Binary Hypothesis Testing

When it is difficult or unnatural to choose costs for a decision problem of interest, a popular alternative formulation is to choose as the decision rule that which maximizes P_{D} subject to a constraint on the maximum allowable P_{F}, i.e.,

$$\max_{\hat{H}(\cdot)} P_{\mathrm{D}} \quad \text{subject to } P_{\mathrm{F}} \leq \alpha.$$

This is referred to as the *Neyman-Pearson* criterion.

When, as in the Bayesian case, we restrict our attention to deterministic decision rules for the time being, we have the following main result.

Theorem 1 (Neymann-Pearson Lemma)**.** *To maximize P_{D} subject to the constraint $P_{\mathrm{F}} < \alpha$, it is sufficient to use an LRT* (1) *where the threshold η is chosen such that*

$$P_{\mathrm{F}} = \mathbb{P}\left(L(\mathbf{y}) \geq \eta \mid \mathsf{H} = H_0\right) = \alpha. \tag{9}$$

Before proceeding with a proof, we emphasize what this means is that the achievable Neyman-Pearson $(P_{\mathrm{D}}, P_{\mathrm{F}})$ operating point can be read off the operating characteristic of the LRT by looking at P_{D} value at which this curve crosses the $P_{\mathrm{F}} = \alpha$ threshold.

Proof. We consider the case of continuous-valued data; the development for discrete-valued data is analogous.

We follow a straightforward Lagrange multiplier approach. To begin, let $P_{\mathrm{F}} = \alpha' \leq \alpha$, and let us consider minimizing

$$\varphi(\hat{H}) = (1 - P_{\mathrm{D}}) + \lambda(P_{\mathrm{F}} - \alpha')$$

with respect to the choice of $\hat{H}(\cdot)$, where λ is the Lagrange multiplier. To obtain our solution, it is convenient to expand $\varphi(\hat{H})$ in the following form

$$\begin{aligned} \varphi(\hat{H}) &= \int_{\mathcal{Y}_0} p_{\mathsf{y}|H}(\mathbf{y}|H_1)\,\mathrm{d}\mathbf{y} + \lambda\left[\int_{\mathcal{Y}_1} p_{\mathsf{y}|H}(\mathbf{y}|H_0)\,\mathrm{d}\mathbf{y} - \alpha'\right] \\ &= \int_{\mathcal{Y}_0} p_{\mathsf{y}|H}(\mathbf{y}|H_1)\,\mathrm{d}\mathbf{y} + \lambda\left[1 - \int_{\mathcal{Y}_0} p_{\mathsf{y}|H}(\mathbf{y}|H_0)\,\mathrm{d}\mathbf{y} - \alpha'\right] \\ &= \lambda(1-\alpha') + \int_{\mathcal{Y}_0}\Big[p_{\mathsf{y}|H}(\mathbf{y}|H_1) - \lambda\,p_{\mathsf{y}|H}(\mathbf{y}|H_0)\Big]\,\mathrm{d}\mathbf{y}. \end{aligned} \tag{10}$$

Now recall from our earlier discussion that specifying $\mathcal{Y}_0$ fully determines $\hat{H}(\cdot)$, so we can view our problem as one of determining the optimum $\mathcal{Y}_0$. From this perspective it is clear we want to choose $\mathcal{Y}_0$ so that it contains precisely those values of $\mathbf{y}$ for which the term in brackets inside the integral in (10) is negative, since this choice makes $\varphi(\hat{H})$ smallest. This statement can be expressed in the form

$$p_{\mathsf{y}|H}(\mathbf{y}|H_1) - \lambda\,p_{\mathsf{y}|H}(\mathbf{y}|H_0) \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} 0,$$

which, in turn, corresponds to an LRT; specifically,

$$\frac{p_{\mathsf{y}|H}(\mathbf{y}|H_1)}{p_{\mathsf{y}|H}(\mathbf{y}|H_0)} \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \lambda, \tag{11}$$

where λ is chosen so that $P_{\mathrm{F}} = \alpha'$. It remains only to determine α'. However, since by Fact 1 the operating characteristic of the likelihood ratio test P_{D} is a monotonically increasing function of P_{F}, the best possible P_{D} is obtained when we let $\alpha' = \alpha$. □

While a natural starting point, it turns out the Neyman-Pearson story is somewhat richer than the above development reflects. In particular, perhaps somewhat surprisingly, it turns out that in some cases restricting our attention to deterministic decision rules is limiting, i.e., better performance can be realized using a *randomized* decision rule. We develop these and other insights in the next section of the notes.