1 Decision Rules, Detection, or Hypothesis Tests

Troels Pedersen, Aalborg University, 2013.

Making decisions is an important part of daily life of anyone — also of engineers and scientists. When we say making decisions, we think of choosing between a number of alternatives on the basis of the information available at the time of decision. The alternative we choose, we call our decision.

Exercise 1. Think of a particular situation where you have to make a decision in your daily life or professionally. Which information do you base your decision on? Can you list the alternative choices you could make?

There are countless ways of making a decision. Daily life decisions can be taken in very personal ways; in most cases, there seem to be no particular rule from which you decide. You simply pick the alternative which seems most attractive at the time, based on the information available to you. Of course, given the same situation, a different person may find one of the other alternatives more attractive. A third person may find all alternatives equally attractive (or unattractive) and pick a random between any of these. Given the same information the same person may not even make the same decision if she had to make the decision again. In some case, there seems to be a particular rule or specified method to get to a decision and in some cases not. This implies that some decisions could be left to a machine while others cannot.

Exercise 2. In your example from before, can you specify a rule for the decision? If so, try to formulate such a rule.

The cases where we *can* formulate decision rules, are the most interesting and useful in engineering and we therefore exclude all other decision methods from the discussion all other decision methods. As engineers, we are mostly interested in making systems work as best they can, and we shall therefore aim at rules that make decision rules which are optimal in some sense. Of course, we have to define in *which* sense a decision rule should be optimum, or put differently, which *optimality criterion* the rule should fulfill.

In the following we formulate a first a probability model for making decisions. This helps us to derive a number of decision rules according to different optimality criteria. Since the decision theory is an ancient topic, many different traditions have evolved. Unfortunately, each tradition has developed its own terminology and therefore the topic is terminology heavy. Do not let this confuse you — the underlying ideas are actually simple (despite the obscure language).

1.1 Probability Model for Decision Problems

We base the decision on the *data*, or *observation*, denoted $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$. This data vector is deterministic but, we *model* it as a realization of a random vector, denoted $\mathbf{X} = [X_1, \ldots, X_N]^T$ with pdf $f_{\mathbf{X}}$. We use lower case notation for the known (deterministic) data $\mathbf{x}$. This helps us to distinguish it from the stochastic *model* for the data $\mathbf{X}$ in upper case.

We decide from one of the K alternatives in the set

$$\mathcal{H} = \{h_0, h_1, h_2, \ldots, h_{K-1}\}.$$

for historical reasons, each alternative h_k is called a *hypothesis*. The number K of hypotheses is at least two, but we could have any finite or (countably) infinite

	$H = h_0$	$H = h_1$	$H = h_2$	$\cdots$
$\hat{H} = h_0$	✔	✗	✗	
$\hat{H} = h_1$	✗	✔	✗	
$\hat{H} = h_2$	✗	✗	✔	
$\vdots$				

Table 1: Decision table. "True" decisions are along the main diagonal. "False" decisions or errors are on the off diagonals.

number of hypotheses. The particular indexing of the hypotheses is arbitrary and we assign no particular meaning to any of the hypothesis (e.g. to h_0).[1] The *true hypothesis* H is unknown to us and we therefore model it as a discrete random variable taking values on $\mathcal{H}$.

Definition 1. A *decision rule* is a function $\hat{H}(\mathbf{x})$ that assigns a hypothesis for each possible observation, i.e. $\hat{H} : \text{range}(\mathbf{X}) \to \mathcal{H}$.

We mark the output of a decision rule with a hat to indicate that the decision $\hat{H}(\mathbf{x})$ is not necessarily equal to the true hypothesis H. This is illustrated in the decision Table 1 where we find the true decisions on the diagonal. Since the decision is a function of the data, we should write $\hat{H}(\mathbf{x})$ all the time, but to simplify the notation we often skip the explicit mention of this dependency and write simply $\hat{H}$ instead when no confusion can occur.

Exercise 3. Benny tosses a coin 10 times and tell you how many heads he got. Now he wonders if the coin is fair or not. What are N, $\mathbf{x}$, $\mathbf{X}$, K and $\mathcal{H}$ in this case? Can you help him to make a decision rule.

Before making the observation, we can speak of the *a priori*[2] probability of a hypothesis to be true. We can then say that the true hypothesis H is h_k with probability $P_k = \mathbb{P}(H = h_k)$. Sometimes these probabilities are also called *prior information.* By definition, the prior probabilities sum to one:

$$P_0 + P_1 + \cdots + P_{K-1} = 1. \tag{1}$$

After making the observation, our view on the probability for a hypothesis to be true changes. It is now $P(H = h_k|\mathbf{x})$, i.e. the conditional probability for $H = h_k$ given observation $\mathbf{x}$. This is called the *a posteriori*[3] probability for h_k to be true. As a consequence, we can write the probability for our decision $\hat{H}$ to be correct (equal to H) as

$$\mathbb{P}(\text{"}\hat{H} \text{ is true given } \mathbf{x}\text{"}) = \mathbb{P}(\hat{H} = H|\mathbf{x}) \tag{2}$$

or conversely, $\mathbb{P}(\text{"}\hat{H} \text{ is not true given } \mathbf{x}\text{"}) = \mathbb{P}(\hat{H} \neq H|\mathbf{x}) = 1 - \mathbb{P}(\hat{H} = H|\mathbf{x})$.

The random vector $\mathbf{X}$ depends on which of the hypothesis are true. Therefore, we work with the conditional pdf of $\mathbf{X}$ under hypothesis h_k and this is denoted $f_{\mathbf{X}|H}(\mathbf{x}|h_k)$. This conditional pdf is called the *likelihood* for h_k. We can use Bayes'

[1] This is in contrast to the tradition in some fields of statistics, where special meaning is assigned to e.g. hypothesis h_0. We avoid this practice since it unnecessary from a formal point of view, and in some situations makes it harder to understand decision problems.

[2] A priori is from Latin and means "from the earlier". We use it in the sense "before the experiment" or "without observing the data".

[3] Again from Latin meaning "from the later". We use in the sense "after the experiment".

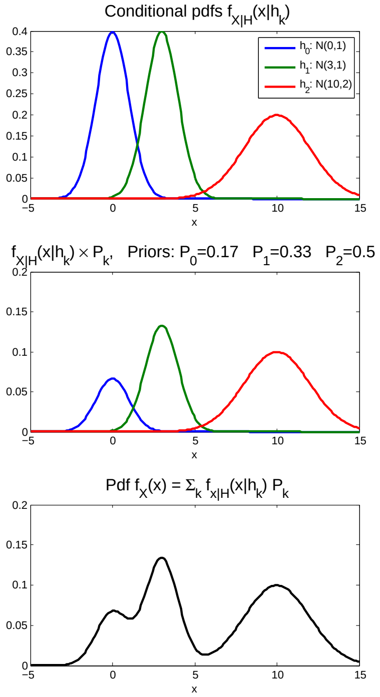

Figure 1: Example with $K = 3$ where $f_{\mathbf{X}|H}$ are Gaussian pdfs with parameters as given in the upper panel and prior probabilities given in the midle panel.

Theorem[4] to connect the prior, posterior and likelihoods:

$$\mathbb{P}(H = h_k|\mathbf{x}) = \frac{f_{\mathbf{X}|H}(\mathbf{x}|h_k)P_k}{f_{\mathbf{X}}(\mathbf{x})} \tag{3}$$

The denominator $f_{\mathbf{X}}(x)$ is the same regardless of which h_k we consider and will not enter the decision rules to be derived.[5]

If we consider the decision rule evaluated at the random data $\mathbf{X}$, then the decision $\hat{H}(\mathbf{X})$ is a random variable. The probability of error P_e is then defined as the probability that $\hat{H}(\mathbf{X}) \neq H$:

$$P_e := \mathbb{P}(\hat{H}(\mathbf{X}) \neq H) = 1 - \underbrace{\mathbb{P}(\hat{H}(\mathbf{X}) = H)}_{P_c}, \tag{4}$$

i.e., one minus the probability of a “true” or “correct” decision.

Exercise 4. How many of the above definitions and concepts can you relate the example in Figure 1?

[4] Named after its discoverer Thomas Bayes. You may have seen the proof in probability theory for probabilities of events A and B: $\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$. The version in (3) involves pdfs, but in principle the proof is the same.

[5] Of course, it can be obtained from the likelihoods and prior probabilities as $f_{\mathbf{X}}(\mathbf{x}) = \sum_{k=1}^{K} f_{\mathbf{X}|H}(\mathbf{x}|h_k)P_k$. To compute this expression is a pain in the neck – the shape of this pdf is typically very complicated (see e.g. Figs. 1, 2, 3), so it is very fortunate that we need not bother about computing it!

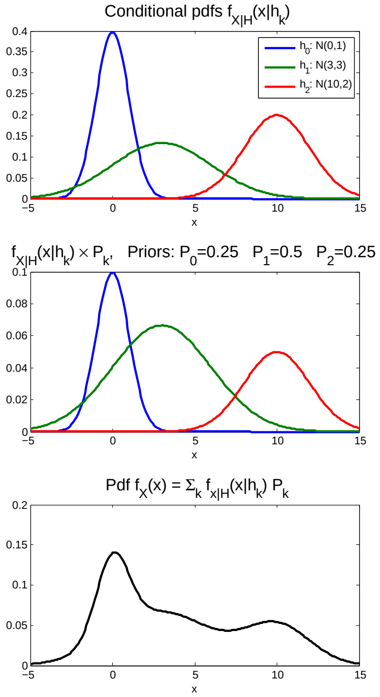

Figure 2: Example with $K = 3$ where $f_{\mathbf{X}|H}$ are Gaussian pdfs with parameters as given in the upper panel and equal prior probabilities.

1.2 Maximum A Posteriori Probability (MAP) Decision Rule

Let us define our first decision rule: Let us pick the hypothesis which has the largest probability of being correct given the data. To make it seem more fancy we call this the “Maximum A posteriori Probability” (MAP) rule:[6]

Definition 2. The Maximum A posteriori Probability (MAP) decision rule:

$$\hat{H}_{\mathrm{MAP}}(\mathbf{x}) := \arg\max_{h_k \in \mathcal{H}} \mathbb{P}(H = h_k|\mathbf{x}) \tag{5}$$

where arg max means “the argument which maximizes”.

The MAP rule can be stated in an alternative (equivalent) form

$$\hat{H}_{\mathrm{MAP}}(\mathbf{x}) = \arg\max_{h_k \in \mathcal{H}} f_{\mathbf{X}|H}(\mathbf{x}|h_k)P_k \tag{6}$$

Exercise 5. Show that (6) is actually the same as (5).

The alternative form (6) shows how the a priori probabilities affect the decision. If $P_k = 0$, the MAP decision will not be h_k, no matter how high the likelihood for

[6] Please do not be scared by the fancy name or the strange mathematical expression. It does exactly pick the hypothesis which is most probably the correct one given the data!

h_k is. In the other extreme, if $P_k = 1$, all other hypothesis have prior probability zero (why is that?), and the MAP rule will always pick h_k. The effect is in many cases reasonable—it essentially means that the more probable a hypothesis is before we observe the data, the more convincing the data needs to be in order to make us change our mind and decide for another hypothesis

Exercise 6. Indicate on Fig. 1 the regions of the x-axis, for which the MAP rule will select h_0, h_1 and h_2, respectively. These regions are called *decision regions.*

Exercise 7. Do the exercise again for Fig. 3. Note the decision region for h_1 is very different from the one in Fig. 1.

By definition, the MAP rule maximizes the probability of a true decision which is the same thing as minimizing the probability of an error. Furthermore, since the MAP rule minimize the probability of error for any $\mathbf{x}$, it also minimizes the probability error when averaged with respect to $f_{\mathbf{X}}$.

$$P_e = 1 - \mathbb{P}(\hat{H}(\mathbf{X}) = H) = 1 - \int \underbrace{\mathbb{P}(\hat{H}(\mathbf{x}) = H|\mathbf{x})}_{\text{Max by def. of MAP}} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \tag{7}$$

Often this important fact is taken as the starting point or optimality criterion for deriving the MAP rule. You would then say that the MAP rule is derived to minimize the probability of error. Obviously this is the same as opting for maximizing the probability of a true decision.

1.3 No prior information: Maximum Likelihood (ML) Decision Rule

The MAP rule is derived under the assumption that we know the apriori probabilities $P_0, \ldots, P_{K-1}$. In some situations these probabilities are unavailable or we may desire not to use them. One possibility is then to assume all hypothesis to be equally probable, which is the same as assuming $P_0 = P_1 = \cdots = P_{K-1} = 1/K$ in the MAP rule. This leads to the Maximum Likelihood (ML) decision rule: pick the hypothesis with the highest likelihood given the data or put in a formula,

Definition 3. The Maximum Likelihood (ML) decision rule:

$$\hat{H}_{\mathrm{ML}}(\mathbf{x}) = \arg\max_{h_k \in \mathcal{H}} f_{\mathbf{X}|H}(\mathbf{x}|h_k) \tag{8}$$

Exercise 8. Mark the decision regions of the ML rule on the plot in Figure 3. Discuss how they are different from the MAP case.

Exercise 9. Let $P_{e,\mathrm{MAP}}$ and $P_{e,\mathrm{ML}}$ be the average error probabilities for the MAP and ML rules, respectively. Which one is smallest? When are they equal?

1.4 Minimizing Costs: Bayes' Decision Rule

Making decisions, particularly making wrong ones, may be costly in terms of money, time, energy, etc. Therefore, it may be a reasonable optimality criterion for a decision rule to choose the hypothesis such that the cost is minimized. In contrast to the MAP rule, which maximized the probability of a true decision, we now want to consider the cost of our decisions. This idea will lead us to the so-called Bayes' decision rule.[7]

[7] Named after its inventor Thomas Bayes, who also discovered Bayes' Theorem. Do not confuse the Bayes' decision rule with Bayes' Theorem).

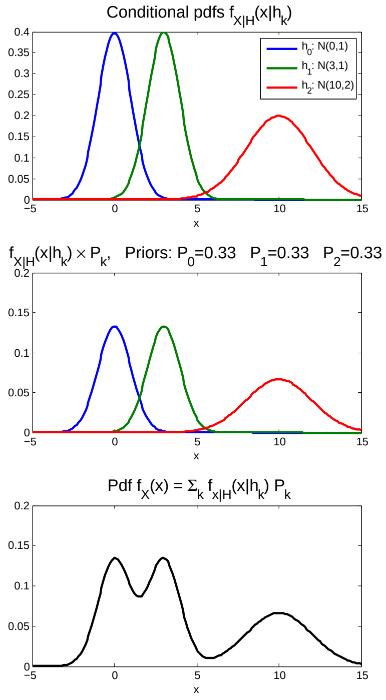

Figure 3: Example with $K = 3$ where $f_{\mathbf{X}|H}$ are Gaussian pdfs with parameters as given in the upper panel and equal prior probabilities.

	$H = h_0$	$H = h_1$	$H = h_2$	$\cdots$
$\hat{H} = h_0$	C_{00}	C_{01}	C_{02}	
$\hat{H} = h_1$	C_{10}	C_{11}	C_{12}	
$\hat{H} = h_2$	C_{20}	C_{21}	C_{22}	
$\vdots$				

Table 2: Cost table or cost matrix for a decision rule. Typically, the entries along the diagonal corresponding to the "true" decisions are smallest (lead to the lowest cost). The matrix is $K \times K$ and asymmetric in general.

We assign costs the decisions as indicated in Table 2, i.e we let $C_{kk'}$ be the cost if we decide h_k if $h_{k'}$ is true. We now seek a decision rule to minimize the cost rather than the probability or error. Often the cost of making a true decision is less than that of making an error and therefore, the costs along the diagonal of the "Cost matrix" are lower than the off diagonal costs. However, this is not necessary for our derivations, so we will not need to assume that (yet). The expected cost for decision h_k given the data (sometimes called the risk) is obtained as the sum of the costs weighted by the conditional probability for hypothesis H given $\mathbf{x}$:

$$C(h_k|\mathbf{x}) := \sum_{k'=0}^{K-1} C_{kk'}\mathbb{P}(H = h_{k'}|\mathbf{x}) \tag{9}$$

Exercise 10. Explain the meaning of (9) to your colleague. Make use of the Cost matrix in Table 2.

Bayes' decision rule is then to choose the hypothesis leading to the lowest expected cost. We can formulate this in mathematical notation as:

Definition 4. Bayes' decision rule

$$\hat{H}_{\text{Bayes}}(\mathbf{x}) = \arg\min_{h_k\in\mathcal{H}} C(h_k|\mathbf{x}) \tag{10}$$

Because Bayes' decision rule (10) minimizes the cost for each specific value $\mathbf{x}$, it also minimizes the average cost defined as $\bar{C} = \mathbb{E}[C(\hat{H}_{\text{Bayes}}|\mathbf{X})]$.

Exercise 11. Write out (9) and (10) for $K = 2$.

Exercise 12. Bayes decision rule reduces to the MAP by assigning the costs such that the cost matrix equals negative the identity matrix, i.e. $C_{kk} = -1$ and $C_{kk'} = 0, k \neq k'$. Check that this is true.
Hint: $\mathbb{P}(H = h_k|\mathbf{x}) = 1 - \mathbb{P}(H \neq h_k|\mathbf{x}) = 1 - \sum_{k'\neq k}\mathbb{P}(H = h_{k'}|\mathbf{x})$.

1.5 Special Case: Binary Decision Rules

A decision problem with only $K = 2$ alternatives is called a *binary decision* problem. The decision rules derived thus far can be simplified in the binary case This yields forms which are simpler to implement. However, these reduced forms makes it harder to spot the original ideas behind the decision rule.

It turns out that any reasonable binary decision rule can be written in a standard form where a *likelihood ratio* denoted by $\Lambda(\mathbf{x})$ is compared to a *threshold* γ which does not depend on $\mathbf{x}$:[8]

$$\hat{H} = \begin{cases} h_0, & \text{if } \Lambda(\mathbf{x}) \geq \gamma \\ h_1 & \text{otherwise.} \end{cases}, \qquad \text{with} \qquad \Lambda(\mathbf{x}) := \frac{f_{\mathbf{X}|H}(\mathbf{x}|H_0)}{f_{\mathbf{X}|H}(\mathbf{x}|H_1)}. \tag{11}$$

Only $\Lambda(\mathbf{x})$ depends on the data whereas γ only depends on a priori information and can be determined beforehand. The decision rules differ only in the choice of threshold value γ. We now show what the MAP, ML and Bayes decision rules may be put in the standard form and derive the corresponding thresholds.

For the binary case, the MAP takes value h_0 if $f_{\mathbf{X}|H}(\mathbf{x}|h_0)P_0 > f_{\mathbf{X}|H}(\mathbf{x}|h_1)P_1$ and h_1 if $f_{\mathbf{X}|H}(\mathbf{x}|h_0)P_0 < f_{\mathbf{X}|H}(\mathbf{x}|h_1)P_1$. The case the two terms are equal we can

[8] Some authors do not follow the same convention for the definition of $\Lambda(\mathbf{x})$ as we do here, but define it as the as the inverse. Therefore, to be sure, you always need to check the definition in the particular references you use.

choose freely. For simplicity, we simply chose h_0 in case the two terms are equal. This rule can be more compactly written as[9]

$$\hat{H}_{\mathrm{MAP}} = \begin{cases} h_0, & \text{if } f_{\mathbf{X}|H}(\mathbf{x}|h_0)P_0 \geq f_{\mathbf{X}|H}(\mathbf{x}|h_1)P_1 \\ h_1 & \text{otherwise.} \end{cases} \tag{12}$$

Exercise 13. Convince yourself that (12) is equivalent to (6) for $K = 2$.

Exercise 14. What is the probability that $f_{\mathbf{X}|H}(\mathbf{x}|h_0)P_0 = f_{\mathbf{X}|H}(\mathbf{x}|h_1)P_1$?

The inequality in (12) can be rearranged as

$$\underbrace{\frac{f_{\mathbf{X}|H}(\mathbf{x}|h_0)}{f_{\mathbf{X}|H}(\mathbf{x}|h_1)}}_{\Lambda(\mathbf{x})} \geq \underbrace{\frac{P_1}{P_0}}_{\gamma_{\mathrm{MAP}}} \tag{13}$$

Consequently, the *binary MAP rule* reads in the standard form

$$\hat{H}_{\mathrm{MAP}} = \begin{cases} h_0, & \text{if } \Lambda(\mathbf{x}) \geq \gamma_{\mathrm{MAP}} \\ h_1 & \text{otherwise.} \end{cases} \tag{14}$$

Exercise 15. Show that the *binary ML decision rule* has $\gamma_{\mathrm{ML}} = 1$, i.e.,

$$\hat{H}_{\mathrm{ML}} = \begin{cases} h_0, & \Lambda(\mathbf{x}) \geq 1 \\ h_1, & \text{otherwise.} \end{cases} \tag{15}$$

The binary Bayes rule can also be brought in the standard form. The binary Bayes decision rule can be stated as

$$\hat{H}_{\mathrm{Bayes}} = \begin{cases} h_0, & C(h_1|\mathbf{x}) \geq C(h_0|\mathbf{x}) \\ h_1, & \text{otherwise.} \end{cases} \tag{16}$$

Inserting the two conditional costs

$$C(h_0|\mathbf{x}) = C_{00}\mathbb{P}(H = h_0|\mathbf{x}) + C_{01}\mathbb{P}(H = h_1|\mathbf{x}) \tag{17}$$

$$C(h_1|\mathbf{x}) = C_{10}\mathbb{P}(H = h_0|\mathbf{x}) + C_{11}\mathbb{P}(H = h_1|\mathbf{x}) \tag{18}$$

in the inequality of (16), using Bayes rule and rearranging the terms we have (assuming that $(C_{00} - C_{01}) > 0$)

$$\underbrace{\frac{f_{\mathbf{X}|H}(\mathbf{x}|h_0)}{f_{\mathbf{X}|H}(\mathbf{x}|H_1)}}_{\Lambda(\mathbf{x})} \geq \underbrace{\frac{(C_{11} - C_{01})P_1}{(C_{00} - C_{10})P_0}}_{\gamma_{\mathrm{Bayes}}} \tag{19}$$

Finally, we see that the *Bayes decision rule for the binary case* is of the same form

$$\hat{H}_{\mathrm{Bayes}} = \begin{cases} h_0 & \Lambda(\mathbf{x}) \geq \gamma_{\mathrm{Bayes}} \\ h_1 & \text{otherwise.} \end{cases} \tag{20}$$

Exercise 16. It is often the case that the entries of $\mathbf{X}$ are assumed i.i.d. In this case, the pdf of $\mathbf{X}$ conditioned on H factorizes as $f_{\mathbf{X}|H}(\mathbf{x}|h_k) = \prod_{n=1}^{N} f_{X|H}(x_n|h_k)$. Show for this case that the likelihood factorizes as

$$\Lambda(\mathbf{x}) = \prod_{n=1}^{N} \frac{f_{X|H}(x_n|h_0)}{f_{X|H}(x_n|h_1)} \tag{21}$$

[9] There exists other specialized notations in the literature, but they all mean the same thing.

In some cases, computations needed in the decision rule may be simplified by transforming the decision inequality. In doing so, we must take care that the inequality is maintained. This is the case if the transformation is monotonously increasing. A transformation g : is monotone if

$$a > b \Leftrightarrow g(a) > g(b), \tag{22}$$

The most common transform to apply is the logarithm. Thus taking the logarithm on both sides of the decision inequality does not change the decision regions and the decision rule is equivalent.

Exercise 17. Derive the log-likelihood of the problem in Exercise 16. State the ML decision rule in the logarithmic version (take the logarithm of both sides of the inequality).

1.6 Some Remarks on Decision Theory

You may have wondered about the use of the "hat" notation for the decision rule, and found odd, since we also use this notation in estimation theory. This notation is chosen intentionally to show this similarity. In Estimation Theory, we attempt to estimate the value of some unknown parameter θ which is considered to be in a subset of the real line (in the scalar case, that is). The estimate is a function of the data denoted by $\hat{\theta} = g(\mathbf{x})$. A decision rule is also a function of the data $\hat{H}(\mathbf{x})$ so wherein lies the difference? Well, the crucial difference is that the set of hypothesis is not a continuum as the parameter space is, but is a countable set. Therefore, we think of it as a discrete random variable. In the estimation case, we seek estimators, which are unbiased and have low mean squared error (MSE). In the case of hypothesis, the notion of bias or MSE does not make sense. Instead, we can operate with the probability of making a correct decision.

It should be remarked that there are many other decision rules than the three we have considered here. Two prominent examples of other decision rules are the Neyman-Pearson rule and the Mini-max rule.

1.7 Further reading

Decision rules hypothesis tests are covered in many Textbooks on Statistics. It also appears in books specialized for certain applications, e.g. radar theory, pattern recognition, machine learning, communication theory, to name a few. A fairly broad introduction from a signal processing point of view is given in the reference below.

- Steven M. Kay, "Fundamentals of Statistical Signal Processing: Vol. II Detection Theory", Prentice Hall, 1998.