Hypothesis Testing

Chapter Intended Learning Outcomes:

(i) Understand the basics of significance testing and binary detection

(ii) Able to apply probability and random variables to formulate and solve simple significance test and binary detection problems

(iii) Able to generate receiver operating characteristic curve

Significance Testing and Detection

A hypothesis is a known probability model, e.g., $\mathcal{N}(5, 10)$ and $\mathcal{U}(-10, 10)$.

Significance testing refers to accepting or rejecting the hypothesis.

The hypothesis is typically called null hypothesis, denoted by $\mathcal{H}_0$. A significance test is performed to determine whether $\mathcal{H}_0$ remains unchanged after a change of the experiment conditions: Accept $\mathcal{H}_0$ if there is high chance that the change has no effect on the probability model; Reject otherwise.

For example, a company wants to check if their weight-loss pills are effective. A group of people are invited to try the pills and $\mathcal{H}_0$ will be accepted if the pills are ineffective.

Detection refers to deciding among two or more possible hypotheses given one or multiple observations.

In binary detection or hypothesis testing, we need to choose between 2 possibilities.

A representative example is signal detection whose task is to choose between the null (noise-only) hypothesis $\mathcal{H}_0$ and alternative (signal-present) hypothesis $\mathcal{H}_1$ given a time segment of $x(t)$:

$$\begin{aligned} \mathcal{H}_0 &: x(t) = n(t), \\ \mathcal{H}_1 &: x(t) = s(t) + n(t), \quad 0 < t < T \end{aligned} \tag{5.1}$$

where $s(t)$ and $n(t)$ denote the signal and noise, respectively.

For multiple hypothesis testing, an example can be deciding which word has been spoken from the admissible set of {"0", "1",..., "9"} given N samples of a recorded speech $x[n]$. That is, we need to choose between 10 possibilities:

$$\begin{aligned}
&\mathcal{H}_0 : x[n] = v_0[n] + w[n], \\
&\mathcal{H}_1 : x[n] = v_1[n] + w[n], \\
&\cdots \quad \cdots \\
&\mathcal{H}_9 : x[n] = v_9[n] + w[n], \quad n = 0, 1, \cdots N-1
\end{aligned} \tag{5.2}$$

where $v_0[n]$, $v_1[n]$,..., $v_9[n]$ represent the speech waveforms of "0", "1",..., "9", respectively, while $w[n]$ is background noise.

Apart from using hypothesis testing and signal detection to describe such problems, classification and discrimination are also commonly used.

Example 5.1

$X_1, \cdots, X_N$ are N independent and identically distributed (IID) zero-mean Gaussian random variables with unknown variance σ^2. Illustrate the difference between significance test and hypothesis test.

For significance test, an example is: whether we accept or reject the hypothesis $\mathcal{H}_0$ that the variance is $\sigma^2 = 10$. That is, $\mathcal{H}_0 : \{X_1, \cdots, X_N\} \sim \mathcal{N}(0, 10)$.

For hypothesis test, an example is: given $X_1, \cdots, X_N$, we need to choose one of the following 3 hypotheses:

$$\mathcal{H}_1 : \{X_1, \cdots, X_N\} \sim \mathcal{N}(0, 10)$$
$$\mathcal{H}_2 : \{X_1, \cdots, X_N\} \sim \mathcal{N}(0, 20)$$
$$\mathcal{H}_3 : \{X_1, \cdots, X_N\} \sim \mathcal{N}(0, 30)$$

Performance Metric for Significance Testing

In significance test, there is only one accuracy measure, called significance level denoted by α.

The sample space S is divided into the acceptance and rejection sets A and R, respectively, such that $A \cup R = S$.

α is the probability of rejecting the hypothesis when it is true:

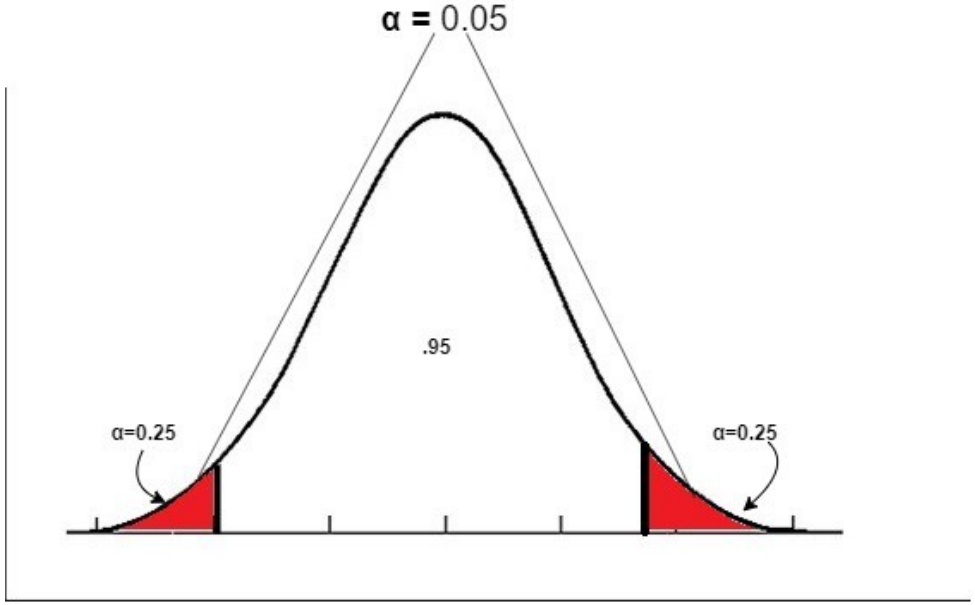
$$\alpha = P(s \in R) \tag{5.3}$$

That is, the observation s corresponds to $\mathcal{H}_0$ but we reject $\mathcal{H}_0$.

Note that the probability of false rejection of $\mathcal{H}_0$, α, is under our control while we have no control on the probability of accepting $\mathcal{H}_0$ when it is false.

Example 5.2

Suppose that on Thursdays between 9:00pm and 9:30pm, the number of tweets in US can be approximately modeled as $T \sim \mathcal{N}(10^7, 10^7)$. The President plans to deliver a 30 min. speech at 9:00pm next Thursday that will be broadcasted on all networks. The null hypothesis $\mathcal{H}_0$ is that the speech does not affect the probability model of tweets. Design a significance test for $\mathcal{H}_0$ at $\alpha = 0.05$.



Hypothesis Testing — 2-tailed test | by Tanwir Khan | Towards Data Science

Now we set:

$$\alpha = P(s \in R) = 0.05$$

The President speech can increase or decrease the number of tweets. Let $10^7 \pm c$ be acceptance region. Then we can formulate and solve the problem using MATLAB `norminv`:

$$\begin{aligned}
& \alpha = P(|T - 10^7| \geq c) = 0.05 \\
\Rightarrow\ & P(T - 10^7 \geq c) = 0.025 \\
\Leftrightarrow\ & P(T - 10^7 < c) = 0.975 = \int_{-\infty}^{10^7+c} \frac{1}{\sqrt{2\pi \cdot 10^7}} e^{-\frac{1}{2\cdot 10^7}\left(u-10^7\right)^2} du \\
\Rightarrow\ & c = 6198
\end{aligned}$$

```
norminv(0.975, 1e7, sqrt(1e7))-1e7
= 6.1980e+03
```

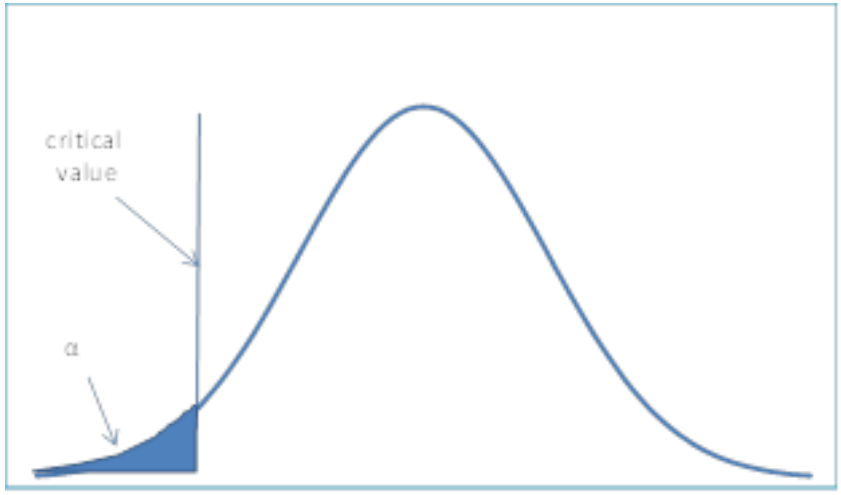
That is, if T is outside $[10^7 - 6198, 10^7 + 6198]$, $\mathcal{H}_0$ is rejected.

This corresponds to two-tail significance test.

What are the differences if a smaller α is used?

Example 5.3

Prior to releasing a diet pill to the public, a drug company runs a test on a group of 64 people. Before testing, the probability model for the people weight is IID Gaussian random variable in pounds $\mathcal{N}(190, 576)$. The null hypothesis $\mathcal{H}_0$ is that the pill is ineffective to reduce body weight. Design a significance test for $\mathcal{H}_0$ at $\alpha = 0.01$ using the sample mean.



Left tailed significance testing | Real Statistics Using Excel (real-statistics.com)

Let X be the sample mean. According to Example 3.11, $\mathbb{E}\{X\} = 190$ and $\text{var}(X) = 576/64 = 9$.

Let c be the weight boundary such that $\mathcal{H}_0$ is rejected if $X \leq c$.

Then we can formulate and solve the problem as follows:

$$\alpha = P(X \leq c) = 0.01 = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi \cdot 9}} e^{-\frac{1}{18}(u-190)^2} du$$

$$\Rightarrow c = 183.02$$

```
norminv(0.01,190,3)
 = 183.0210
```

That is, if $X \leq 183.02$ pounds, then $\mathcal{H}_0$ is rejected or the pill is effective to reduce body weight.

This corresponds to one-tail or left-tail significance test.

Note that another one-tail test refers to the right-tail case.

Performance Metrics for Binary Detection

We first consider the problem of signal detection in noise.

1. Probability of detection

Probability of making the correct decision given that the signal is present, i.e., when signal is present, we decide that signal is present.

Denote the event of making the signal presence decision as T, we can write $P_{\mathrm{D}} = P(T|\mathcal{H}_1)$, which is also sensitivity (靈敏度).

2. Probability of false alarm

Probability of making the incorrect decision when the signal is absent, i.e., when there is no signal, we decide that signal is present.

We can express it as $P_{\mathrm{FA}} = P(T|\mathcal{H}_0)$, which is also known as Type I error, and can be computed as 1-specificity (特異度).

Note that the false rejection of $\mathcal{H}_0$ in significance testing, α, corresponds to Type I error.

3. Probability of miss

Probability of making the incorrect decision when signal is present, i.e., when signal is present, we decide that there is no signal.

We can express it as $P_{\mathrm{M}} = P(\overline{T}|\mathcal{H}_1)$, which is also known as Type II error, and can be computed as $1 - P_{\mathrm{D}}$ or 1-sensitivity.

An ideal detector should have:

- Largest P_{D}
- Smallest P_{FA} and P_{M} (if P_{D} is considered, P_{M} is not required as maximizing P_{D} implies minimizing P_{M})

However, it is not possible to reduce P_{FA} and P_{M} at the same time under uncertainty.

We start with a simple problem based on (5.1) in deciding whether there is a constant $A = 1$ using single observation $x[0]$:

$$\begin{aligned} \mathcal{H}_0 &: x[0] = w[0], \\ \mathcal{H}_1 &: x[0] = A + w[0] \end{aligned} \tag{5.4}$$

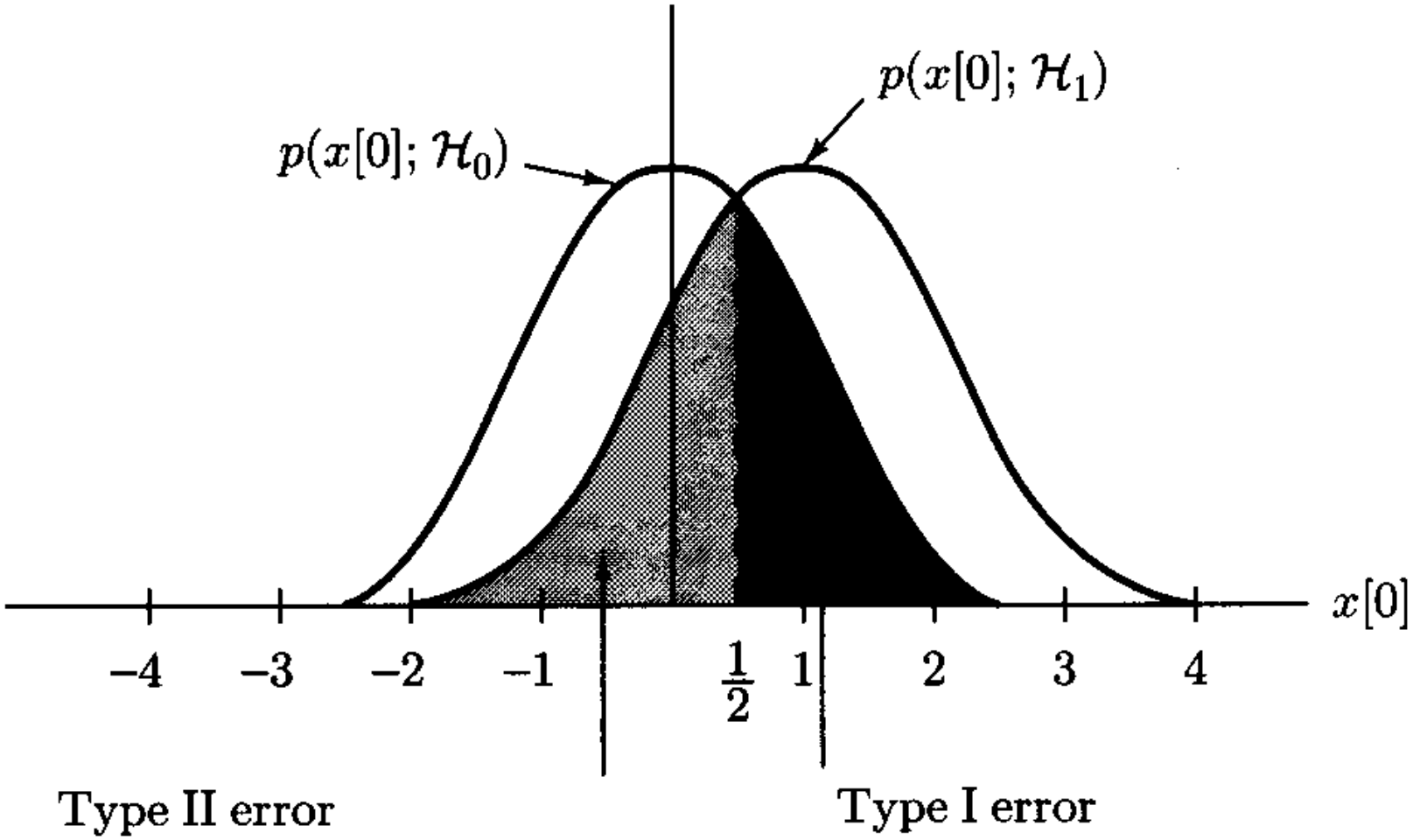
where $w[0] \sim \mathcal{N}(0, 1)$.

Using (2.13), the probability density function (PDF) of $x[0]$ under each of the two hypotheses is constructed as:

$$\begin{aligned} P_{x[0]|\mathcal{H}_0}(x[0]) &= p(x[0]; \mathcal{H}_0) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2[0]} \\ P_{x[0]|\mathcal{H}_1}(x[0]) &= p(x[0]; \mathcal{H}_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x[0]-1)^2} \end{aligned} \tag{5.5}$$

A simple detector is to utilize $x[0]$ directly. That is we choose $\mathcal{H}_1$ if $x[0] > \gamma$ and choose $\mathcal{H}_0$ if $x[0] < \gamma$ where γ is the threshold.

When $\gamma = 0.5$:



$P_{\text{FA}} = P_{\text{M}}$

P_{D}, P_{FA} and P_{M} at $\gamma = 0.5$ are:

$$P_{\mathrm{D}} = \int_{0.5}^{\infty} p(x; \mathcal{H}_1) dx = \int_{0.5}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-1)^2} dx$$

$$\begin{aligned} P_{\mathrm{M}} &= \int_{-\infty}^{0.5} p(x; \mathcal{H}_1) dx = \int_{-\infty}^{0.5} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-1)^2} dx \\ &= \int_{0.5}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy, \quad -y = x - 1 \end{aligned}$$

$$P_{\mathrm{FA}} = \int_{0.5}^{\infty} p(x; \mathcal{H}_0) dx = \int_{0.5}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

Recall that the cumulative distribution function (CDF) of Gaussian random variable:

$$F(\gamma) = P(X \leq \gamma) = \int_{-\infty}^{\gamma} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

can be computed using MATLAB command `normcdf`.

For P_{M}, we use `normcdf(0.5,1,1)` where the first and second `1` correspond to mean and standard deviation, respectively.

```
>> normcdf(0.5,1,1)
ans = 0.3085
```

For P_{D}:

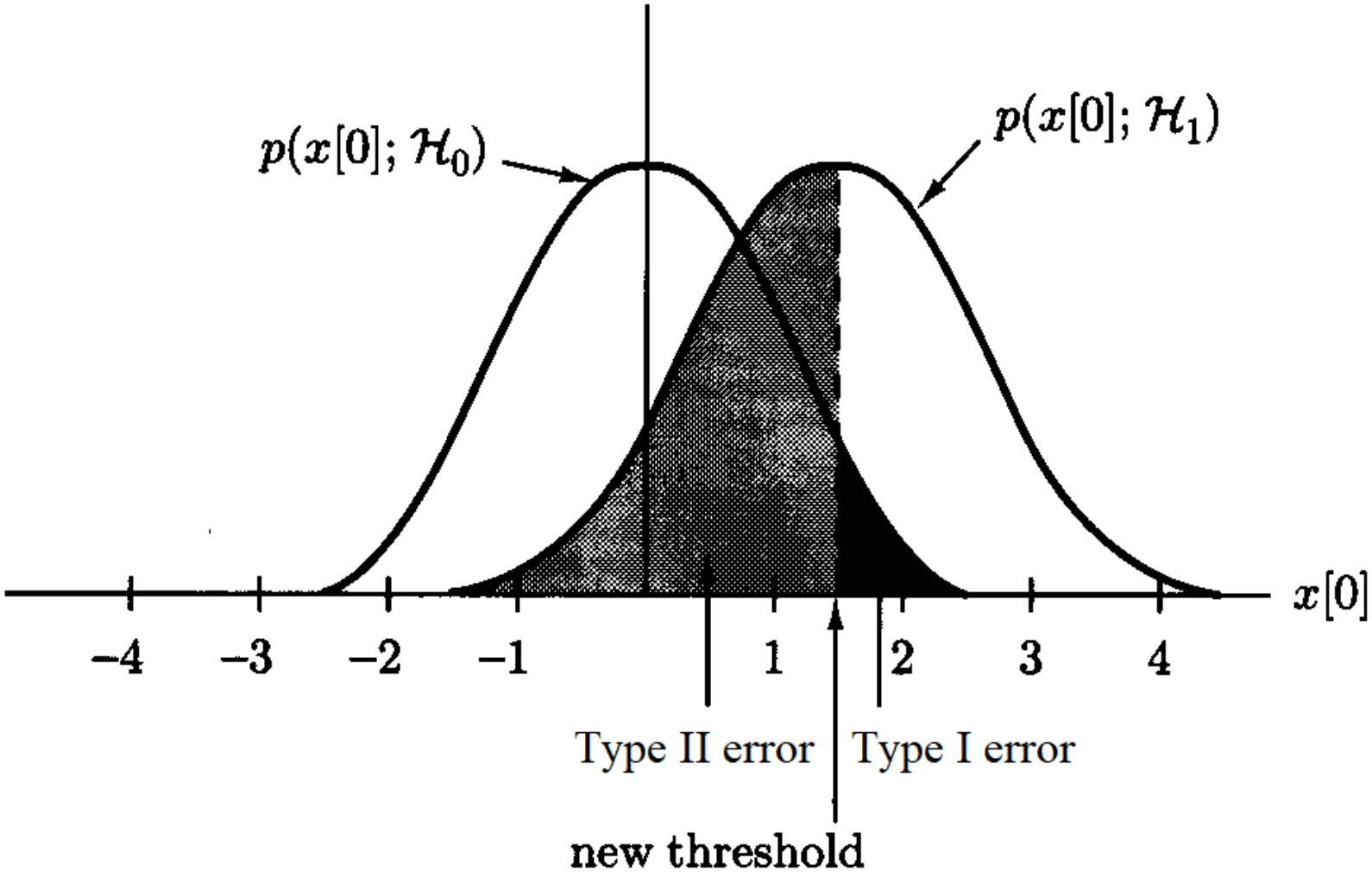
```
>> 1-normcdf(0.5,1,1)
ans = 0.6915
```

For P_{FA}:

```
>> 1-normcdf(0.5,0,1)
ans = 0.3085
```

When the standard deviation is increased from 1 to 10, what do you expect about the probability of detection, probability of false alarm and probability of miss?

When $\gamma > 0.5$:



$\downarrow P_{\text{FA}}$ but $\uparrow P_{\text{M}}$

$\gamma = 0.7$:

For P_{M}:

```
>> normcdf(0.7,1,1)
ans = 0.3821
```

For P_{D}:

```
>> 1-normcdf(0.7,1,1)
ans = 0.6179
```

For P_{FA}:

```
>> 1-normcdf(0.7,0,1)
ans = 0.2420
```

$\gamma = 0.3$:

For P_{D}:

```
>> 1-normcdf(0.3,1,1)
ans = 0.7580
```

For P_{FA}:

```
>> 1-normcdf(0.3,0,1)
ans = 0.3821
```

Because of the tradeoff between P_{D} and P_{FA}, we may determine γ for a fixed value of P_{FA} or P_{D}.

In military, a very small value such as $P_{\mathrm{FA}} = 10^{-8}$ is required because if we falsely decide an enemy aircraft is present, an attack is initiated.

In medicine, we need to ensure a high P_{D} and thus a larger P_{FA} may be allowed (in Example 1.16, $P_{\mathrm{FA}} = 10^{-1}$ and $P_{\mathrm{D}} = 0.9$ for a certain γ depending on how to interpret the test results)

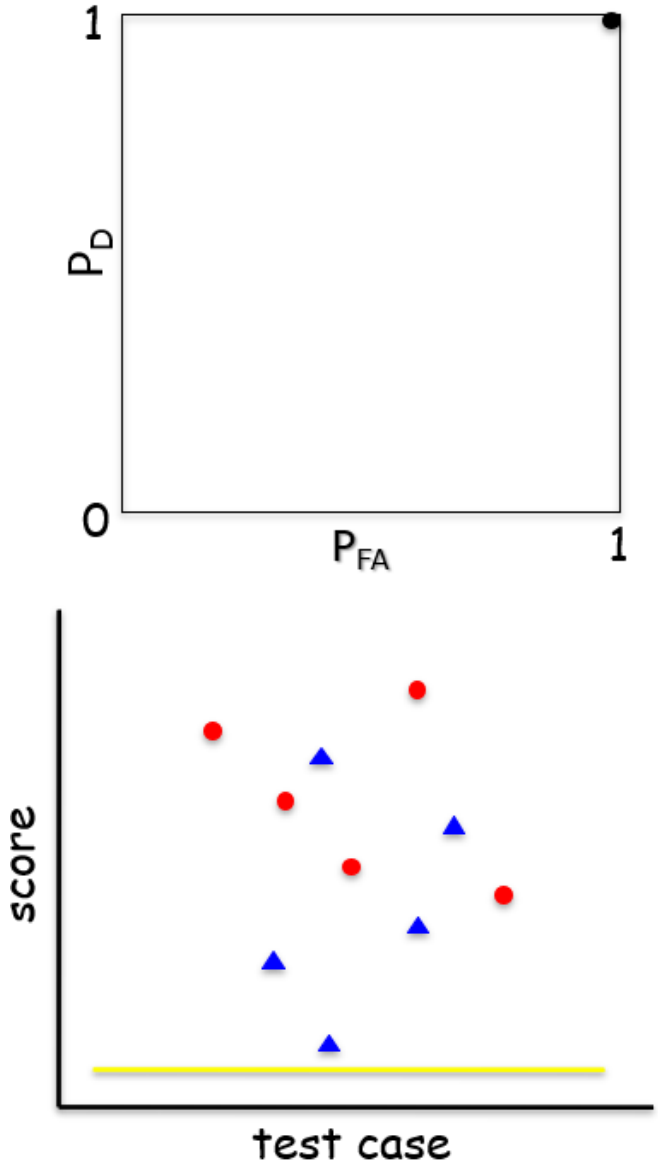
Another example is a typical Covid-19 rapid antigen test kit with sensitivity of 98.1% and specificity of 99.8%:

120007883-v1-Panbio-COVID-19-Ag-Nasal-AsymptomaticSe.pdf (abbott.com)

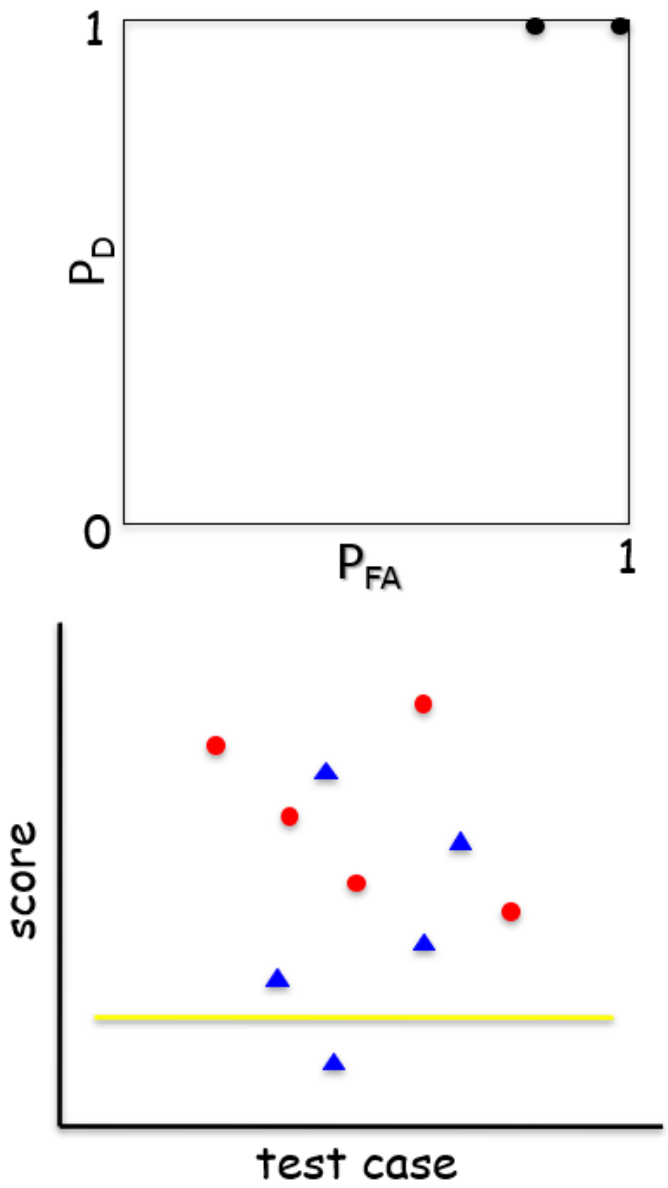
To assess the performance of a detector, receiver operating characteristic (ROC) can be used, which is a curve of P_{D} versus P_{FA}. Hence we can use the ROC to determine the desired operating point $(P_{\mathrm{FA}}, P_{\mathrm{D}})$.

If the expressions of P_{D} and P_{FA} are known, the ROC may be calculated in a theoretical manner. But it is simpler to generate numerous data corresponding to $\mathcal{H}_0$ and $\mathcal{H}_1$, and evaluate experimental probabilities at different thresholds.

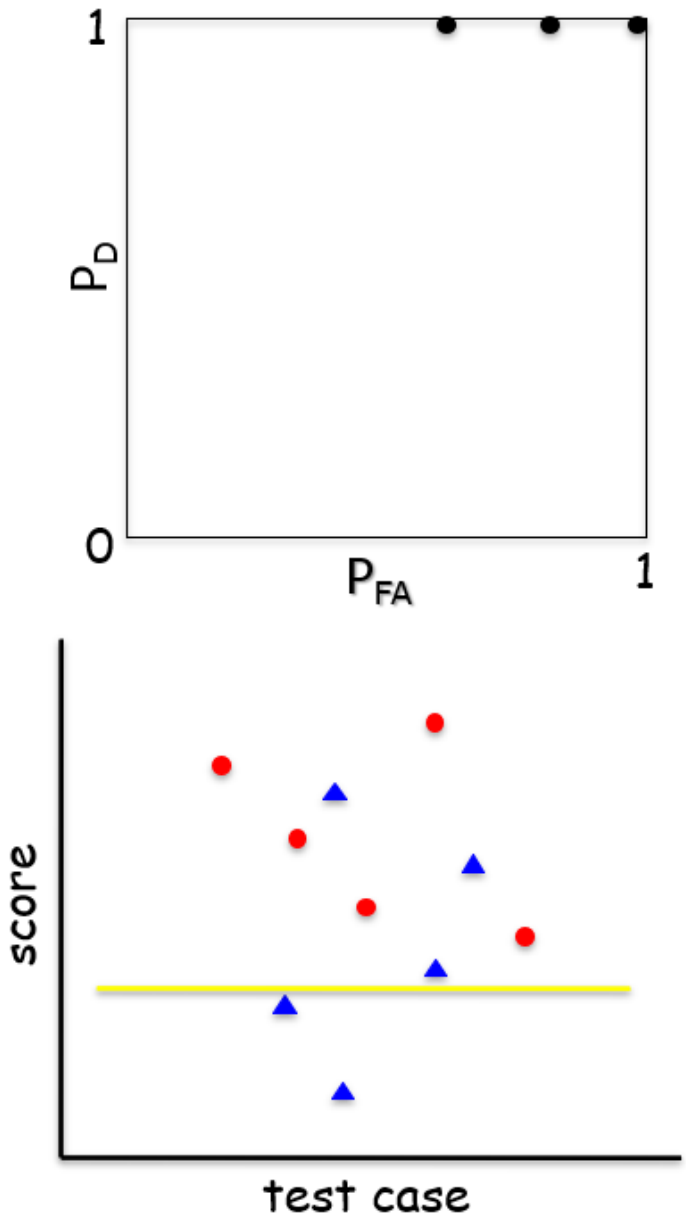
An illustration of constructing a ROC using 5 data from both $\mathcal{H}_0$ (blue) and $\mathcal{H}_1$ (red), i.e., 10 test cases, by considering the threshold value from small to large is given as follows. In practice, much more data are needed.



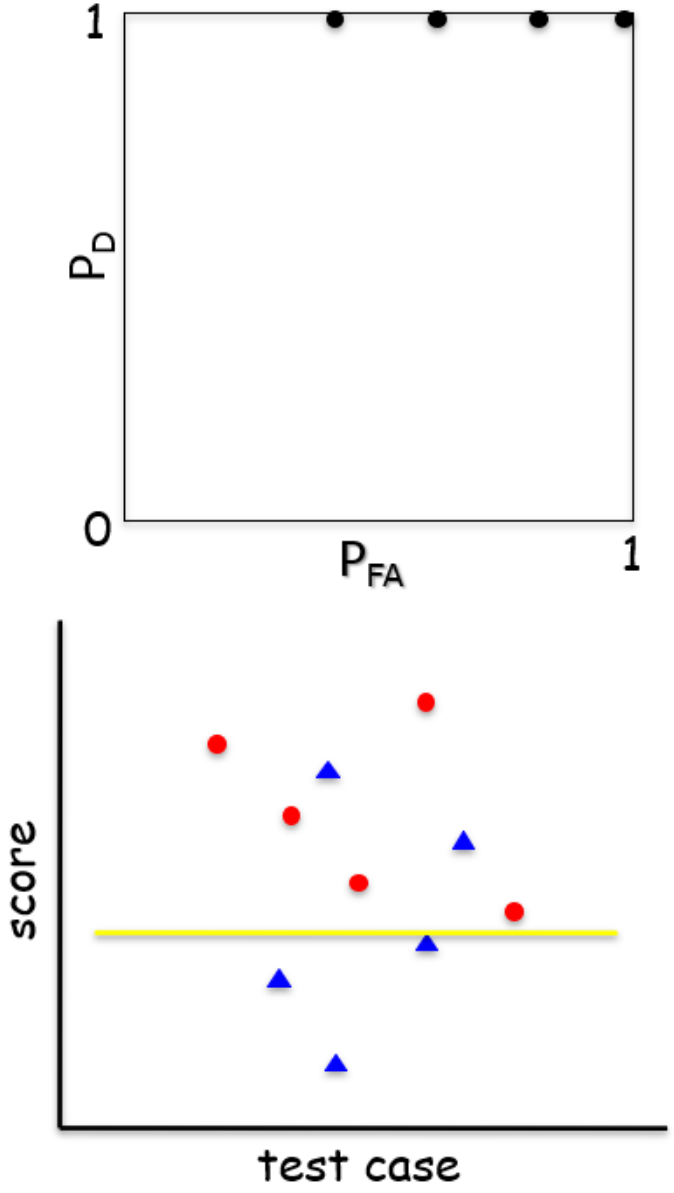
$P_{\mathrm{D}} = 1,\ P_{\mathrm{FA}} = 1$



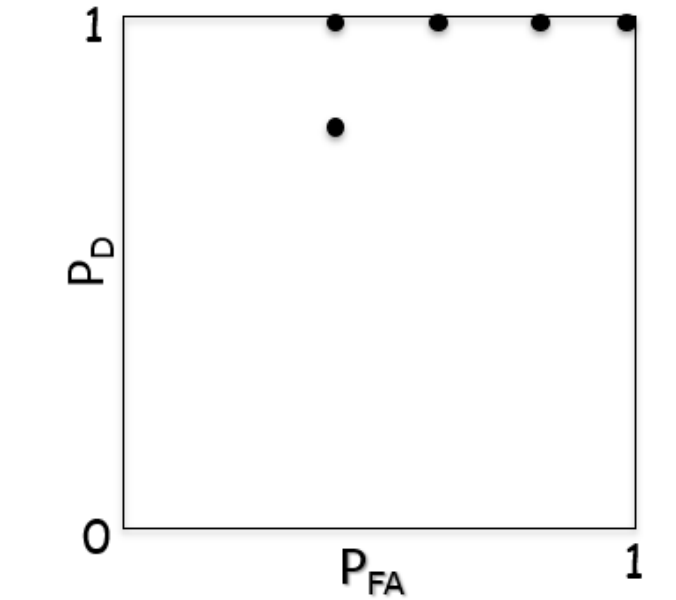
$P_{\mathrm{D}} = 1,\ P_{\mathrm{FA}} = 0.8$

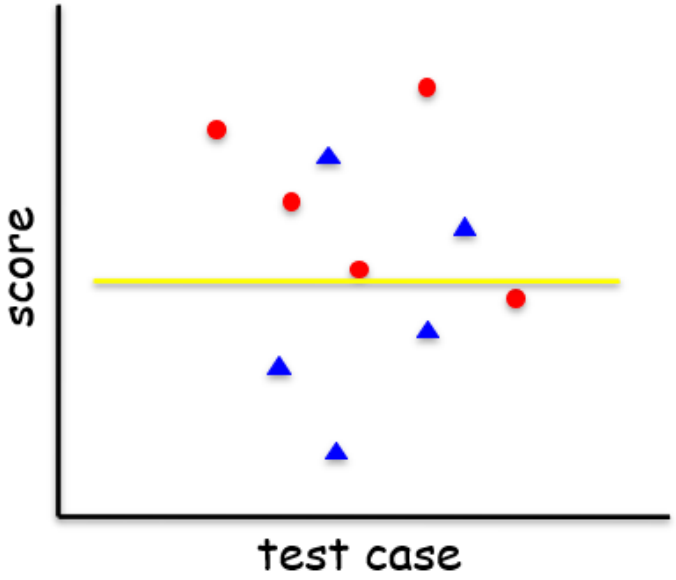


$P_{\mathrm{D}} = 1,\ P_{\mathrm{FA}} = 0.6$

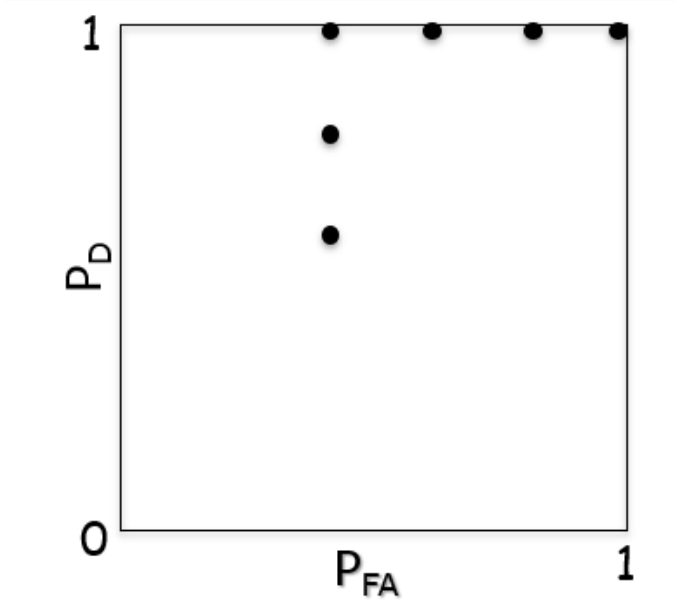


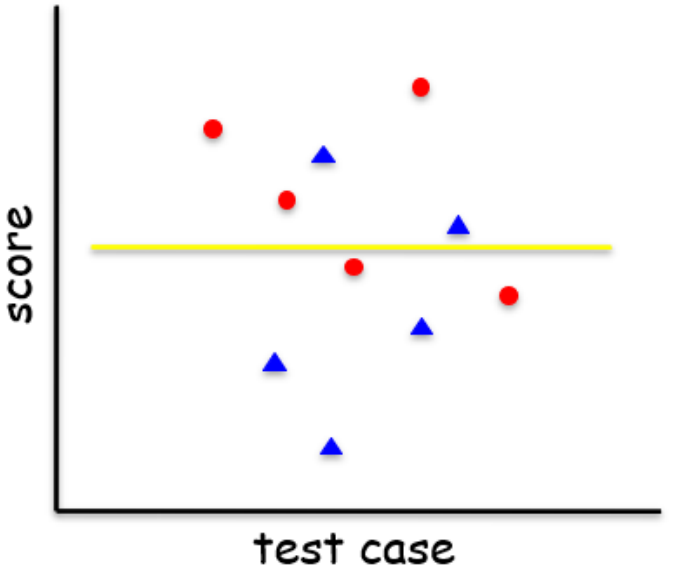
$P_{\mathrm{D}} = 1,\ P_{\mathrm{FA}} = 0.4$



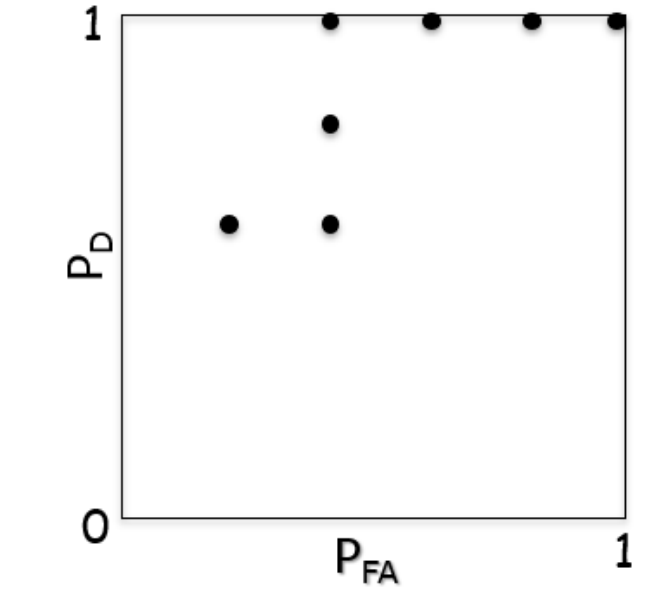


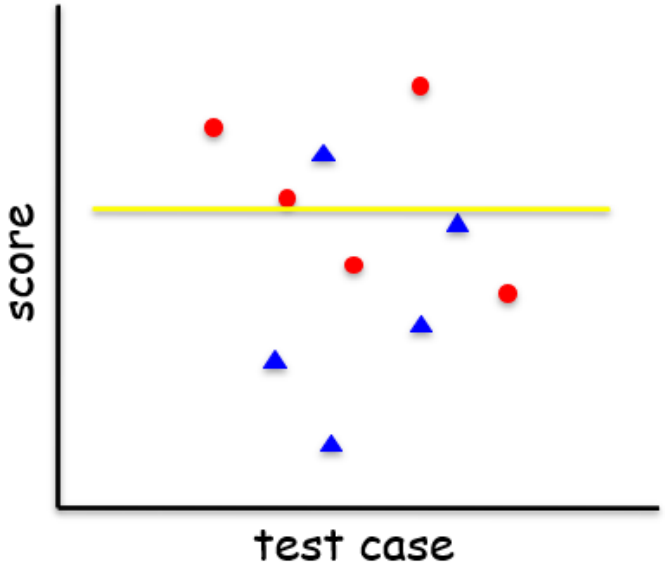
$P_{\mathrm{D}} = 0.8,\ P_{\mathrm{FA}} = 0.4$



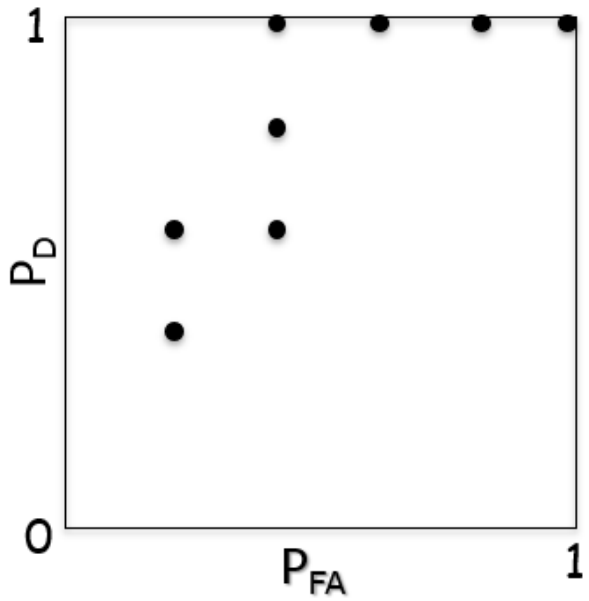


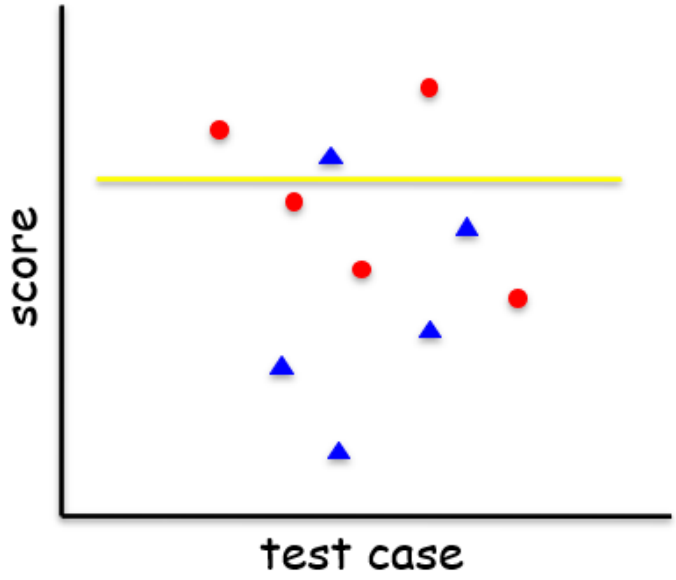
$P_{\mathrm{D}} = 0.6,\ P_{\mathrm{FA}} = 0.4$



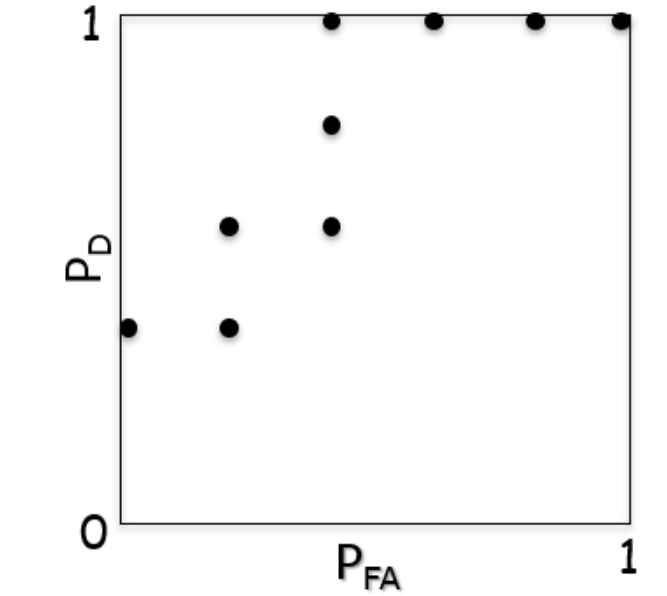


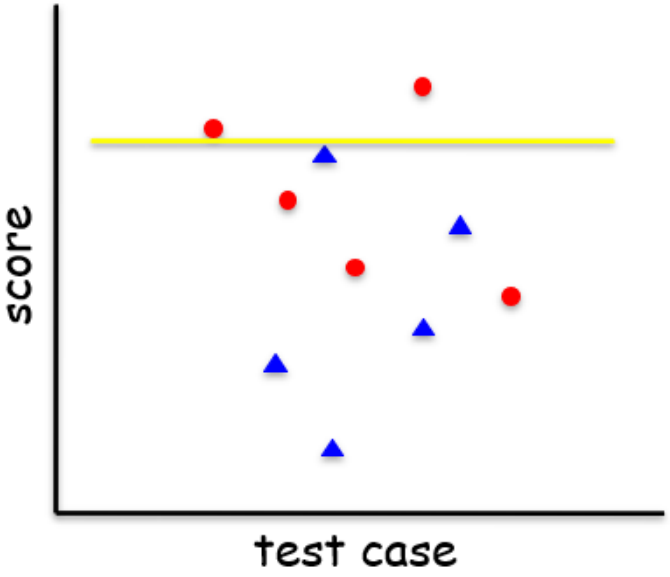
$P_{\mathrm{D}} = 0.6,\ P_{\mathrm{FA}} = 0.2$



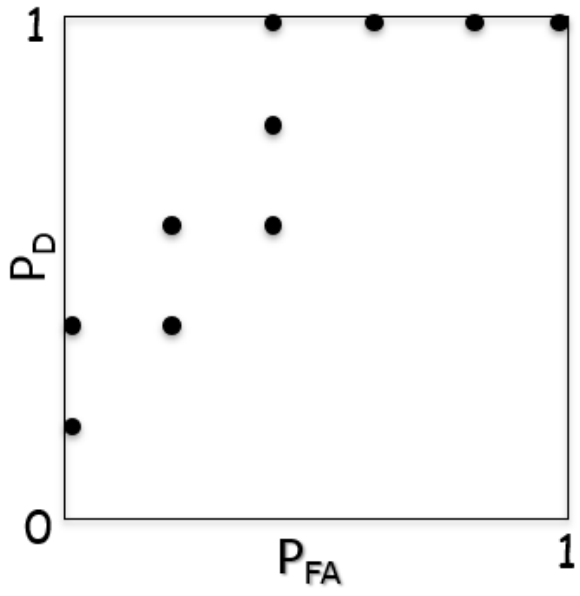


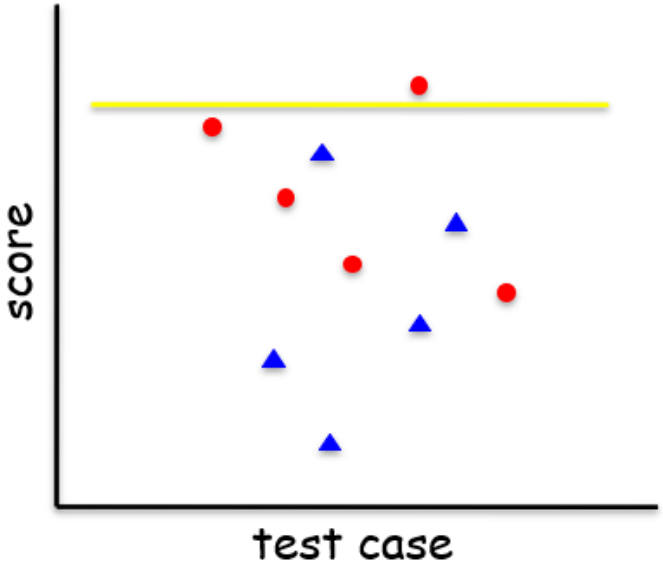
$P_{\mathrm{D}} = 0.4,\ P_{\mathrm{FA}} = 0.2$



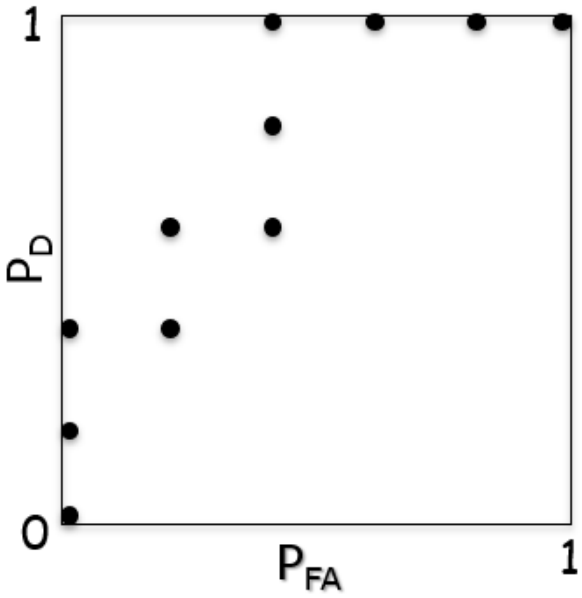


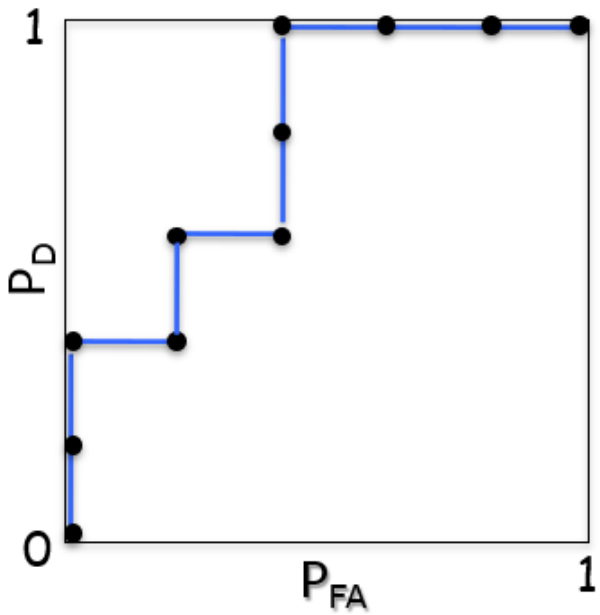
$P_{\mathrm{D}} = 0.4,\ P_{\mathrm{FA}} = 0$



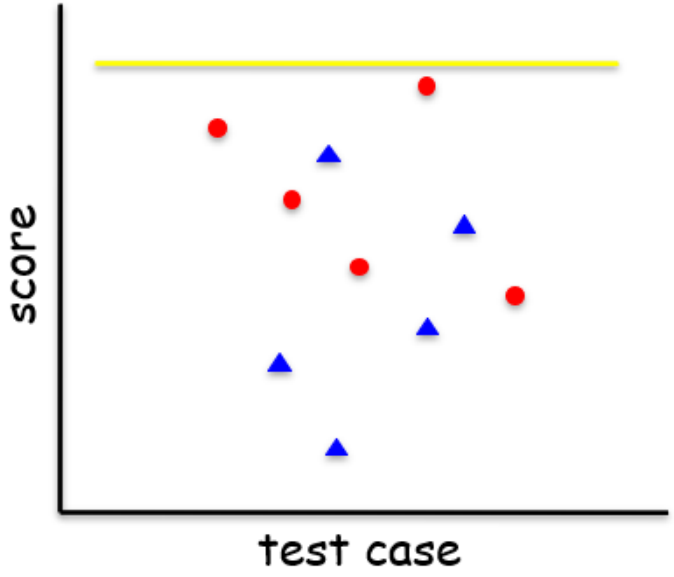


$P_{\mathrm{D}} = 0.2,\ P_{\mathrm{FA}} = 0$





Connecting the dots gives ROC



$P_{\mathrm{D}} = 0, \; P_{\mathrm{FA}} = 0$

Neyman-Pearson Theorem

It is reasonable and straightforward to directly adopt $x[0]$ as the decision statistic in the binary detection problem of (5.2).

In fact, this is the optimal choice according to the Neyman-Pearson theorem.

Given a set of observations $\boldsymbol{x} = [x[0]\ x[1]\ \cdots\ [N-1]]^T$ with joint PDFs under $\mathcal{H}_0$ and $\mathcal{H}_1$, i.e., $p(\boldsymbol{x}; \mathcal{H}_0)$ and $p(\boldsymbol{x}; \mathcal{H}_1)$.

To maximize $P_{\rm D}$ for a given $P_{\rm FA} = \alpha$, $\mathcal{H}_1$ should be decided if

$$L(\boldsymbol{x}) = \frac{p(\boldsymbol{x}; \mathcal{H}_1)}{p(\boldsymbol{x}; \mathcal{H}_0)} > \gamma_{\rm NP} \tag{5.6}$$

where $L(\boldsymbol{x})$ is called the likelihood ratio which measures how probable $\mathcal{H}_1$ is true relative to $\mathcal{H}_0$ is true given $\boldsymbol{x}$.

The threshold γ_{NP} is chosen to satisfy:

$$P_{\text{FA}} = P(L(\boldsymbol{x}) > \gamma_{\text{NP}}; \mathcal{H}_0) = \alpha \tag{5.7}$$

Example 5.4

Show that using $x[0]$ as the decision statistic for the signal detection problem in (5.4) is an optimal choice according to the Neyman-Pearson theorem. Determine γ_{NP} and P_{D} for $P_{\text{FA}} = 10^{-3}$ with $A = 1$ and $w[0] \sim \mathcal{N}(0, 1)$.

Generalizing (5.5) with $A > 0$ and $w[0] \sim \mathcal{N}(0, \sigma^2)$, we have

$$p(\boldsymbol{x}; \mathcal{H}_0) = p(x[0]; \mathcal{H}_0) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}x^2[0]}$$

$$p(\boldsymbol{x}; \mathcal{H}_1) = p(x[0]; \mathcal{H}_1) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x[0]-A)^2}$$

We apply (5.6) to obtain:

$$L(\boldsymbol{x}) = \frac{p(\boldsymbol{x};\mathcal{H}_1)}{p(\boldsymbol{x};\mathcal{H}_0)} = \frac{e^{-\frac{1}{2\sigma^2}(x[0]-A)^2}}{e^{-\frac{1}{2\sigma^2}x^2[0]}} = e^{-\frac{1}{2\sigma^2}\left(-2Ax[0]+A^2\right)} > \gamma_{\mathrm{NP}}$$

Even A and σ^2 are unknown, we can proceed as follows:

$$\begin{aligned}
&-\frac{1}{2\sigma^2}\left(-2Ax[0]+A^2\right) > \ln(\gamma_{\mathrm{NP}}) \\
&\Rightarrow 2Ax[0] - A^2 > 2\sigma^2 \ln(\gamma_{\mathrm{NP}}) \\
&\Rightarrow x[0] > \frac{A}{2} + \frac{\sigma^2}{A}\ln(\gamma_{\mathrm{NP}}) \Rightarrow x[0] > \gamma, \quad \gamma = \frac{A}{2} + \frac{\sigma^2}{A}\ln(\gamma_{\mathrm{NP}})
\end{aligned}$$

Note that using both decision statistics $e^{-\frac{1}{2\sigma^2}\left(-2Ax[0]+A^2\right)}$ and $x[0]$ should yield the same performance for the same value of γ_{NP}.

Consider $A = \sigma^2 = 1$. To compute γ_{NP}, we can obtain γ first:

$$P_{\text{FA}} = \int_{\gamma}^{\infty} p(x; \mathcal{H}_0) dx = \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.001$$
$$\Rightarrow \int_{-\infty}^{\gamma} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.999 \Rightarrow \gamma = 3.0902$$

```
>> norminv(1-0.001,0,1)
ans = 3.0902
```

This also aligns with

```
>> 1-normcdf( 3.0902,0,1)
ans = 0.0010
```

Then we get:

$$P_{\text{D}} = 0.0183$$
$$\gamma_{\text{NP}} = e^{\gamma - 0.5} = 13.3329$$

```
>> 1-normcdf(3.0902,1,1)
ans = 0.0183
```

Example 5.5

Consider the binary detection problem using multiple observations with IID $w[n] \sim \mathcal{N}(0, \sigma^2)$:

$$\begin{aligned}\mathcal{H}_0 &: x[n] = w[n], \\ \mathcal{H}_1 &: x[n] = A + w[n], \quad n = 0, \cdots, N-1\end{aligned}$$

Determine a decision statistic according to the Neyman-Pearson theorem. Assume that $A > 0$ and σ^2 are unknown.

Let $\boldsymbol{x} = [x_0 \ \cdots \ x_{N-1}]^T$. According to (3.38), we have:

$$p(\boldsymbol{x}; \mathcal{H}_0) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} x^2[n]}$$

$$p(\boldsymbol{x}; \mathcal{H}_1) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} (x[n]-A)^2}$$

We apply (5.6) to obtain:

$$L(\boldsymbol{x}) = \frac{p(\boldsymbol{x};\mathcal{H}_1)}{p(\boldsymbol{x};\mathcal{H}_0)} = \frac{e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2}}{e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]}} > \gamma_{\text{NP}}$$

$$\Rightarrow -\frac{1}{2\sigma^2}\left[\sum_{n=0}^{N-1}(x[n]-A)^2 - \sum_{n=0}^{N-1}x^2[n]\right] > \ln(\gamma_{\text{NP}})$$

$$\Rightarrow \frac{A}{\sigma^2}\sum_{n=0}^{N-1}x[n] - \frac{NA^2}{2\sigma^2} > \ln(\gamma_{\text{NP}})$$

$$\Rightarrow \frac{1}{N}\sum_{n=0}^{N-1}x[n] > \frac{A}{2} + \frac{\sigma^2}{NA}\ln(\gamma_{\text{NP}}) = \gamma$$

Therefore, an optimal decision statistic is to use the sample mean.

Can we use $\sum_{n=0}^{N-1}x[n]$ as the optimal decision statistic?

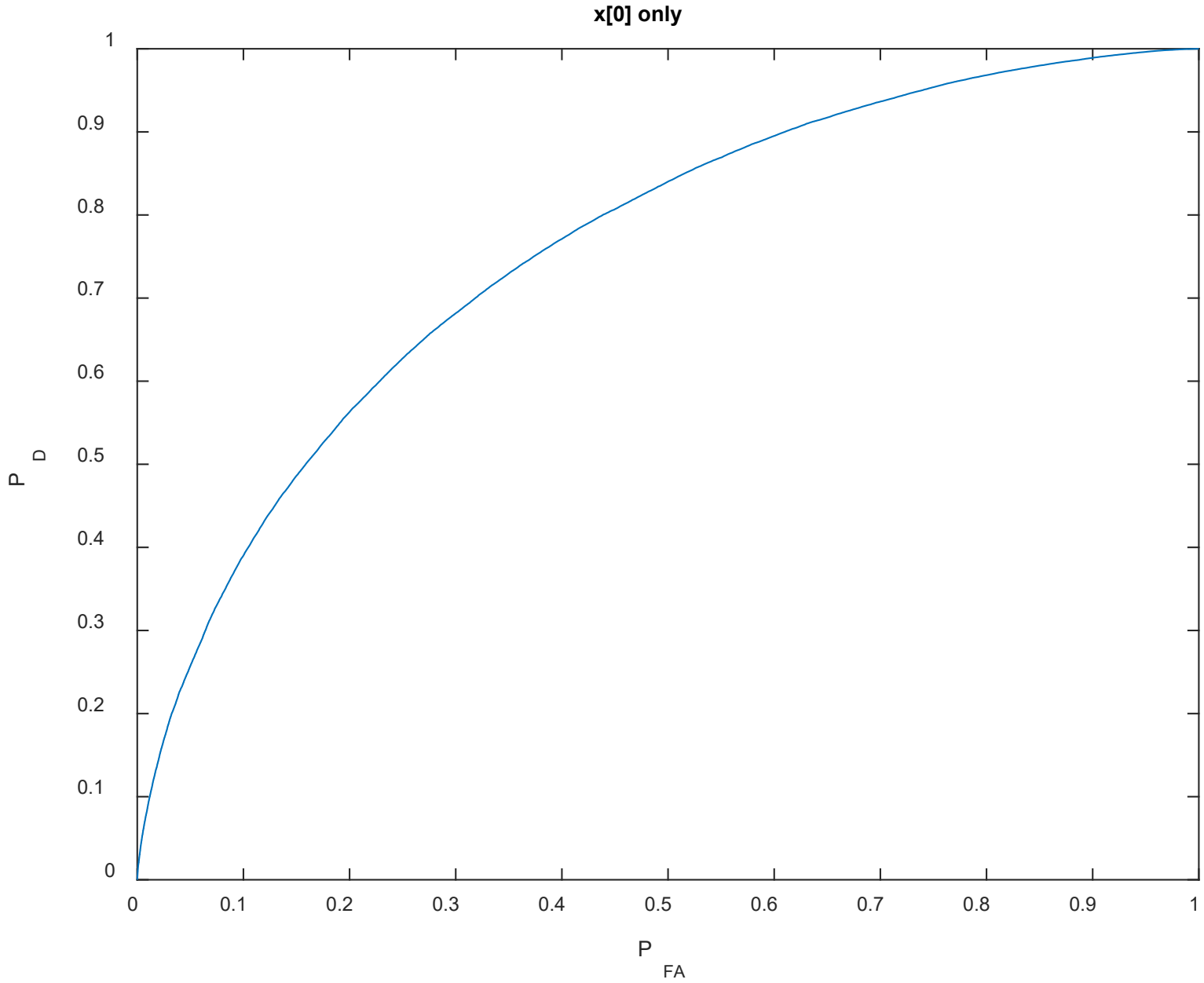
Example 5.6

Consider the following hypotheses:

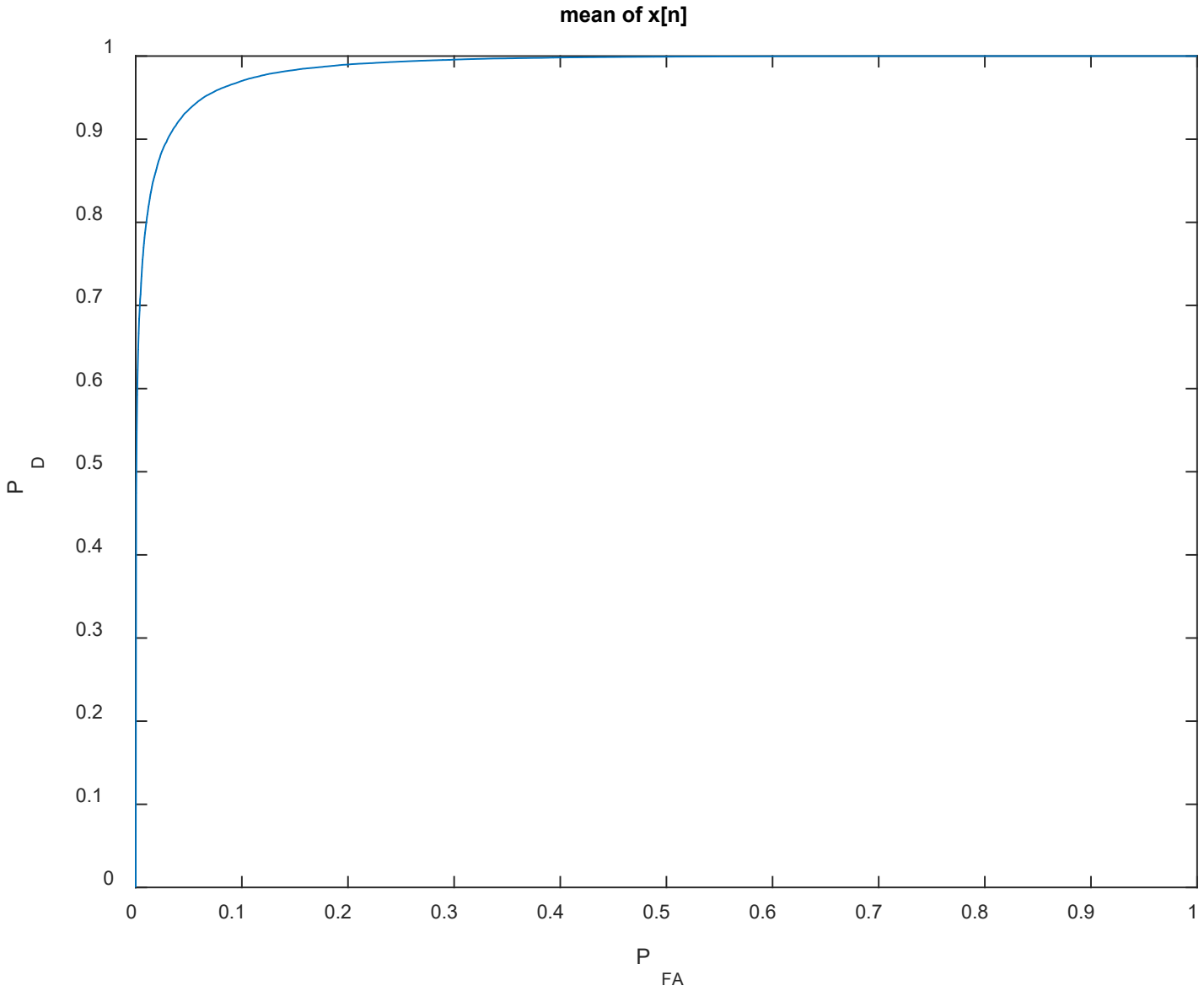
$$\mathcal{H}_0 : x[n] = w[n],$$
$$\mathcal{H}_1 : x[n] = A + w[n], \quad n = 0, \cdots, N-1$$

where $w[n] \sim \mathcal{N}(0,1)$. Two detection statistics, namely, $x[0]$ and $\overline{x} = \sum_{n=0}^{N-1} x[n]/N$ are suggested. Plot the ROCs using MATLAB command `perfcurve` with $A = 1$ and $N = 10$.

```
A = 1;
n = 100000;
h0 = randn(1,n); % noise only
h1 = randn(1,n) + A; % signal plus noise
h = [h0,h1].'; % concatenate H0 and H1
label0 = zeros(1,n);
label1 = ones(1,n);
label = [label0,label1]; % concatenate labels
[X,Y] = perfcurve(label,h,1); %1 means positive case H1
plot(X,Y)
```


x[0] only
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P D
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
P FA

```
A=1;
n = 100000;
h0 = randn(10,n); % noise only
e0 = mean(h0, 1); % calculate mean
h1 = randn(10,n) + A; % signal plus noise
e1 = mean(h1, 1); % calculate the mean
e = [e0,e1].';
label0 = zeros(1,n);
label1 = ones(1,n);
label = [label0, label1];
[X,Y] = perfcurve(label, e, 1);
plot(X,Y)
```



Which detector is better?

For the PDF of $\overline{x} = \sum_{n=0}^{N-1} x[n]/N$, we apply the results in Example 3.11 to obtain $\mathbb{E}\{\overline{x}\} = 0$ for $\mathcal{H}_0$, $\mathbb{E}\{\overline{x}\} = 1$ for $\mathcal{H}_1$, and $\mathrm{var}(\overline{x}) = 1/N$ for both cases. That is:

$$p(\overline{x}; \mathcal{H}_0) = \frac{1}{\sqrt{2\pi/N}} e^{-\frac{N}{2}\overline{x}^2}$$

$$p(\overline{x}; \mathcal{H}_1) = \frac{1}{\sqrt{2\pi/N}} e^{-\frac{N}{2}(\overline{x}-1)^2}$$

As the variance is reduced by N, the spread of these PDFs is smaller, and they have smaller overlap, making two hypotheses easier to differentiate.

Analogously, the estimation performance can be improved by using all available data appropriately, as indicated in Example 3.18.

References:

1. R. D. Yates and D. J. Goodman, *Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineer*, Wiley, 2014
2. S. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*, Pearson, 1998
3. M. Stamp, *Introduction to Machine Learning with Applications in Information Security*, Chapman & Hall/CRC, 2017
   http://www.cs.sjsu.edu/~stamp/ML/powerpoint/