Department of Mechanical
and Industrial Engineering

TPK4450 - Data Driven Prognostics and Predictive Maintenance

Semester work I

Author:
Joachim Nilsen Grimstad

September 18, 2020

Foreword

All the code is written in Python and can found well commented in my gitHub repository. All the generated data uses numpy.random.distribution(parameters) to generate data from various distributions, since this is a pseudo random algorithm, I have seeded all the code with the $seed = 4450$ so the results can be reproduced. Any PDF plotted from generated data, is plotted using seaborn.kdeplot() which uses kernel density estimation to plot the PDF of a dataset. Note that the probability density (y-value) in a PDF can be larger than one, since the area under the curve (probability) needs to be 1.

$$CDF = \int_{-\infty}^{\infty} PDF = 1$$

Reflections

Using the sample mean as a decision metric makes a lot of sense for a few reasons.

First lets assume the measurements X_i are independent and of the same <u>unknown</u> distribution. From the central limit theorem we know that the sample mean:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{1}$$

is normally distributed:

$$\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}}) \tag{2}$$

if n is "sufficiently large": $n > 30$ (Wikan and Kristensen 2018, p. 173) or $n \geq 20$ (Løvås 2013, p. 199).

To illustrate this I generated some data sets of sample means $\overline{X}$ with different underlying distributions and sample sizes. When $n = 1$ the distribution of $\overline{X}$ is equal to the underlying distribution of X, however as n increases, the distribution of $\overline{X}$ trends towards a normal distribution as seen in Figure 1 and Figure 2. It is therefore possible to set alarm bounds for any distribution if we use the sample mean as the decision metric.

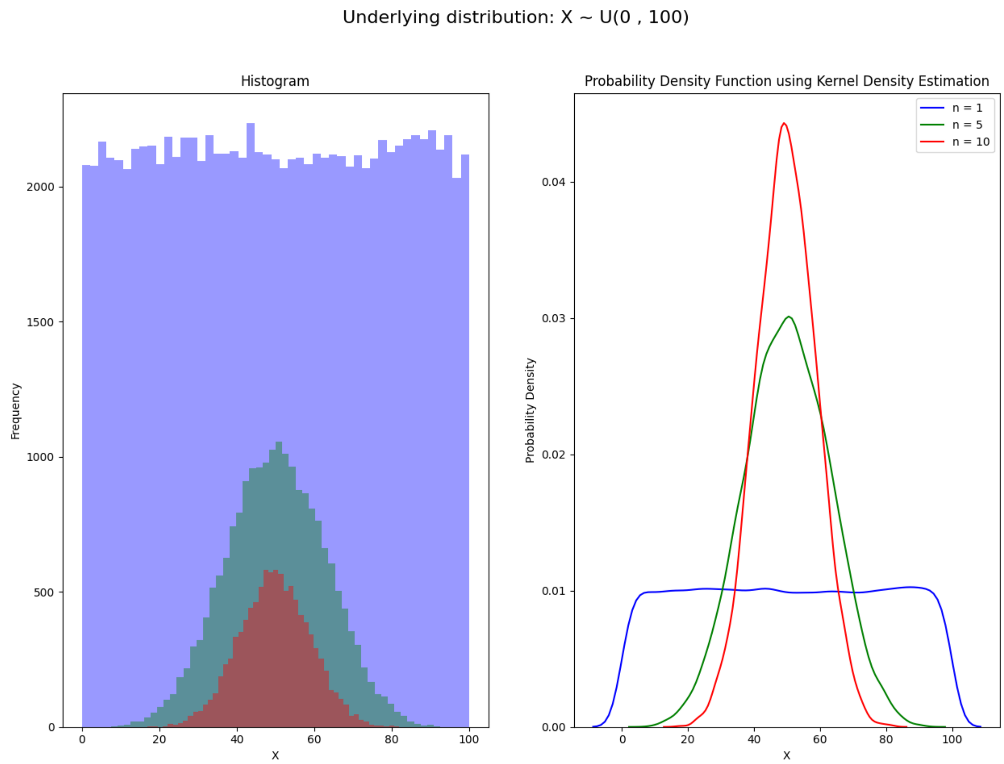

Figure 1: Histogram and PDF of $\overline{X}$ for $n \in [1, 5, 10]$

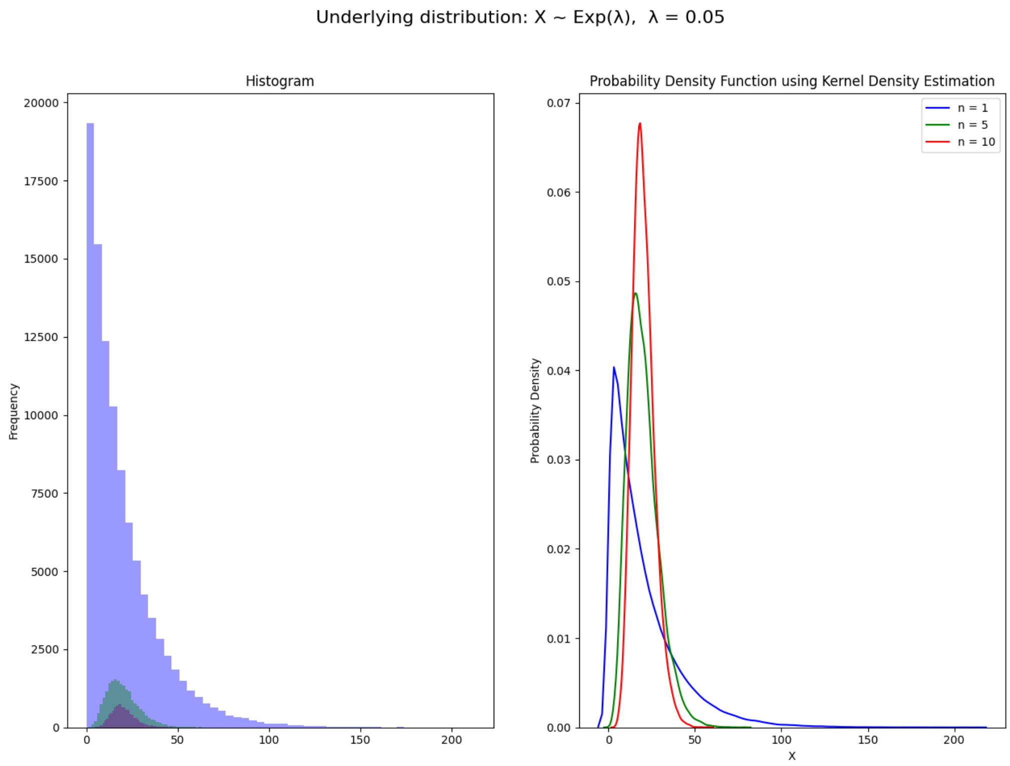

Figure 2: Histogram and PDF of $\overline{X}$ for $n \in [1, 5, 10]$

In this case however, we know that the underlying distribution is $X \sim \mathcal{N}(\mu, \sigma)$. Thus the sample mean $\overline{X}$ is normally distributed, regardless of sample size n, but with increasing sample size the standard deviation is reduced as shown in Figure 3. This can cause less false alarms, but may also require more faulty measurements in a sample to identify as faulty, potentially increasing the detection time and non-detection.

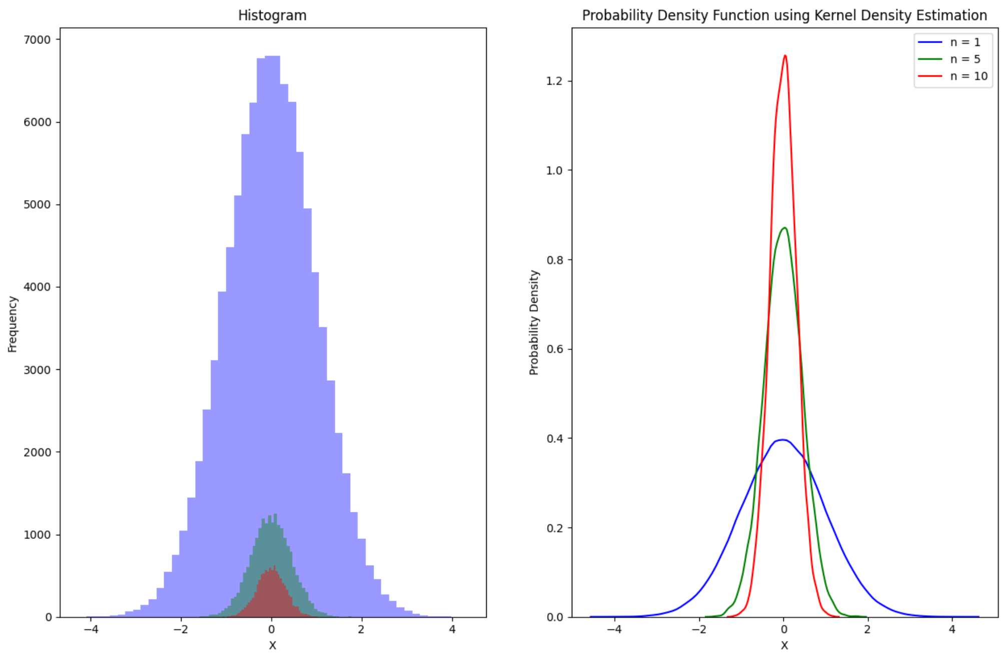

Figure 3: Histogram and PDF of $\overline{X}$ for $n \in [1, 5, 10]$

Problem a)

1)

Simple statistics notation:

$$\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

$$\overline{X} \sim \mathcal{N}(0, \frac{2}{\sqrt{10}})$$

$$\alpha = 0.05 \implies z_\alpha = 1.645$$

$$X_{max} = \mu + (z_\alpha \times \frac{\sigma}{\sqrt{n}}) = \frac{3.29}{\sqrt{10}}$$

$$X_{max} = \underline{\underline{1.04038935}}$$

2)

Plotted using *scipy.stats.norm.pdf*$(domain, \mu, \frac{\sigma}{\sqrt{n}})$ where the domain means $\overline{X} \in D$ where: $D = \{-3, -2.9994, -2.9988, ..., 3\}$, $|D| = 10000$. This draws a PDF of the decision metric between -3 and 3 with 10000 points, making for a nice looking PDF rather than using KDE on estimated data.

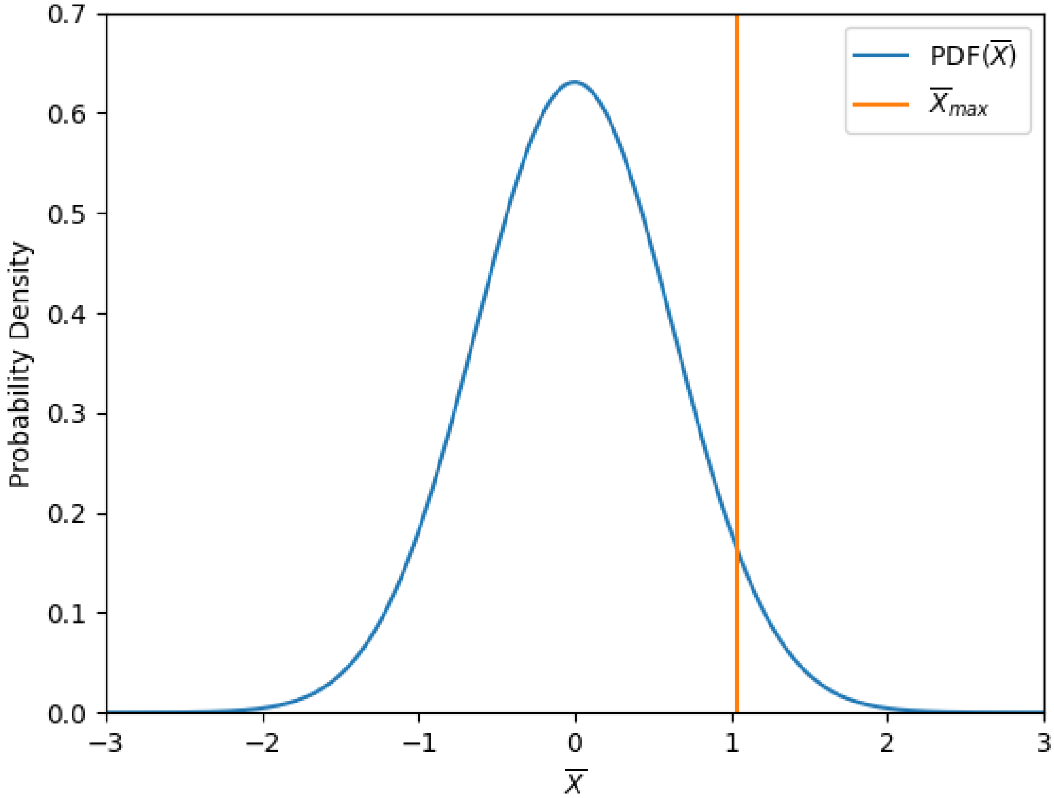

Figure 4: $PDF(\overline{X})$ with $\overline{X}_{max}$

3)

Empirically, with only 100 samples leaves room for errors due to uncertainty, increasing i will reduce the error.

$$\overline{X}_i \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$false\ alarm(\overline{X}_i) = \left\{ \begin{array}{ll} 1, & \text{for } \overline{X}_i \geq \overline{X}_{max} \\ 0, & \text{for } \overline{X}_i < \overline{X}_{max} \end{array} \right\}$$

$$false\ alarms = \sum_{i=1}^{100} false\ alarm(\overline{X}_i) = \underline{\underline{9}}$$

$$\implies \quad healthy = \underline{\underline{91}}$$

Analytically

$$\overline{X}_i \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$false\ alarm(\overline{X}_i) = \left\{ \begin{array}{ll} 1, & \text{for } \overline{X}_i \geq \overline{X}_{max} \\ 0, & \text{for } \overline{X}_i < \overline{X}_{max} \end{array} \right\}$$

$$percentage\ false\ alarms = \lim_{n \to +\infty} \left(\frac{\sum_{i=1}^{n} false\ alarm(\overline{X}_i)}{n} \times 100\% \right) = \underline{5\%}$$

$$\implies \quad percentage\ healthy = \underline{95\%}$$

Problem b)

4)

The figure becomes a little "wonky" with KDE of only 100 samples.

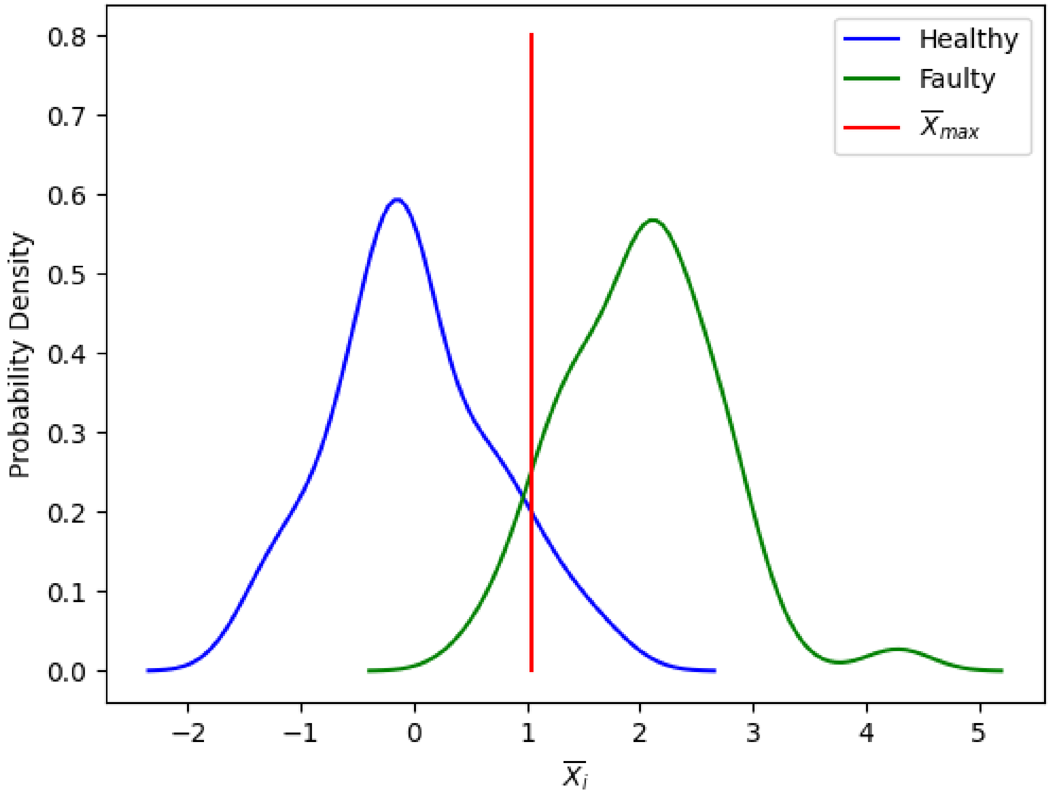

Figure 5: Healthy and faulty PDF with $\overline{X}_{max}$

Empirically, since I seeded the generator, the false alarms are the same as 3)

$$\overline{X}_i \sim \mathcal{N}\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$$

$$false\ alarm(\overline{X}_i) = \left\{ \begin{array}{ll} 1, & \text{for } \overline{X}_i \geq \overline{X}_{max} \\ 0, & \text{for } \overline{X}_i < \overline{X}_{max} \end{array} \right\}$$

$$false\ alarms = \sum_{i=1}^{100} false\ alarm(\overline{X}_i) = \underline{\underline{9}}$$

$$\implies \quad healthy = \underline{\underline{91}}$$

$$\overline{X}_i \sim \mathcal{N}\left(\mu_1, \frac{\sigma}{\sqrt{n}}\right)$$

$$non - detection(\overline{X}_i) = \left\{ \begin{array}{ll} 1, & \text{for } \overline{X}_i < \overline{X}_{max} \\ 0, & \text{for } \overline{X}_i \geq \overline{X}_{max} \end{array} \right\}$$

$$non - detections \ = \sum_{i=1}^{100} non - detection(\overline{X}_i) = \underline{\underline{7}}$$

$$\implies \ detections = \underline{\underline{93}}$$

As you can see by the plot Figure 5 the closer to the $\overline{X}_{max}$ you get the more problematic it is to classify. The trouble is approximately in the interval $\{\overline{X}_i \in \mathbb{R} \mid -0.4 \leq \overline{X}_{max} \leq 2.7\}$. In this interval there is a chance for non-detection and false alarms.

5)

Hypothesis:

$$\begin{cases} H_0 \ : \ \overline{X} \sim p_0(\overline{x}) \\ \\ H_1 \ : \ \overline{X} \sim p_1(\overline{x}) \end{cases}$$

Likelihood ratio:

$$\Lambda(x) = \frac{p_1(\overline{x})}{p_0(\overline{x})}$$

Decision making:

$$\delta(x) = \begin{cases} D_0 \ if \ \Lambda(x) < \lambda_{NP} \\ \\ D_1 \ if \ \Lambda(x) \geq \lambda_{NP} \end{cases}$$

Finding λ_{NP}:

$$\alpha = Pr(\Lambda(x) > \lambda_{NP} \mid H_0)$$

$$\alpha = Pr\left(\frac{\frac{1}{\frac{\sigma}{\sqrt{n}}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\frac{\sigma}{\sqrt{n}}}\right)^2}}{\frac{1}{\frac{\sigma}{\sqrt{n}}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right)^2}} > \lambda_{NP} \mid H_0 \right)$$

$$\alpha = Pr\left(e^{\frac{1}{2}\left(\frac{(x-\mu_0)^2-(x-\mu_1)^2)}{\left(\frac{\sigma}{\sqrt{n}}\right)^2}\right)} > \lambda_{NP} \mid H_0 \right)$$

$$since \ \mu_0 = 0 \ , \ \mu_1 = 2 \ , \ \sigma = 2 \ and \ n = 10$$

$$\implies \alpha = Pr\left(e^{5(x-1)} > \lambda_{NP} \mid H_0\right)$$

$$\alpha = Pr\left(5(x-1) > ln(\lambda_{NP}) \mid H_0\right)$$

$$\alpha = Pr\left(x > \frac{ln(\lambda_{NP})}{5} + 1 \mid H_0\right)$$

$$\alpha = \int_{\frac{ln(\lambda_{NP})}{5}+1}^{+\infty} p_0(x)\, dx$$

$$\alpha = F_0(\infty) - F_0\left(\frac{ln(\lambda_{NP})}{5} + 1\right)$$

$$\alpha = 1 - F_0\left(\frac{ln(\lambda_{NP})}{5} + 1\right)$$

$$F_0\left(\frac{ln(\lambda_{NP})}{5} + 1\right) = 1 - \alpha$$

$$ln(\lambda_{NP}) = 5\left(F_0^{-1}(1-\alpha) - 1\right)$$

$$from\ task\ 1) : (1-\alpha) = \overline{X}_{max}$$

$$\lambda_{NP} = e^{5(\overline{X}_{max}-1)} = 1.22378284 \approx \underline{\underline{1.224}}$$

Finding β:

$$\beta = \int_{-\infty}^{\frac{ln(\lambda_{NP})}{5}+1} p_1(x)\, dx$$

$$\beta = \int_{-\infty}^{\frac{ln(\lambda_{NP})}{5}+1} \left(\frac{1}{\frac{\sigma}{\sqrt{n}}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\frac{\sigma}{\sqrt{n}}}\right)^2}\right) dx$$

solved it numerically, code is in the gitHub repository:

$$\beta \approx 0.0645982979827704 \approx \underline{\underline{0.065}}$$

6)

Plotted the evolution of the probability of false alarms and non-detections as a function of λ_{NP}. The plots are generated using: $\lambda_{NP} \in D$ where $D = \{0.004, 0.008, 0.012, ..., 4\}$, $|D| = 1000$. and the formulas:

$$Pr(false\ alarm) = \alpha = 1 - \int_{-\infty}^{\frac{ln(\lambda_{NP})}{5}+1} p_0(x)\ dx$$

$$Pr(non-detection) = \beta = \int_{-\infty}^{\frac{ln(\lambda_{NP})}{5}+1} p_1(x)\ dx$$

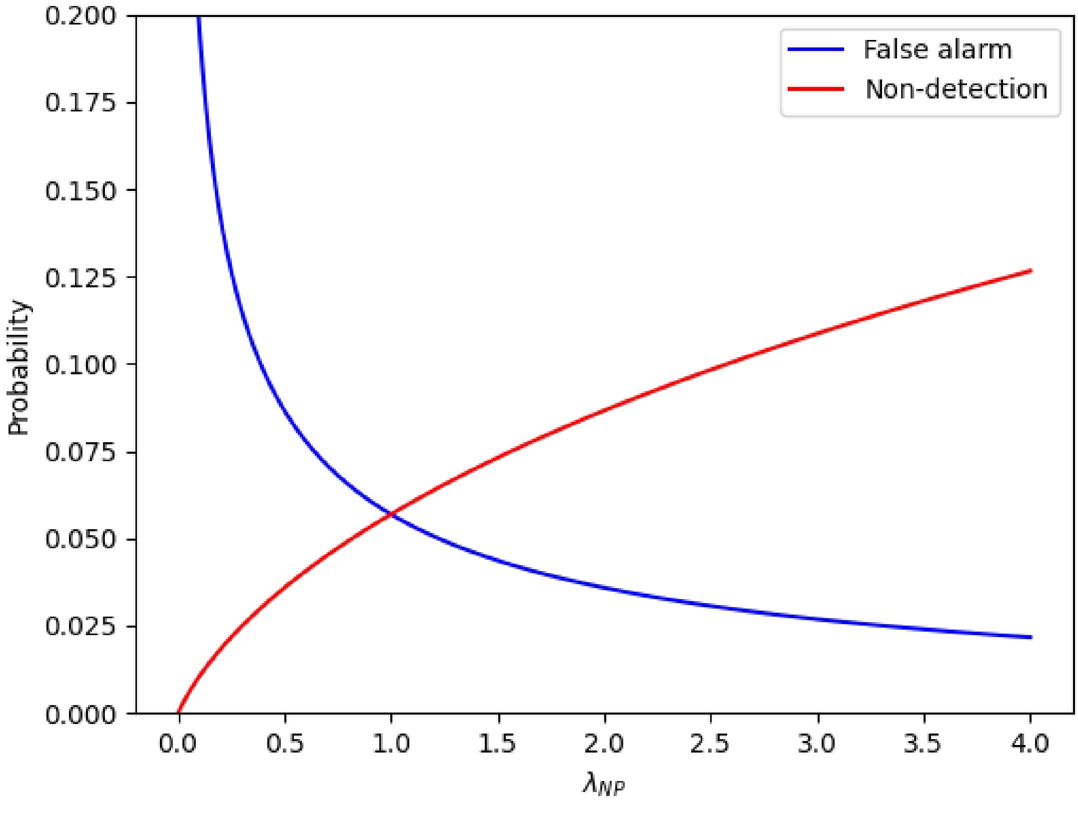

Figure 6: $Probability/\lambda_{NP}$ study

7)

Plotted as described in the problem, with formulas:

$$Pr(false\ alarm) = \alpha = 1 - \int_{-\infty}^{\frac{ln(\lambda_{NP})}{5}+1} p_0(x)\ dx$$

$$Pr(non-detection) = \beta = \int_{-\infty}^{\frac{ln(\lambda_{NP})}{5}+1} p_1(x)\ dx$$

$$Pr(detection) = 1 - \beta$$

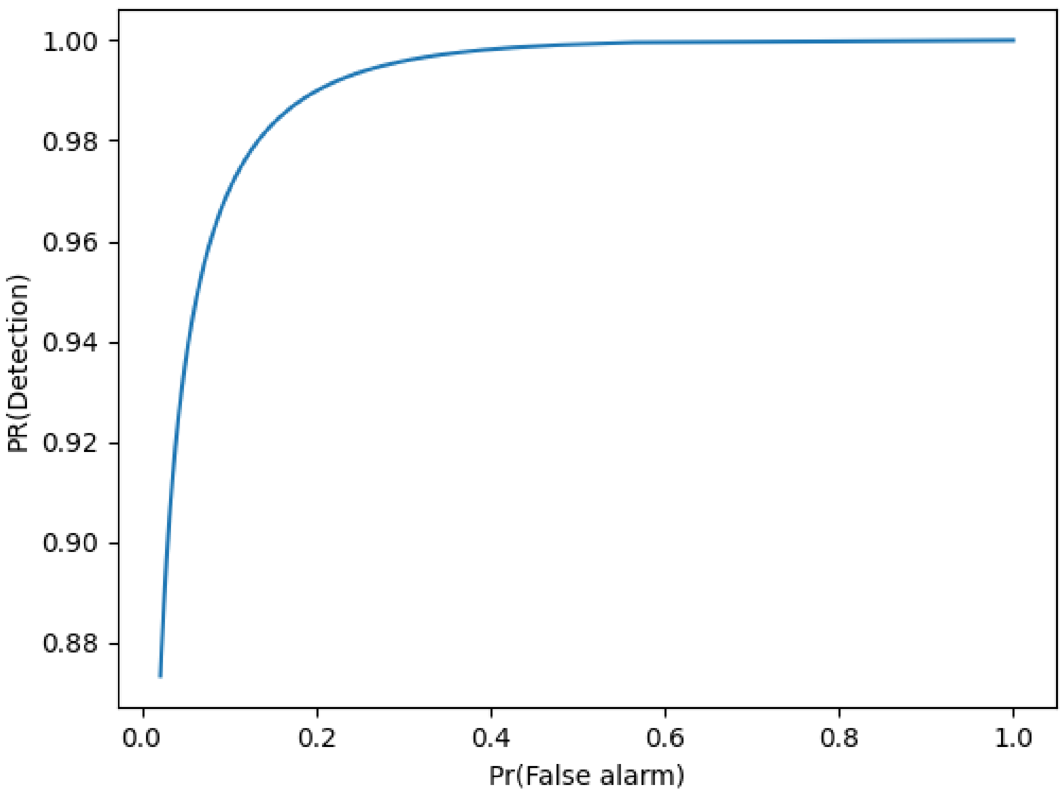

Figure 7: Receiver Operating Characteristics (ROC)

Problem C)

8)

Scatterplot of the classifications

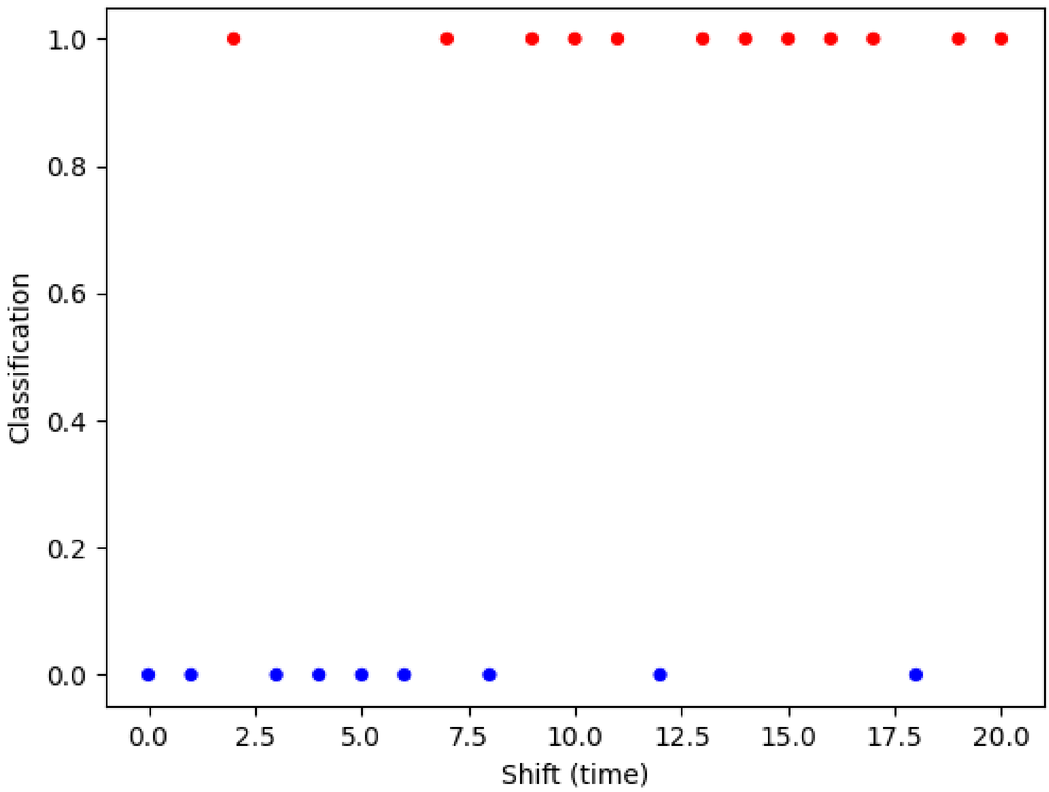

Figure 8: Scatter plot

References

Løvås Gunnar, G. (2013). *Statistikk for Universiteter og Høgskoler*. 3nd. Universitetsforlaget. ISBN: 978-82-15-01807-2.

Wikan, A. and Ø. Kristensen (2018). *Sannsynlighetsregning Og Statistikk*. 2nd. Fagbokforlaget. ISBN: 978-82-450-1938-4.