DAV LAB 3 Report

Aditya Joglekar **201401086**

Rajdeep Pinge **201401103**

Load data_lab3.mat. ‘ammonia_concentration’ provides the ammonia concentration values (in mg/L) from a water treatment plant during a certain period.

1.Suppose you. are asked to assume that this data is normally distributed. How will you confirm/reject this assumption using boxplots, histograms and estimated data statistics (mean, standard deviation etc.). You should explicitly list the features/aspects that you considered in order to arrive at your answer

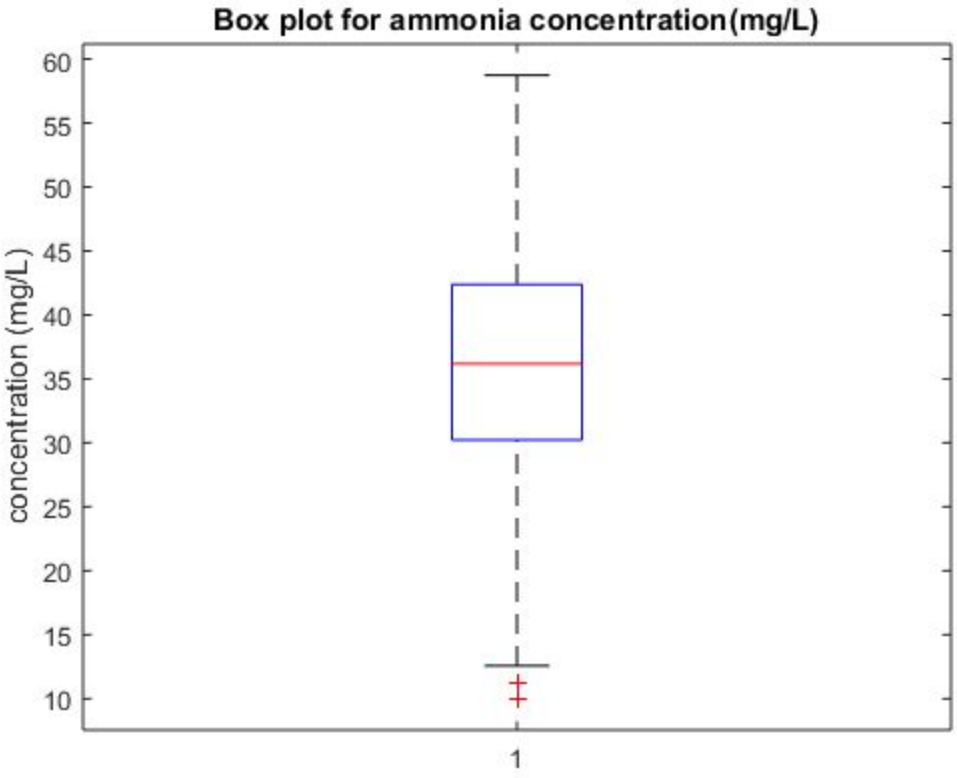
We use the characteristics of the normal distribution to check if the normal distribution assumption is valid or not.

DAV LAB 3 Report

Aditya Joglekar **201401086**

Rajdeep Pinge **201401103**

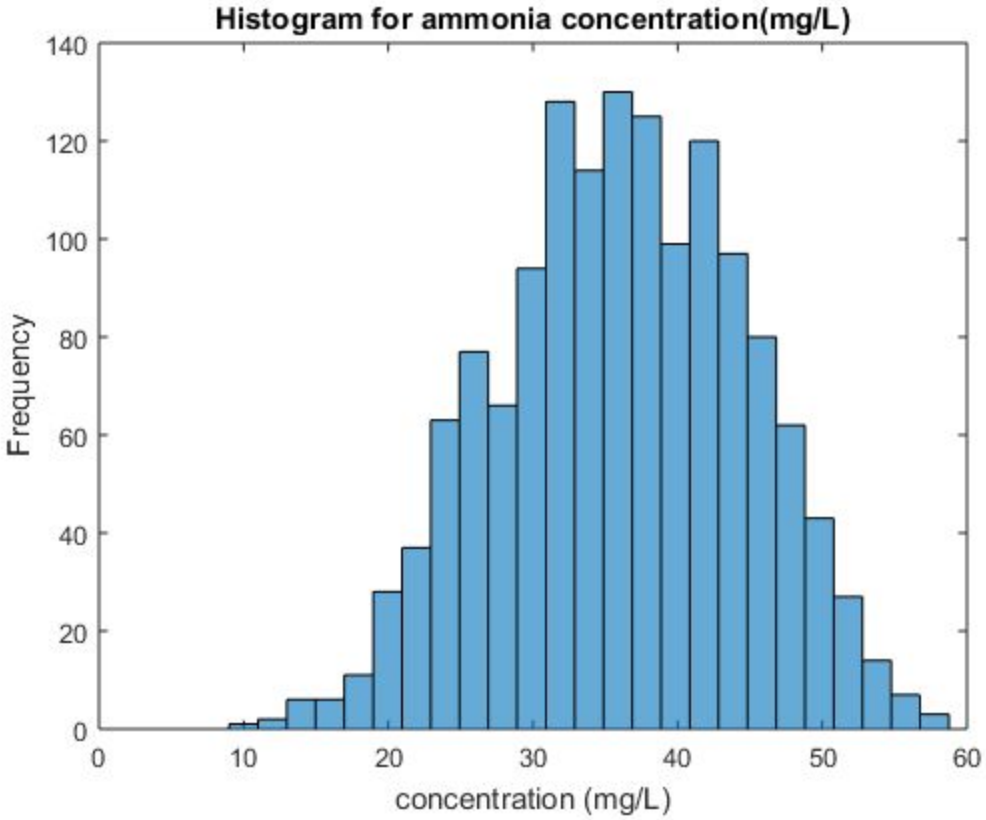
The box plot and the histogram for the given ammonia concentration data are



DAV LAB 3 Report

Aditya Joglekar **201401086**

Rajdeep Pinge **201401103**



From the box plot we see that:

1) The median roughly lies in the middle of the data range.
2) The median lies in the middle of the inter quartile range. This indicates that the data is not skewed towards one direction.
3) The interquartile range is concentrated in a narrow band roughly in the middle of the data range. This means that about 50% of the data is concentrated in a very narrow region in the middle.

The box plot seems to indicate that the data is symmetric and that the majority of the data is concentrated in the middle. It does not provide us any information about the mean or the mode.

DAV LAB 3 Report

Aditya Joglekar **201401086**

Rajdeep Pinge **201401103**

From the histogram we see that:

1) At a glance, one can observe that the distribution “looks” like a gaussian distribution.
2) The median, mean and the mode are roughly the same
3) The further we are from the centre the amount of data goes on decreasing.
4) The distribution is symmetric.

The mean, median and mode values obtained are as follows:

Mean = 36.0950

Median = 36.1800

Mode = 32.6900

From the above, we can say that the data distribution behaves more or less like a gaussian distribution. However, this is still an approximation. For instance the median, mean and mode are not exactly the same as that in the ideal gaussian distribution.

2.Notwithstanding your answer in the previous part, assuming that the data is normally distributed, estimate the probability that ammonia concentration is greater than 40 mg/L in two cases: (a) using only the data provided (i.e. not computing any statistics from given data), (b) by using estimated statistics from the data. Does the answer in the two cases agree, yes or no? Provide reason(s) for your choice.

DAV LAB 3 Report

Aditya Joglekar **201401086**

Rajdeep Pinge **201401103**

a) The simplest way to approximate the probability that the concentration is greater than 40 is to , count the samples greater than 40 and divide it by the total number of samples.

Probability estimated by this method = 0.3444

b) Now we use the normal distribution approximation. We assume that data distribution is normal. We further standardize the distribution by subtracting the calculated mean and dividing by the standard deviation.

The mean and standard deviations for the given data are : 36.0950 and 8.5189

Thus 40 is rescaled to 0.4584

We need to find the P (x>40). So we need to find the P(x>0.4584).

Thus,

P(x>0.4584)= 1- (P(x<=0.4584))

P(x<=0.4584) can be obtained using z-score table of the standard normal distribution. It is basically equal to cdf(x=0.4584) of a standard normal distribution.

We have used the MATLAB normcdf to directly obtain the above cdf value.

The answer which we obtain is 0.3233. It is close to the answer obtained directly from the data. The answer is not accurate as we have made an assumption that the data is normal. But the data is not exactly normal. Hence, we get a close enough answer. But the answers don't match exactly.

DAV LAB 3 Report

Aditya Joglekar **201401086**

Rajdeep Pinge **201401103**

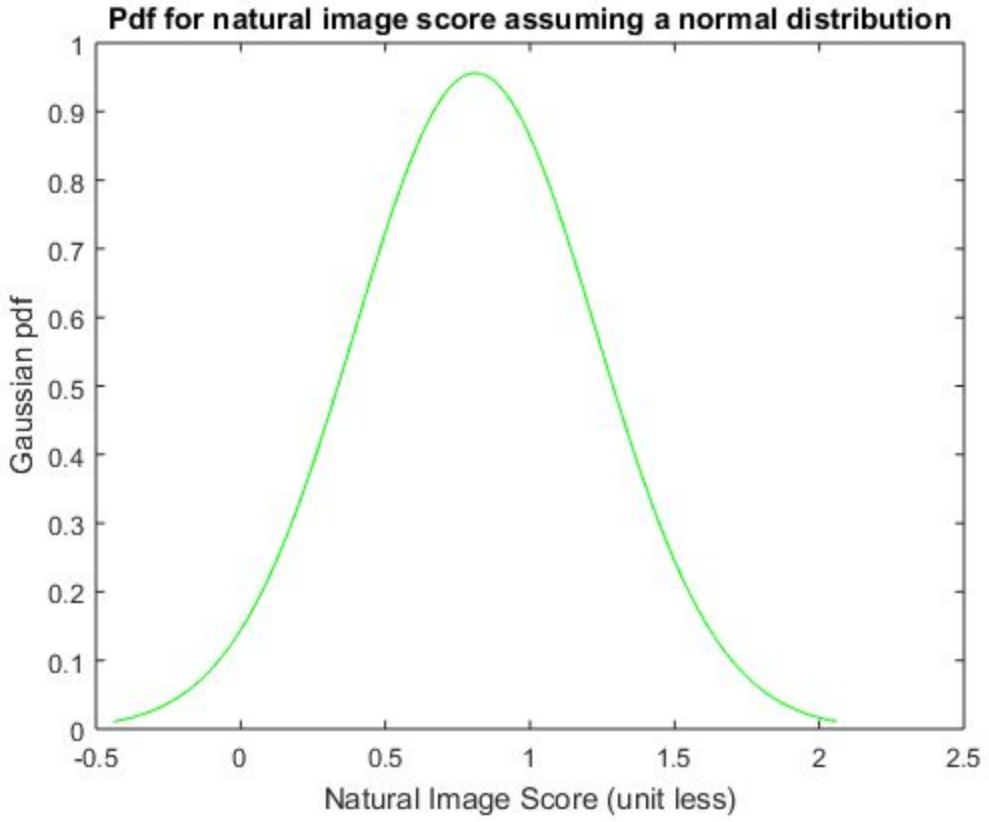
Q2.1 : Assuming that the data in 'score_natural_model' is normally distributed, obtain the corresponding normal distribution curve. Note that the normal pdf is given by :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

DAV LAB 3 Report

Aditya Joglekar **201401086**

Rajdeep Pinge **201401103**



Since the resulting pdf is derived from a set of natural images, it can be used to evaluate the likelihood of how natural a new test image is.

(a) as the pdf value at a point does not give probability, how can we make relative comparisons from the pdf? (hint: think in terms of similarity between pdf and histogram)

⇒ A pdf although does not give probability directly, the area under the pdf does provide with the probability. The area under the pdf till a point on x-axis depict the probability that the random variable will take value less than the value on x-axis. Mathematically, $P(X < x)$ (probability that random variable X will take value less than x) is obtained from the area under curve.

A histogram is often used as discrete representation of a pdf.
So given any two points on a pdf, we can compare them by finding the area under the curve, that how likely a point is to have a value less than the given x-value.

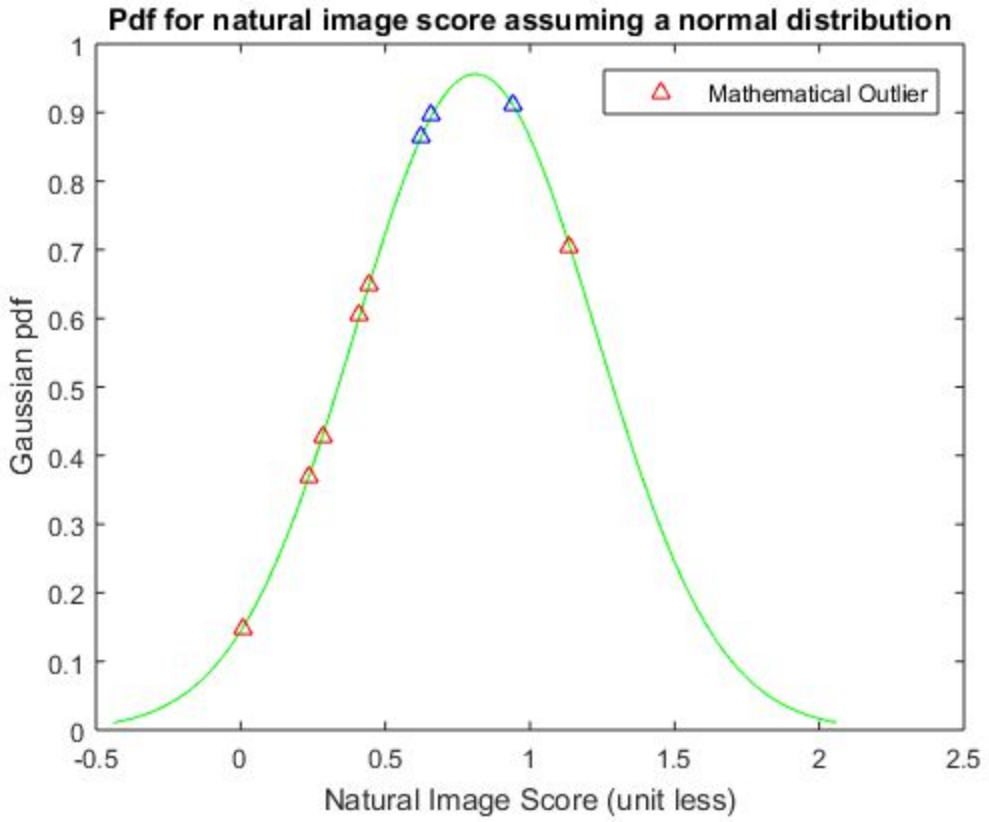
DAV LAB 3 Report

Aditya Joglekar **201401086**

Rajdeep Pinge **201401103**

We can compare two pdfs by looking at the shape of the curves.

(b) Based on your answer to the previous part, evaluate the resulting pdf at 9 values provided in 'score_test'. Note that these 9 values correspond to the provided images img1, img2,…img9 respectively. Does the output from the modeled pdf agree with subjective (human) opinion of naturalness of the given 9 test images? Justify your answer.



Here, the RED dots are the outliers, and the BLUE points are the natural test images.
The boundary considered by us after cross-checking visually the images given to us, is :
⇒ Any point lying outside ±0.5*STDEV from the mean (peak) is an outlier.

OUTLIERS : Fig 1, Fig 2, Fig 3, Fig 6, Fig 8, Fig 9

Now assume that the given data follows a Rayleigh distribution. Note that the Rayleigh pdf is given by

DAV LAB 3 Report

Aditya Joglekar **201401086**

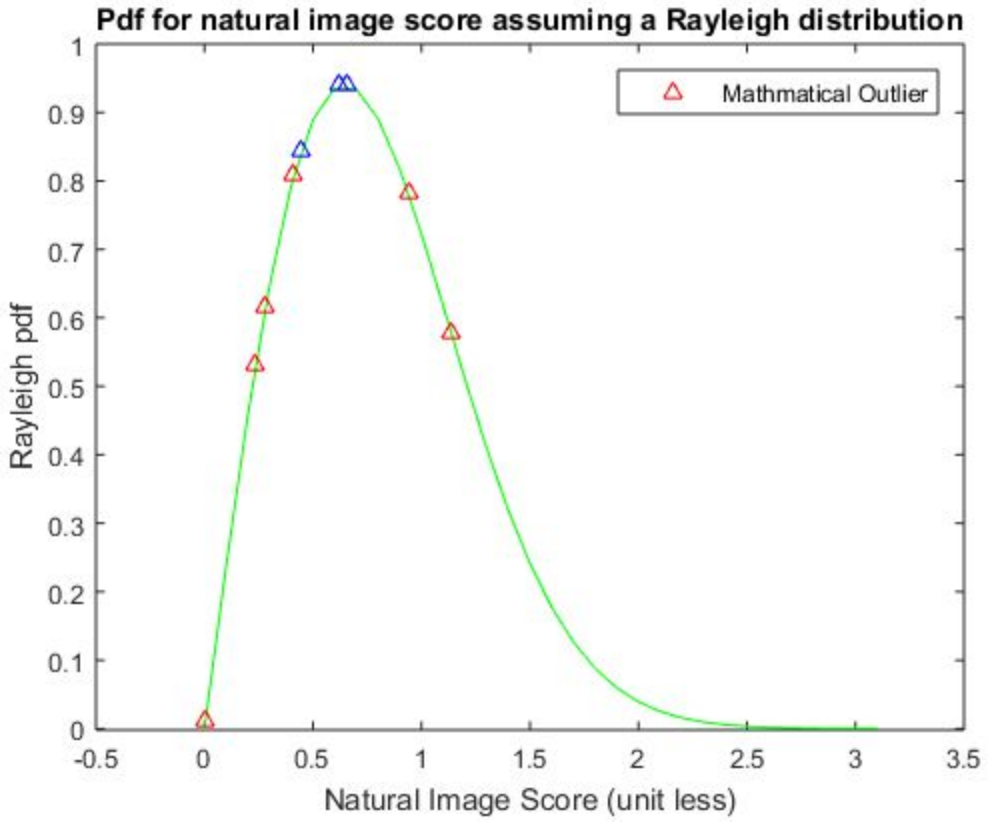
Rajdeep Pinge **201401103**

:

$$f(x) = \frac{x}{\sigma^2} e^{\frac{-x^2}{2\sigma^2}}, x \geq 0$$

In this case, represents the scale parameter which is estimated from the data. For the given problem .

Similar to the previous part, evaluate the resulting pdf at the 9 test values provided in 'score_test'. Does the output from the modeled Rayleigh pdf agree with subjective opinion of naturalness of the given 9 test images? Compare your answers with the previous case where the data was assumed to be normally distributed.



Observing the results, we observe that rayleigh slightly better classifies the images as the outliers as it includes the Fig 4. in outliers which is an image with very high contrast. Fig 4 was not classified as an outlier by the normal distribution method.

DAV LAB 3 Report

Aditya Joglekar **201401086**

Rajdeep Pinge **201401103**

OUTLIERS : Fig 1 , Fig 2,Fig 3 , Fig 4, Fig 8, Fig 9

Method : Here we classified the images outside the ± 0.5*STDEV window from the peak to be classified as outliers.
We took the window surrounding the peak here, as the peak represents that the corresponding x value is the most likely to be a score of a "natural" image.
NOTE : We kept 0.5*STDEV as the window for the both the above methods, in order to compare between them.

However, both the methods are not able to classify properly the Fig. 8, which is an extremely smooth image despite being natural. So this is acceptable.

Q1

```
clear
close all

load('data_lab3.mat');

ammo_conc= ammonia_concentration;

boxplot(ammo_conc)
title('Box plot for ammonia concentration(mg/L)')
ylabel('concentration (mg/L)')

figure
histogram(ammo_conc,25) % orig none
```

DAV LAB 3 Report

Aditya Joglekar 201401086

Rajdeep Pinge 201401103


```
title('Histogram for ammonia concentration(mg/L)')
xlabel('concentration (mg/L)')
ylabel('Frequency')


mu= mean(ammo_conc)
sigma= std(ammo_conc)
med= median(ammo_conc)
mode= mode(ammo_conc)

prob_greater40 = length(find(ammo_conc > 40)) / length(ammo_conc)

stan_ammo = ammo_conc;

stan_ammo = (stan_ammo - mu)/sigma; % standardzied

x = (40 - mu)/sigma

%mean(stan_ammo)
%std(stan_ammo) succ standardized

p = normcdf([-10 x]);
p_less40= p(2)-p(1);
p_more40= 1- p_less40
```

Q2 a) Gaussian distribution

```
clear
close all

load('data_lab3.mat')
arr = score_natural_model;
mean_arr = mean(arr);
std_arr = std(arr);
```

DAV LAB 3 Report

Aditya Joglekar 201401086

Rajdeep Pinge 201401103


```
start = mean_arr-3*std_arr;
ending = mean_arr + 3*std_arr;

dt = 0.01;

x = start:dt:ending;
y = zeros(size(x));
for i = 1:size(x,2)
        y(i) = (1/(std_arr*sqrt(2*pi)))*exp(-0.5*((x(i) - mean_arr)/std_arr)^2);
end

figure
plot(x,y,'g');
title('Pdf for natural image score assuming a normal distribution')
xlabel('Natural Image Score (unit less)')
ylabel('Gaussian pdf')


y1 = zeros(size(score_test));
test = score_test;
outlier = zeros(size(test));
%lets assume the ones beyond 1.5*sigma as outliers

for i = 1:size(test,1)
        hold on;

        if abs(test(i) - mean_arr) > 0.5*std_arr
        y1(i) = (1/(std_arr*sqrt(2*pi)))*exp(-0.5*((test(i) - mean_arr)/std_arr)^2);
        plot(test(i),y1(i),'r^');
        outlier(i) = 1;
        else
        y1(i) = (1/(std_arr*sqrt(2*pi)))*exp(-0.5*((test(i) - mean_arr)/std_arr)^2);
        plot(test(i),y1(i),'b^');
        end


end

legend
```

DAV LAB 3 Report

Aditya Joglekar **201401086**

Rajdeep Pinge **201401103**

Outlier

//

Q2 b) Rayleigh distribution

```
load('data_lab3.mat')
arr = score_natural_model;
mean_arr = mean(arr);
std_arr = std(arr);
sigma = 0.6451;

start = min(arr);
ending = max(arr);

dt = 0.1;

x = start:dt:ending;
y = zeros(size(x));
for i = 1:size(x,2)
        y(i) = (x(i)/(sigma^2))*exp(-0.5*((x(i)/sigma)^2));
end
figure

plot(x,y,'g');
title('Pdf for natural image score assuming a Rayleigh distribution')
xlabel('Natural Image Score (unit less)')
ylabel('Rayleigh pdf')


y1 = zeros(size(score_test));
test = score_test;
outlier = zeros(size(test));
%lets assume the ones beyond 1.5*sigma as outliers

for i = 1:size(test,1)
        hold on;

        if abs(test(i) - sigma) >0.5*std_arr
```

DAV LAB 3 Report

Aditya Joglekar **201401086**

Rajdeep Pinge **201401103**

```
        y1(i) =  (test(i)/(sigma^2))*exp(-0.5*((test(i)/sigma)^2));
        plot(test(i),y1(i),'r^');
        outlier(i) = 1;
        else
        y1(i) =  (test(i)/(sigma^2))*exp(-0.5*((test(i)/sigma)^2));
        plot(test(i),y1(i),'b^');
        end


end

outlier
```