Computer exercise 2

SF1900 Probability Theory and Statistics

Sareh Jalalizad
sarehj@kth.se

Table of contents:

Introduction

Probability and statistics, the branches of mathematics concerned with the laws governing random events, including the collection, analysis , interpretation, and display of numerical data.[1]
In this part we will briefly define definitions and theorems used in the exercises.

- Rayleigh distributed : A mathematical statement, usually applied to frequency distributions of random variables, for the case in which two orthogonal variables are independent and normally distributed with unit variance.[2]
- Density function : A function whose integral is calculated to find probabilities associated with a continuous random variable.
- ML : In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of an assumed probability distribution, given some observed data.
- LS : The method of least squares is about estimating parameters by minimizing the squared discrepancies between observed data, on the one hand, and their expected values on the other.[3]
- Confidence interval : A confidence interval is how much uncertainty there is with any particular statistic. Confidence intervals are often used with a margin of error.
- Linear regression : Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data.
- Distributions : The distribution provides a parameterized mathematical function that can be used to calculate the probability for any individual observation from the sample space.[4]
- Normal distributions : The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side. The area under the normal distribution curve represents probability and the total area under the curve sums to one.[5]
- Jarque-Bera Test : The Jarque-Bera Test,a type of Lagrange multiplier test, is a test for normality. The Jarque-Bera test is usually run before one of these tests to confirm normality. It is usually used for large data sets, because other normality tests are not reliable when n is large.[6]
- Moore’s law : Moore's Law states that the number of transistors on a microchip doubles about every two years, though the cost of computers is halved. In 1965, Gordon E. Moore, the co-founder of Intel, made this observation that became known as Moore's Law. [7]

- Multiple linear regression (MLR) : Also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.[8]

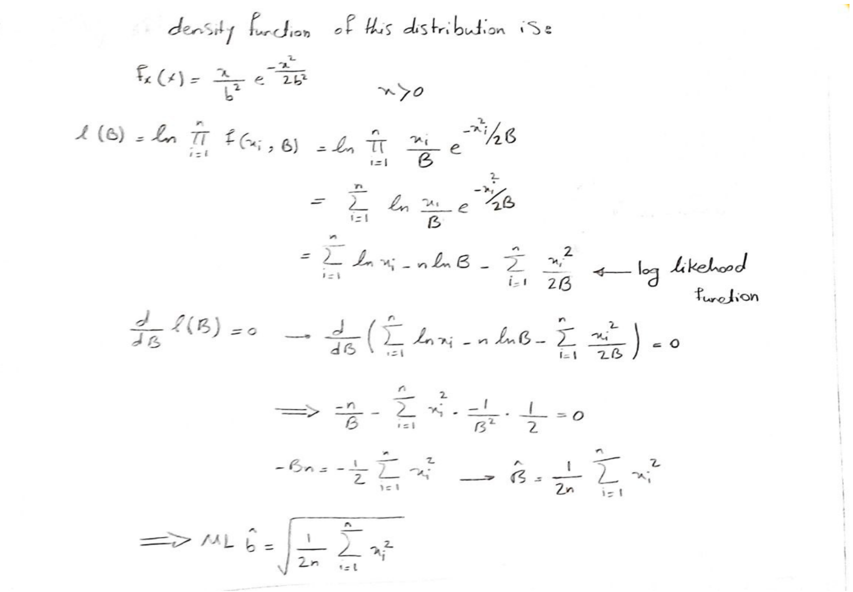
1. Preparatory exercises

1. When a random variable X has the density function

$$f_X(x) = \frac{x}{b^2} e^{-\frac{x^2}{2b^2}}, \qquad x \geq 0.$$

then it is said to be Rayleigh distributed. Assume that you have observed outcomes of n independent Rayleigh distributed random variables.

- Determine the ML estimate of the parameter.



density function of this distribution is:

$$f_X(x) = \frac{x}{b^2} e^{\frac{-x^2}{2b^2}} \qquad x>0$$

$$l(B) = \ln \prod_{i=1}^{n} f(x_i; B) = \ln \prod_{i=1}^{n} \frac{x_i}{B} e^{-x_i^2/2B}$$

$$= \sum_{i=1}^{n} \ln \frac{x_i}{B} e^{-x_i^2/2B}$$

$$= \sum_{i=1}^{n} \ln x_i - n \ln B - \sum_{i=1}^{n} \frac{x_i^2}{2B} \longleftarrow \text{log likehood function}$$

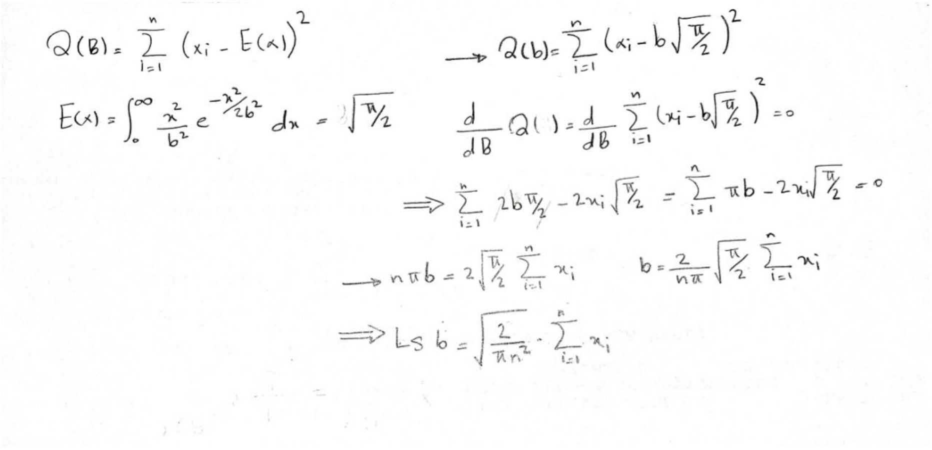
$$\frac{d}{dB} l(B) = 0 \longrightarrow \frac{d}{dB}\left(\sum_{i=1}^{n} \ln x_i - n \ln B - \sum_{i=1}^{n} \frac{x_i^2}{2B} \right) = 0$$

$$\Longrightarrow \frac{-n}{B} - \sum_{i=1}^{n} x_i^2 \cdot \frac{-1}{B^2} \cdot \frac{1}{2} = 0$$

$$-Bn = -\frac{1}{2} \sum_{i=1}^{n} x_i^2 \longrightarrow \hat{B} = \frac{1}{2n} \sum_{i=1}^{n} x_i^2$$

$$\Longrightarrow \text{ML } \hat{b} = \sqrt{\frac{1}{2n} \sum_{i=1}^{n} x_i^2}$$

- Determine the LS estimate of the parameter



$$Q(B) = \sum_{i=1}^{n} (x_i - E(x))^2 \longrightarrow Q(b) = \sum_{i=1}^{n} \left(x_i - b\sqrt{\pi/2}\right)^2$$

$$E(x) = \int_0^\infty \frac{x^2}{b^2} e^{-x^2/2b^2} dx = \sqrt{\pi/2}$$

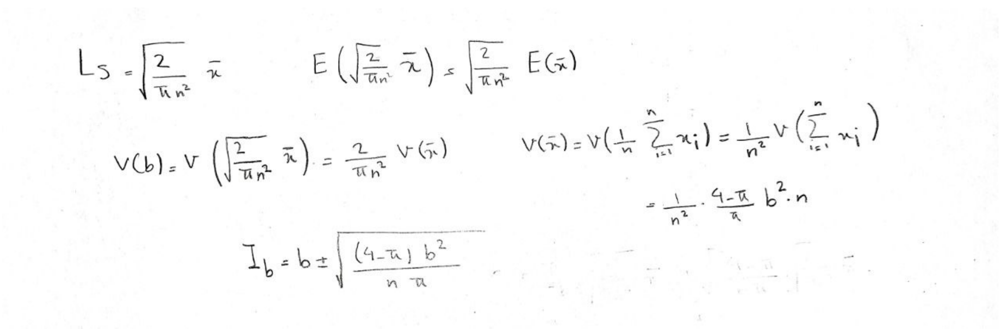
$$\frac{d}{dB} Q(\ ) = \frac{d}{dB} \sum_{i=1}^{n} \left(x_i - b\sqrt{\pi/2}\right)^2 = 0$$

$$\Longrightarrow \sum_{i=1}^{n} 2b\,\pi/2 - 2x_i\sqrt{\pi/2} = \sum_{i=1}^{n} \pi b - 2x_i\sqrt{\pi/2} = 0$$

$$\longrightarrow n\pi b = 2\sqrt{\pi/2} \sum_{i=1}^{n} x_i \qquad b = \frac{2}{n\pi}\sqrt{\pi/2} \sum_{i=1}^{n} x_i$$

$$\Longrightarrow \text{LS } \hat{b} = \sqrt{\frac{2}{\pi n^2}} \cdot \sum_{i=1}^{n} x_i$$

2. Derive a confidence interval for the parameter b with approximate con-fidence level 1−α.



$$L_S = \sqrt{\frac{2}{\pi n^2}}\, \bar{x} \qquad E\left(\sqrt{\frac{2}{\pi n^2}}\, \bar{x}\right) = \sqrt{\frac{2}{\pi n^2}}\, E(\bar{x})$$

$$V(b) = V\left(\sqrt{\frac{2}{\pi n^2}}\, \bar{x}\right) = \frac{2}{\pi n^2} V(\bar{x}) \qquad V(\bar{x}) = V\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{n^2} \cdot \frac{4-\pi}{\pi} b^2 \cdot n$$

$$I_b = b \pm \sqrt{\frac{(4-\pi)\, b^2}{n\, \pi}}$$

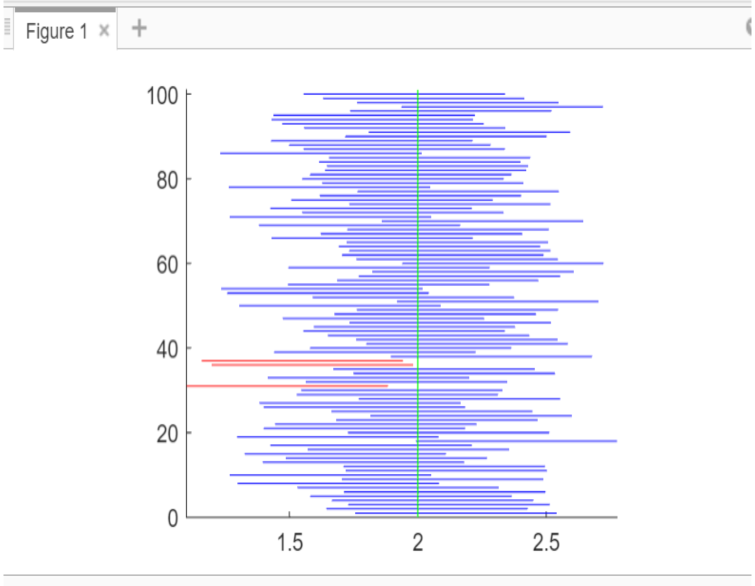
3. Describe the idea behind linear regression. Describe how the MATLAB command regress can be used to obtain estimates of the parameters in the following model

wk= log(yk) =β0+β1xk+εk

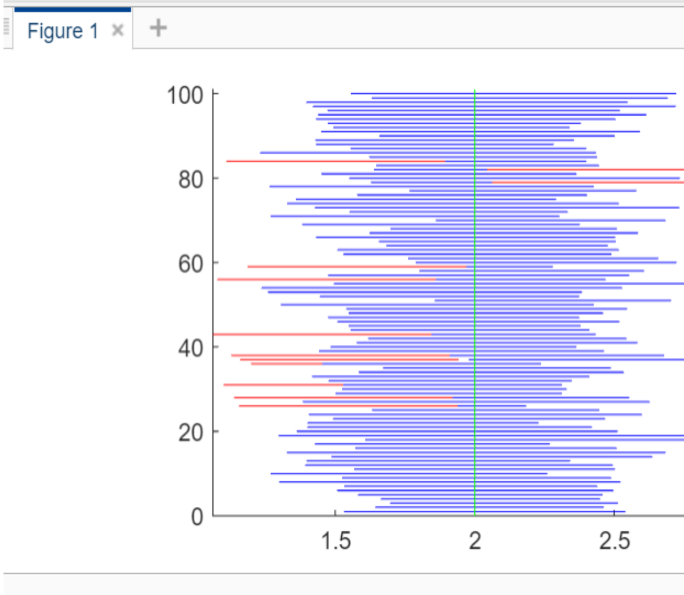
In statistics, an estimated regression equation is a formula that is used to represent the relationship between dependent and independent variables. A hypothesis about the relationship between the dependent and independent variables are first stated as a single or multiple regression model.
The least squares method is the most often used technique for estimating model parameters.
For simple linear regression, the least squares estimates of the model parameters β0 and β1 are denoted b0 and b1. An estimated regression equation is created using these estimates: = b0 + b1x.
For simple linear regression, the graph of the estimated regression equation is a straight line approximation of the relationship between y and x.

Problem 1

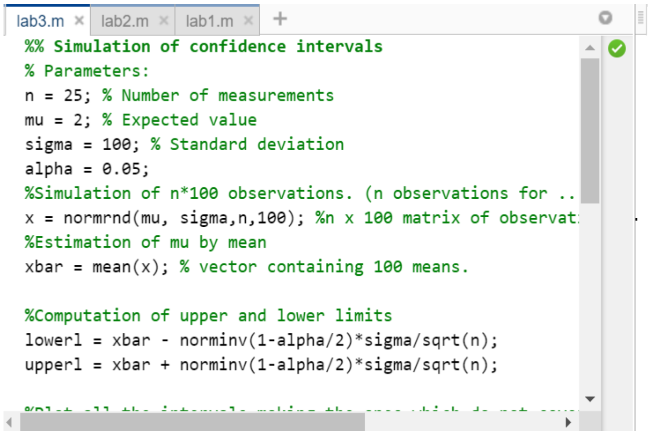
Running the simulation for the first time



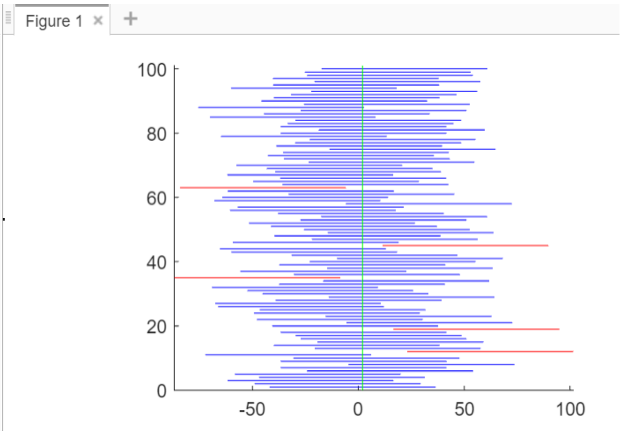
Running the simulation for the several times



Changing sigma 1 to 100

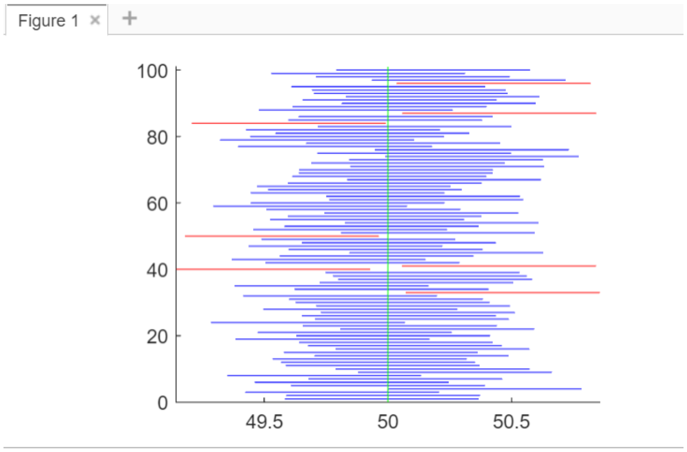


```
%% Simulation of confidence intervals
% Parameters:
n = 25; % Number of measurements
mu = 2; % Expected value
sigma = 100; % Standard deviation
alpha = 0.05;
%Simulation of n*100 observations. (n observations for ..
x = normrnd(mu, sigma,n,100); %n x 100 matrix of observat:
%Estimation of mu by mean
xbar = mean(x); % vector containing 100 means.

%Computation of upper and lower limits
lowerl = xbar - norminv(1-alpha/2)*sigma/sqrt(n);
upperl = xbar + norminv(1-alpha/2)*sigma/sqrt(n);

%Plot all the intervals making the ones which do not cove
```

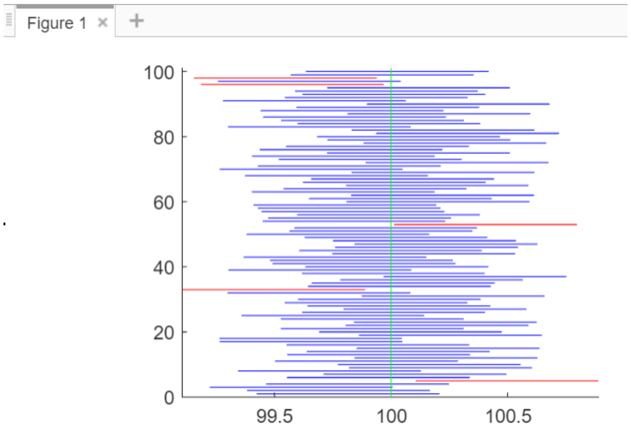


Sigma is the standard deviation and the width of a curve is determined by the sigma of a normal distribution. When sigma increases, so does the width of the confidence interval. The higher the sigma, the wider the graph.

Changing mu 2 to 50

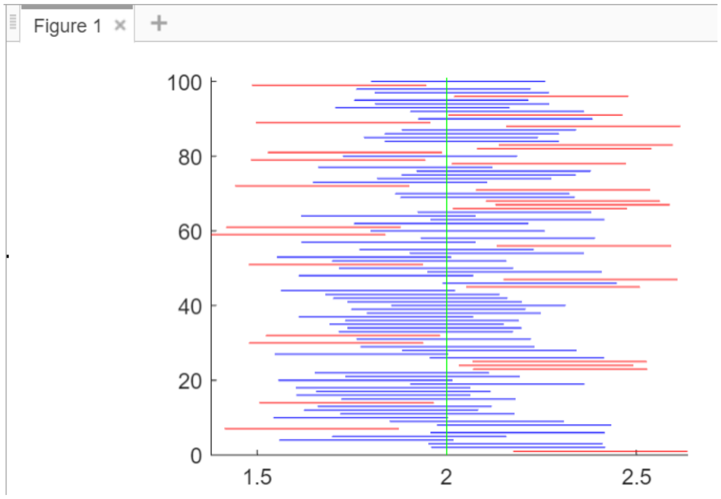


Changing mu 2 to 100

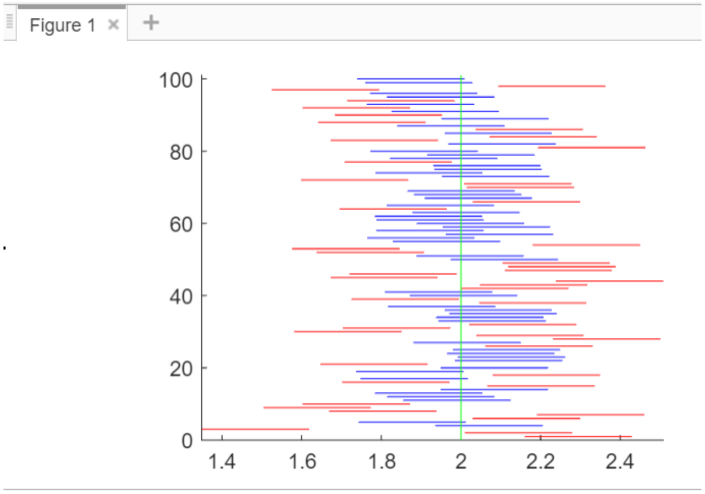


Mu has a reverse relation with Normal distribution as Normal distribution(Z) = (X-mu) / sigma.

Changing alpha 0.05 to 0.25

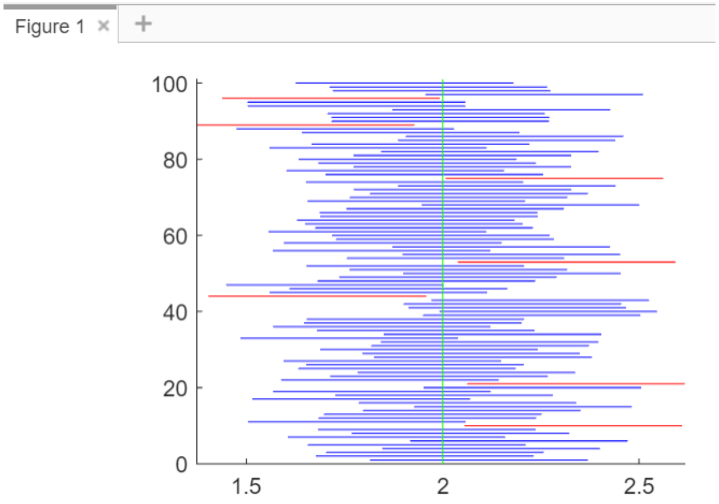


Changing alpha 0.05 to 0.50

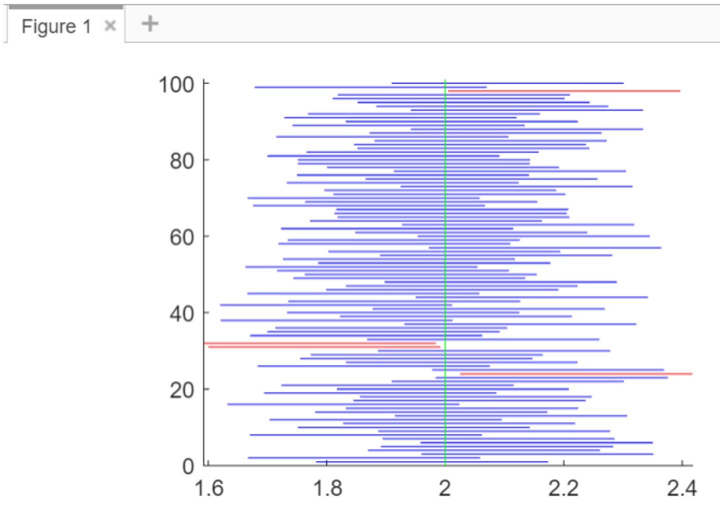


Confidence intervals are smaller and more accurate when the confidence level (1-alpha) is lower, implying that alpha is larger.

Changing n 25 to 50



Changing n 25 to 100



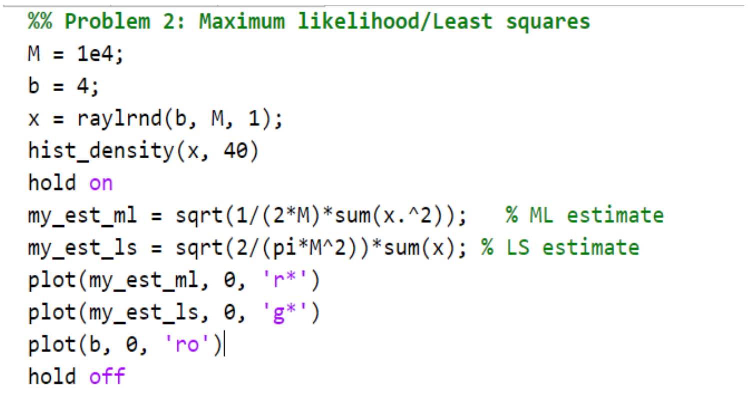
The interval narrows as n increases, resulting in a more accurate estimate.
What do the horizontal lines and the vertical line indicate? The vertical line is mu (Expected value) and horizontal lines are intervals.

Code for problem 1:

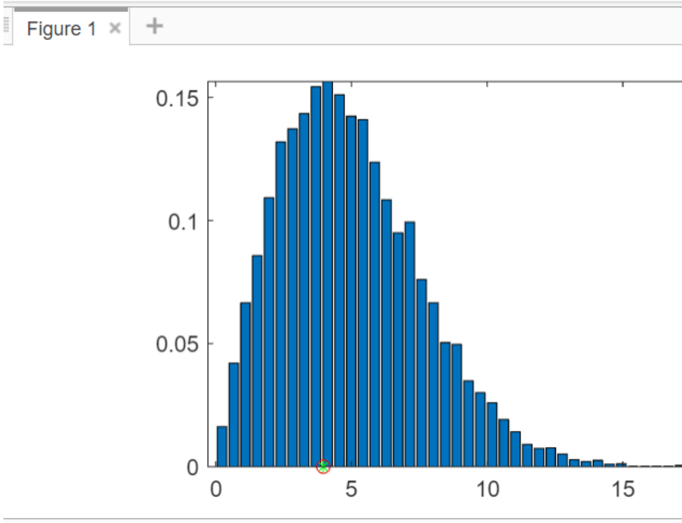
```
%% Simulation of confidence intervals
% Parameters:
n = 25; % Number of measurements
mu = 2; % Expected value
sigma = 1; % Standard deviation
alpha = 0.05;
%Simulation of n*100 observations. (n observations for ...each interval and 100 intervals)
x = normrnd(mu, sigma,n,100); %n x 100 matrix of observations
%Estimation of mu by mean
xbar = mean(x); % vector containing 100 means.
%Computation of upper and lower limits
lowerl = xbar - norminv(1-alpha/2)*sigma/sqrt(n);
upperl = xbar + norminv(1-alpha/2)*sigma/sqrt(n);
%Plot all the intervals making the ones which do not cover ...the true value red
figure(1)
hold on
for k=1:100
  if upperl(k) < mu
     plot([lowerl(k) upperl(k)],[k k],'r')
  elseif lowerl(k) > mu
     plot([lowerl(k) upperl(k)],[k k],'r')
  else
    plot([lowerl(k) upperl(k)],[k k],'b')
  end
```

```
end
%b1 and b2 are only used to make the figure look nice.
b1 = min(xbar - norminv(1 - alpha/2)*sigma/sqrt(n));
b2 = max(xbar + norminv(1 - alpha/2)*sigma/sqrt(n));
axis([b1 b2 0 101]) % Minimizes amount of unused space in ...the figure
%Plot the true value
plot([mu mu],[0 101],'g')
hold off
```

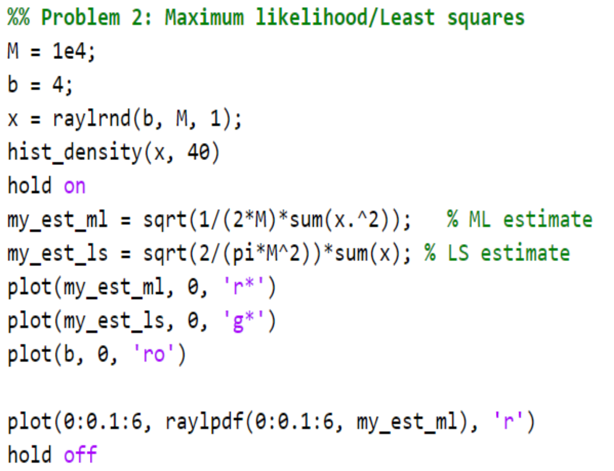
Problem 2



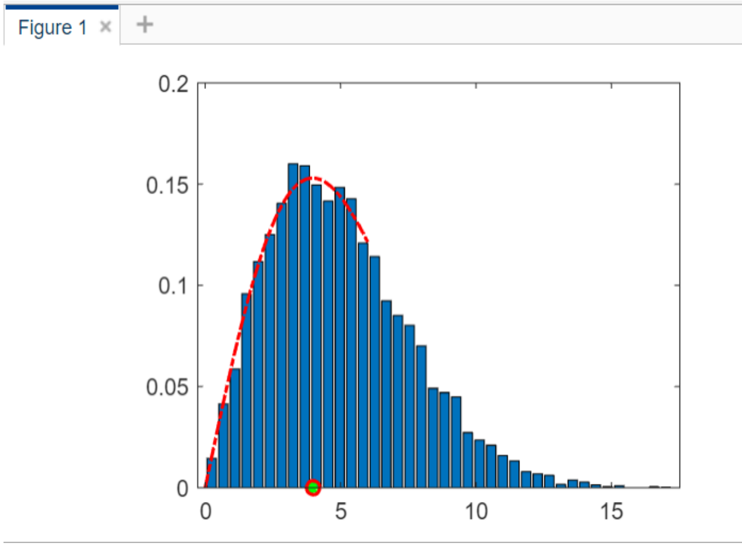
```
%% Problem 2: Maximum likelihood/Least squares
M = 1e4;
b = 4;
x = raylrnd(b, M, 1);
hist_density(x, 40)
hold on
my_est_ml = sqrt(1/(2*M)*sum(x.^2));   % ML estimate
my_est_ls = sqrt(2/(pi*M^2))*sum(x); % LS estimate
plot(my_est_ml, 0, 'r*')
plot(my_est_ls, 0, 'g*')
plot(b, 0, 'ro')
hold off
```



density function looks like by plotting it along with your estimate:



```
%% Problem 2: Maximum likelihood/Least squares
M = 1e4;
b = 4;
x = raylrnd(b, M, 1);
hist_density(x, 40)
hold on
my_est_ml = sqrt(1/(2*M)*sum(x.^2));   % ML estimate
my_est_ls = sqrt(2/(pi*M^2))*sum(x); % LS estimate
plot(my_est_ml, 0, 'r*')
plot(my_est_ls, 0, 'g*')
plot(b, 0, 'ro')

plot(0:0.1:6, raylpdf(0:0.1:6, my_est_ml), 'r')
hold off
```

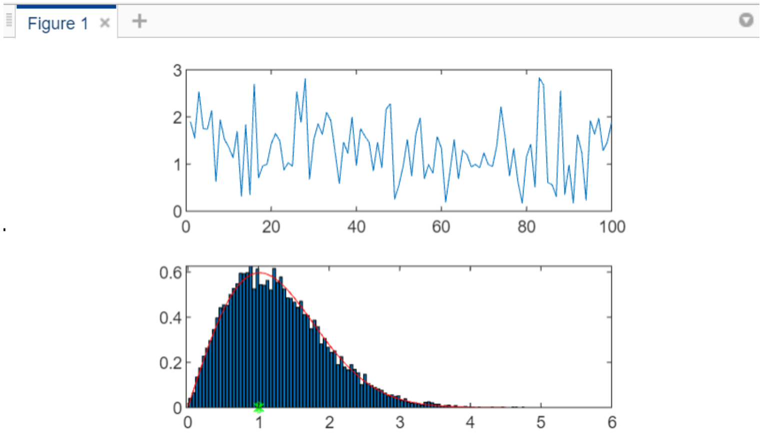


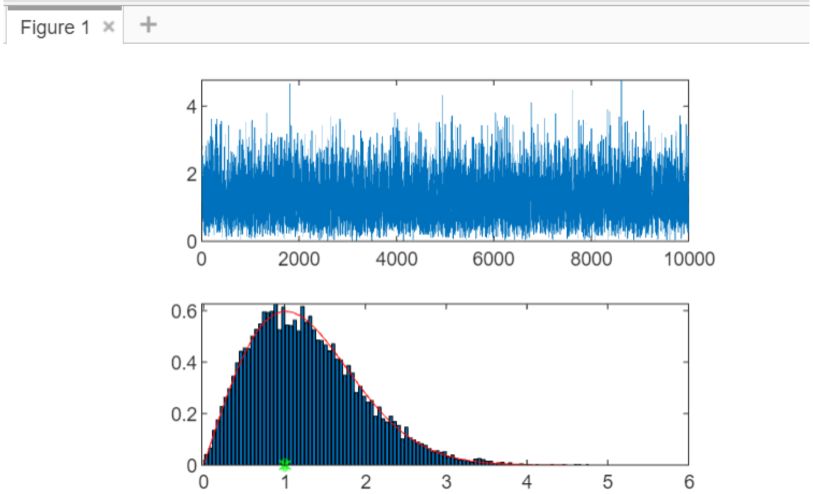
As it is obvious from the figure our estimation was very good as the density function corresponds with our estimation. Furthermore, as displayed in the figure, the blue lines have the same pattern as the red line which plays similarly to our estimation.

Code problem 2:

```
%% Problem 2: Maximum likelihood/Least squares
M = 1e4;
b = 4;
x = raylrnd(b, M, 1);
hist_density(x, 40)
hold on
my_est_ml = sqrt(1/(2*M)*sum(x.^2));   % ML estimate
my_est_ls = sqrt(2/(pi*M^2))*sum(x); % LS estimate
plot(my_est_ml, 0, 'r*')
plot(my_est_ls, 0, 'g*')
plot(b, 0, 'ro')
plot(0:0.1:6, raylpdf(0:0.1:6, my_est_ml), 'r')
hold off
```

Problem 3





change y(1:100) to y(1:end)

The distribution is very close to the data and fits it positively correlated.

And intl_leng=(upper_bound-lower_bound);" was found is 0.026793969.

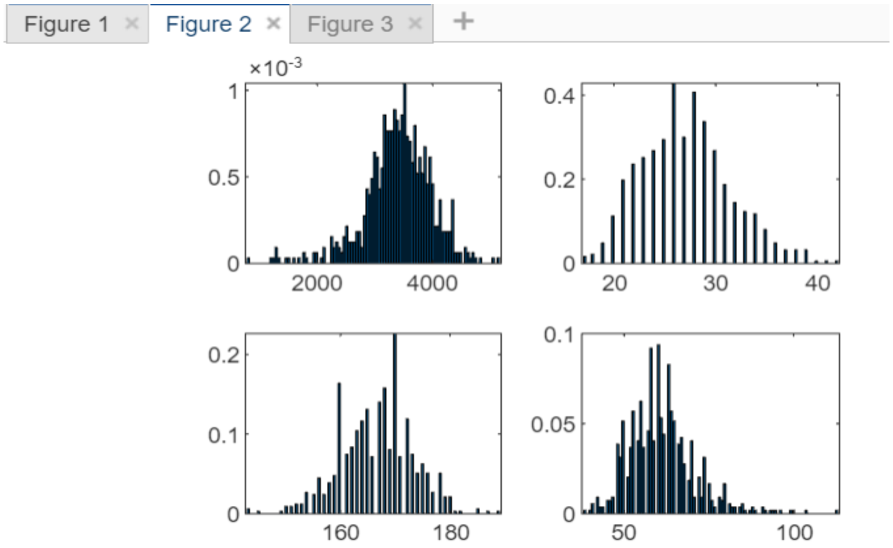
Code problem 3

```
%% Problem 3: Confidence interval for Rayleigh distribution
load wave_data.mat
subplot(2,1,1), plot(y(1:end))   %y(1:end)
subplot(2,1,2), hist_density(y)
hold on     % holds the current plot
n = length(y);
ybar = mean(y);
my_est = sqrt(2/(pi*n^2))*sum(y);
sigma = sqrt((4 - pi)/n *pi ) * my_est;
```

```
lower_bound = my_est -norminv(1-alpha/2) * sigma;
upper_bound = my_est + norminv(1-alpha/2) * sigma;
plot(lower_bound, 0, 'g*')
plot(upper_bound, 0, 'g*')
plot(0:0.1:6, raylpdf(0:0.1:6, my_est), 'r')
hold off
int_leng=(upper_bound-lower_bound);
```

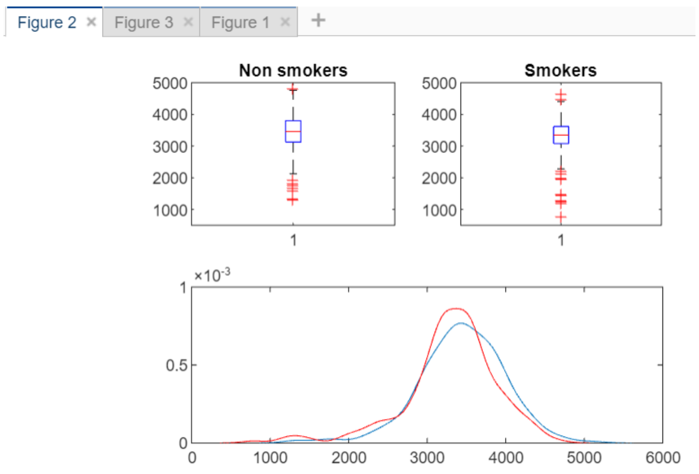
Problem 4

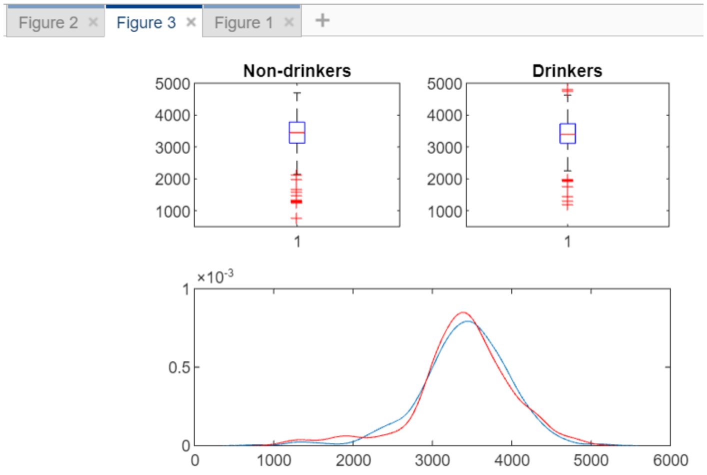
The distributions of the birth weight of the child and the age of the mother



The distributions of the length of the mother, and the weight of the Mother

The left figure shows the difference between the distribution of birth weights of children whose mothers smoked and whose mothers didn't smoke and the right one shows the difference between the distribution of birth weights of children whose mothers drank alcohol and whose mothers didn't drink.





In the first figure , the major differences between the distribution of birth weights of children whose mothers were cigarette smokers during pregnancy in comparison to those who did not smoke while pregnant.

The second figure points out the differences between the distribution of birth weights of children whose mothers drank alcoholic beverages during their pregnancy, and those who did not drink.

The blue lines are indicators for nonsmokers and nondrinkers and the red lines represent the drinkers/smokers. The plots are there to provide information and help differentiate by comparing the weights of children whose parents smoked or drank versus those to which their parents were nondrinkers and nonsmokers.

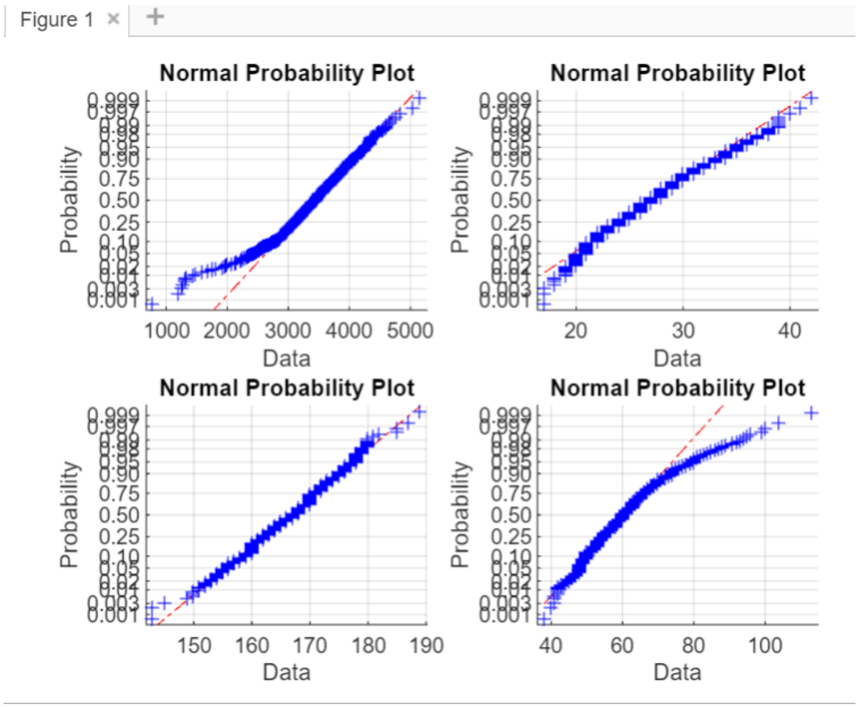
As we evaluate the graphs, we can clearly see that drinking and smoking are categorical variables that have a direct effect on the birth weight of the children.

Code Problem 4

```
%% Problem 4: Distributions of given data
load birth.dat
figure(1)
subplot(2,2,1), hist_density(birth(:,3));
subplot(2,2,2), hist_density(birth(:,4));
subplot(2,2,3), hist_density(birth(:,16));
subplot(2,2,4), hist_density(birth(:,15));

x = birth(birth(:, 20) < 3, 3); %non-smokers
y = birth(birth(:, 20) == 3, 3); %smokers
figure(2)
subplot(2,2,1), boxplot(x),
axis([0 2 500 5000])
```

```
title('Non smokers')
subplot(2,2,2), boxplot(y),
axis([0 2 500 5000])
title('Smokers')
subplot(2,2,3:4), ksdensity(x),
hold on
[fy, ty] = ksdensity(y);
plot(ty, fy, 'r')
hold off

xx=birth(birth(:,26) < 2, 3); %non drinkers
yy=birth(birth(:,26) == 2, 3); %drinkers
figure(3)
subplot(2,2,1), boxplot(xx),
axis([0 2 500 5000])
title('Non-drinkers')
subplot(2,2,2), boxplot(yy),
axis([0 2 500 5000])
title('Drinkers')
subplot(2,2,3:4), ksdensity(xx),
hold on
[fy, ty] = ksdensity(yy);
plot(ty, fy, 'r')
hold off
```

Problem 5



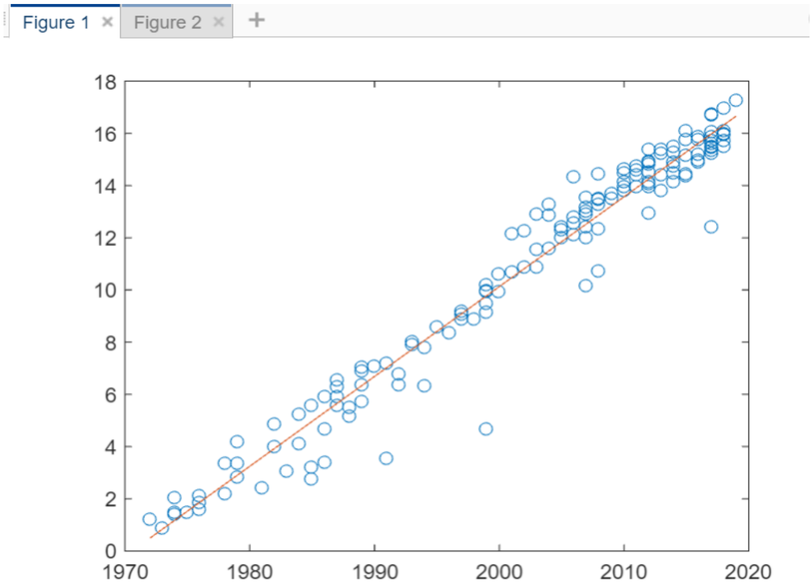
We have tried to take advantage of commands normplot which compare the empirical quantiles of the data set with the quantiles of a normal distribution to demonstrate the figures of distribution and the only normal distribution here shows to be the weight of the mother which is shown in the bottom figure. The distributions which have been done using the normplot methods are shown in the figure.

Code for problem 5

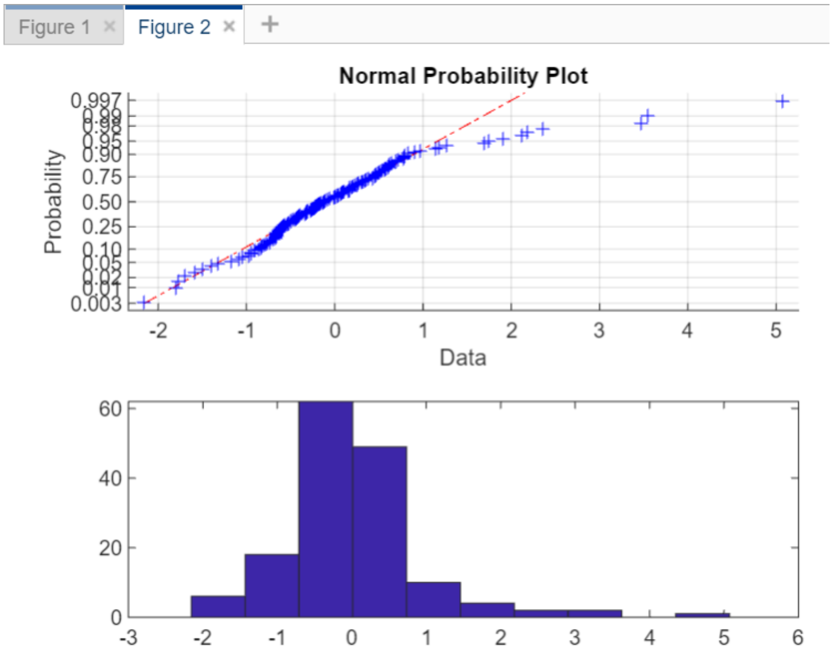
```
%% Problem 5 - Testing for normality
load birth.dat
figure(1)
```

```
subplot(2,2,1), normplot(birth(:,3)) %%weight of the child
subplot(2,2,2), normplot(birth(:,4)) %%age of the mother
subplot(2,2,3), normplot(birth(:,16)) %%length of the mother
subplot(2,2,4), normplot(birth(:,15)) %%weight of the mother
jbtest(birth(:,3))      %%weight of the child
jbtest(birth(:,4))      %%age of the mother
jbtest(birth(:,16))     %%length of the mother
jbtest(birth(:,15))     %%weight of the mother
```

Problem 6



In the graphs, the number of transistors per unit area is represented by the alphabetical letter Y while the year is represented by the letter X. this points out to us that if we plot Y against X, we will end up with a plot of the number of transistors/unit area's growth over time, and as we know that Y increases exponentially, the logarithm of the number Y is bound to increase on the same line as time passes.



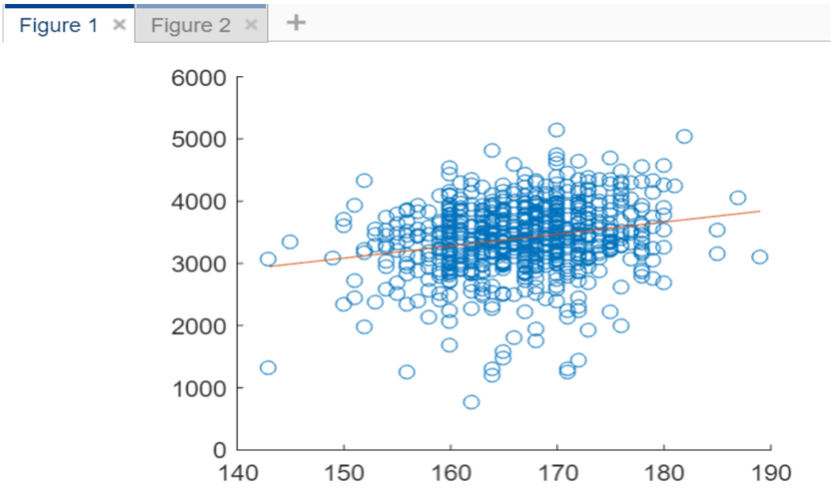
We discovered that the distribution is exponential considering the deviation observed in the figure 2, thus we can use the exponential distribution formula to estimate for the year 2025, which is 1.3599e+08. And we can see from the command Window the value of R^2 which is equal to 0,9586

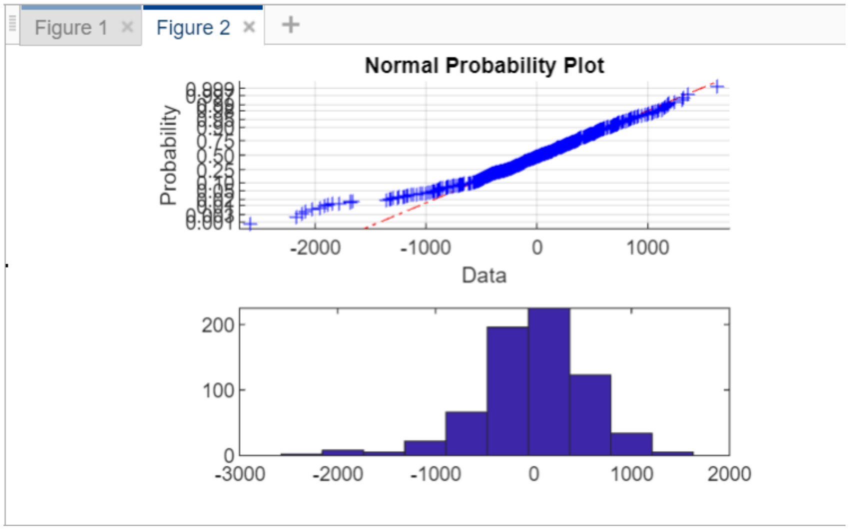
Code for problem 6

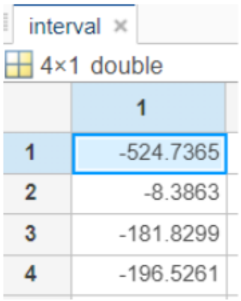
```
%% Problem 6: Regression
load moore.dat
l=length(moore);
x1=ones(l,1);
X=[x1,moore(:,1)];
Y=[moore(:,2)];
y1=log(Y);
[beta_hat,bint,~,rint,stats]=regress(y1,X); % β= (β0,β1)T
figure (1)
plot(X(:,2),y1,'o')
hold on
plot(X(:,2),X*beta_hat)
res =  X*beta_hat -y1;
R = stats(1)
figure(2)
subplot(2,1,1), normplot(res)
subplot(2,1,2), hist(res)
g=exp([1 2025]*beta_hat); %estimate of transistors/ area units in year 2025
```

Problem 7

Figure 1 represents a basic linear regression model for the relationship between a child's birth weight and the mother's length. And we created a multiple linear regression model with mother's weight, smokingbehaviors, and drinking habits.







	1
1	-524.7365
2	-8.3863
3	-181.8299
4	-196.5261

The residuals of the multiple regression model using normplot.

Figure above shows the examined factors have minor to little impact on a child’s weight since the distribution lays around 0 and gets lowered as time passes and you get closer to the extremes of the weight effect. Furthermore, based on the interval almost 95 percent of values are between -524.7365 and 196.5261, showing that most of the children’s impacted weights are between these non significant ranges.

Code for problem 7

```
%% Problem 7: Multiple linear regression
load birth.dat
mom_l=birth(:,16);
child_w=birth(:,3);
x1=ones(length(birth),1);
X=[x1,mom_l];
[beta,~,~,~,stats1]=regress(child_w,X);
figure(1)
scatter(mom_l,child_w) %child weight to a mothers length
hold on
plot(mom_l,X*beta) %estimated weight based off of mothers length
mom_w=birth(:,15); %weight
mom_s=birth(:,20); %smok
mom_d=birth(:,26); %drink
mom_s(mom_s<3)=0;
mom_s(mom_s==3)=1;
mom_d(mom_d<2)=0;
mom_d(mom_d==2)=1;
var=[x1, m_w, mom_s, mom_d];
[b,bint,res,rint,stats]=regress(child_w,var);
bint
R2 = stats(1);
interval = abs(bint(:,1)-bint(:,2));
figure(2)
subplot(2,1,1), normplot(res)
subplot(2,1,2), hist(res)
```

Summary

While carrying out this exercise, we applied MatLab to visualize and see many topics that we covered in class. As examples to those theories, we can mention LS, ML, confidence interval, normal distribution, exponential distribution, probability density function (PDF) and also comprehended a vast variety of new topics such as Rayleigh distribution and Jarque-Bera Test.

Both problems 1 and 3 cover the same basics which were simulating confidence intervals. “a parameter will fall between a set of values”. With the difference that in problem 1 we were provided with a set of codes that taught us how the confidence intervals work and function. However, in problem 3, Rayleigh distribution was being used and we wrote codes in which the data files were provided for us.

In problem 2, we figured out the value of Rayleigh distribution that was going to be used in problem 3. That was done with the help of using ML and LS.
ML and LS were both calculated in preparatory exercises. The codes to these problems were also provided for us so we only fit them in the MatLab script.

Problem 4 had a different scenario and we used the birth.dat file. This file included different variables which had a direct impact on the weight of the babies. We used different graphs which showed the effects of variables on one another.

Problem 5 was about normal distributions and compared the variables in birth.dat to normal distributions. And with the help of the Jarque bera test we could see if they had passed.
The purpose of the question was to test if a variable is normally distrusted at a 0.5 significant level .

Problem 6 is about linear regression in which we used a regress function.
Similar to this problem.

problem 7 was also about linear regression in which we used the linear regression model and multiple linear regression model.

These exercises helped me to grasp a better understanding of the concepts involved in the theory of this course and be challenged more often. The benefit of MatLab is that by showing you the graphs and tables it makes the information more comprehensive.
The purpose of this exercise was to compare the correct graphs with our own calculations and estimates in which MatLab helped us dearly with.

References :

1-https://www.britannica.com/science/probability

2-https://glossary.atis.org/glossary/rayleigh-distribution/?char=R&page_number=1&sort=ASC

3- https://stat.ethz.ch/~geer/bsa199_o.pdf

4- https://machinelearningmastery.com/statistical-data-distributions/

5-https://www.simplypsychology.org/normal-distribution.html

6-https://www.statisticshowto.com/jarque-bera-test/

7-https://www.investopedia.com/terms/m/mooreslaw.asp

8-https://www.investopedia.com/terms/m/mlr.asp