

University of Wales Trinity Saint David

School of Engineering



Analysis of Student Grades using Statistical Methods in

MATLAB

Engineering Application & Study Skills

ACMA4007

Gethin Dyer Davies

2307576

05/2024

Contents

1	Introduction	2
1.1	Objective	2
1.2	Dataset Overview	2
2	Data Import and Summary Statistics.....	3
2.1	Data Import.....	3
2.2	Summary Statistics.....	3
2.2.1	Summary Statistics Table	4
2.2.2	Limitation of Summary Statistics.	5
3	Correlation Analysis.....	6
3.1	Pairwise Linear Correlation Coefficients	6
3.2	Hypothesis Test	7
4	Grouped Data and Exam 1, Exam 2, and O/All Results	9

4.1	Distribution of Frequencies	9
4.2	Histograms	9
4.3	Ogives	10
5	Discussion and Conclusion	12
5.1	Discussion of Findings and Implications.....	12
5.2	Strengths and Limitations of Statistical Methods.....	12
5.3	Conclusion	12
6	References.....	14
7	Appendix	16

1 Introduction

1.1 Objective

This assignment objective is to gain insight into the factors affecting student performance and understanding the statistical techniques used to analyse such data. This is carried out by writing a MATLAB coding for applying various statistical methods such as correlation coefficients, hypothesis test, and visualisation of grouped data using histogram. The aim in this report is to analysis and uncover patters or relationships between gender, attendance, age, and the overall exam scores from the dataset supplied.

1.2 Dataset Overview

The data set provided to be analysed named “student _grades.csv” contains data from 100 students. The data presented in columns are:

- Student ID
- Gender: Male or Female
- Age
- Attendance: Attendance in percentage
- Exam_1: 1st exam score
- Exam_2: 2nd exam score

This data is used to see if there are a common factor or factors that may contribute to the exam results.

2 Data Import and Summary Statistics

2.1 Data Import

The dataset was downloaded on to the laptop MATLAB documents and using the 'readtable' function in MATLAB. 'readtable' function creates a table by reading column-oriented data from the Excel data sheet downloaded (MATLAB, Documentation. 2024)

2.2 Summary Statistics

To add the overall exam score for each student to the summary statistics table a new column had to be created "All" and a command for the exam score 1 and exam score 2 of each student separately to be added together and then divided by two to retrieve the exams overall average score as shown in *Figure 1* below.

```
% Moved from task C to add to table2array  
SGD("All") = (SGD("Exam_1") + SGD("Exam_2")) / 2; % Adding O/All score. Being the average of both exam score.
```

Figure 1 - The command used to calculate the overall average score and present in the dataset.

The numerical data (Age, Attendance, Exam 1, Exam 2, and Overall Exam Score) was calculated by the MATLAB coding by converting the numerical data to array and then using the individual commands as shown in *Figure 2*. The statistics calculated are,

- Mean - Found by adding all the figure in the column and dividing by the number of students.
- Median – Found by placing the figures in numerical order and selecting the middle figure.
- Mode – The number that occurs the most amount.
- Range of Min and MAX (Range of Both) - Finding the lowest number in its column and the highest number
- Standard Deviation – How far the data is from its mean.

(Khan Academy, 2024)

```

ND=table2array(SGD(:,3:end));% Convert Numerical Data to Array

% b) Compute summary for each numerical variable
% i) Mean
Mean = mean(ND)';

% ii) Median
Median = median(ND)';

% iii) Mode
Mode = mode(ND)';

% iv) Range
Min_Range = min(ND)';
Max_Range = max(ND)';
R_O_B = [Min_Range, Max_Range]; %Range Of Both%   %Min and Max Range of All Variables

% v) Standard Deviation
STD = std(ND)';

```

Figure 2 - table to array command and the summary statistic command in MATLAB

The results of these calculated central tendencies and dataset distribution are presented in the *Table 1* below.

2.2.1 Summery Statistics Table

Table 1 - Summery Statistics

Variable	Mean	Median	Mode	R_O_B		STD
"Age"	36.65	37.5	48	19	53	9.7643
"Attendance"	60.58	57	27	20	100	24.249
"1st Exam Score"	58.27	58	59	7	98	21.673
"2nd Exam Score"	55.08	53.5	51	4	98	22.428
"O/All Score"	56.675	54.5	25.5	9.5	95	21.452

The table reveals that the median and means are close to each other indicating that data distribution is relatively symmetrical, the central attendances show that the average exam score is 56.675 %.

Standard deviation reveals that the data in “Attendance, 1st Exam Score, 2nd Exam Score and O/All score (Overall Exam Score)” is at a wider range from their mean compared to “Age”.

2.2.2 Limitation of Summery Statistics.

Summary statistic offer a quick and simple overview of the data set but do not capture the complexity of data distributions. These are some of the limitations.

- The limitation of mean is that it can be influenced by skewed distribution and any data point that differs significantly from others.
- Standard Deviation gives more weight to extreme values and less on those near the mean, giving a less equal representative of the data and doesn't indicate the direction of variability.
- Median doesn't use all the observations in the data set in its calculations and this may result in median and mean figures becoming far apart for example data set (1,2,3,99,100). The median would be 3 but the mean would be 41.

With these examples and viewing the table as well, summer statistics condenses the data into few key outputs. This is potentially hiding important details, patterns and doesn't provide information on relationships between variables and trends. (Kappal, 2018)

3 Correlation Analysis

Correlation analysis is a statistical method used to evaluate the strength and direction of the linear relationship between two variables. The correlation analysis helps us understand the relationship between variables. This is important to understand what factors effect student performance and how to improve their scores. *Figure 3* is the correlation matrix this was visualised using ‘corrplot’ function providing a graphical representation of the relationship between the variables, this help to understand the results. (Akoglu, 2018)

3.1 Pairwise Linear Correlation Coefficients

The pairwise linear correlation coefficients indicate’ the degree of linear associated between each pair of variables. The values close to 1 indicate a strong linear relationship, while the values close to 0 indicate a weak or a no linear relationship.

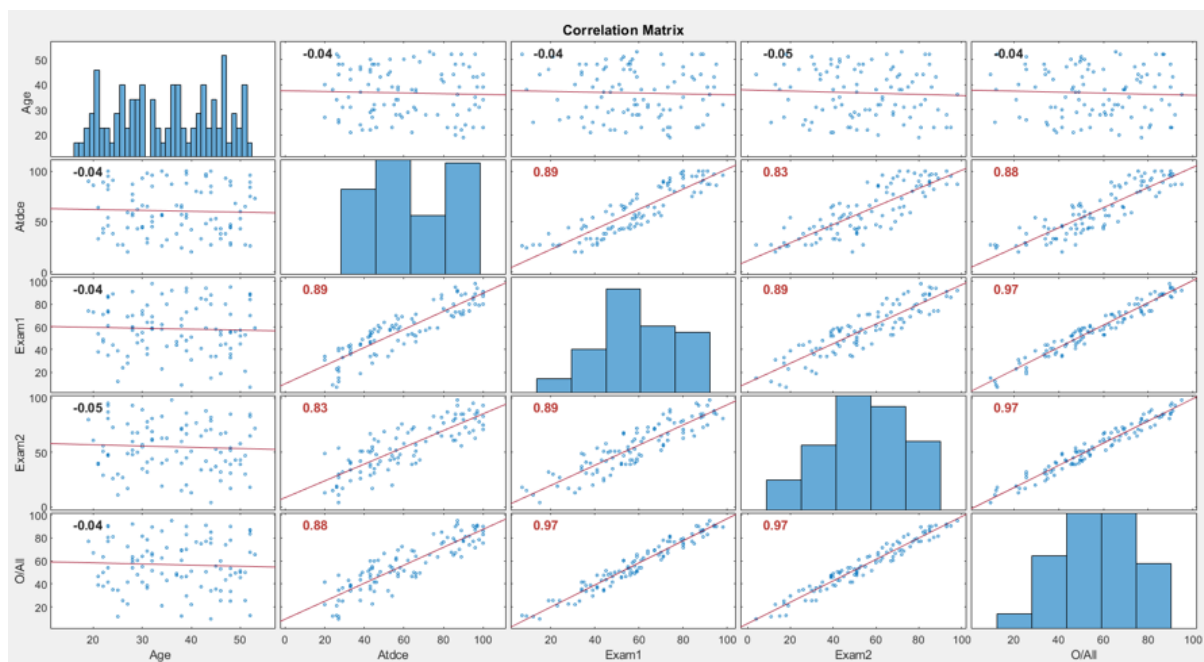


Figure 3- Results of the pairwise linear correlation coefficients.

The results show that there is a strong positive relationship between attendance and Exam 1, Exam 2, and overall exam score (O/All), with the lowest reading at 0.83 (Exam 2). This indicates that the

students with the higher attendance have a higher exam score compared to the ones with the lower attendance percentage. Exam 1 also has a strong correlation with Exam 2 showing that the student that achieved a high score in exam have scored high in Exam 2, were as the students that had a lower score in Exam 1 also had a lower score in Exam 2. This shows the student performance are consistent.

The matrix also show that Age had a very weak correlation indicating that Age has no significant factor in determining the students' performance.

This data highlights the importance of attendance in relation to student performance. This finding can be used to encourage students to attend. Lecture would also see the importance of attendance and may seek support for individual students and review ways of improving attendance.

3.2 Hypothesis Test

Hypothesis test is used in this report to see if there is a difference between the females overall score to compare with the males overall score. The "hypothesis test is used to assess whether a difference between two samples represents a real difference between the populations from which the samples were taken." (Walker, 2019).

Using the function of 'ttest2' on MATLAB, this tests the null hypothesis that the two data are from populations with equal means. This test also returns a p-value as a measure of the evidence against the null hypothesis. The results are shown in *Figure 4* below as calculated and displayed by MATLAB.

```
H =  
0  
  
P =  
0.4370
```

Figure 4 - ttest2 results in MATLAB.

$H=0$ states that the data is a Null Hypothesis, Null Hypothesis is a statement that there's no difference in the overall scores between the genders. This shows that being a Female or a Male would be a disadvantage when it comes to exam results in this dataset.

$P=0.437$ is equal to a p-value of 43.7% probability of observing the sample data. This states that there's a high indication that the data observed is true. This confirms that the findings in the Null Hypothesis is strong and conclude that the observed differences in the data are likely due to random variation.

4 Grouped Data and Exam 1, Exam 2, and O/All Results

4.1 Distribution of Frequencies

The data of Exam 1, Exam 2 and O/All results was distributed into classes of width 10 ranging from the lowest score possible in an exam of 0% up to the highest score possible 100%, this was carried out using the 'histcounts' function in MATLAB. The advantage of this is that it simplifies the interpretation of large dataset by organising the data into related groups. This allows us to see the distribution of frequencies more clearly that can help identify central tendencies within the data.

4.2 Histograms

Using the 'histcounts' results, Histograms are used to visualize the distribution of the data. We can clearly view the distribution of scores of Exam 1, Exam 2 and O/All scores in *Figure 5* below.

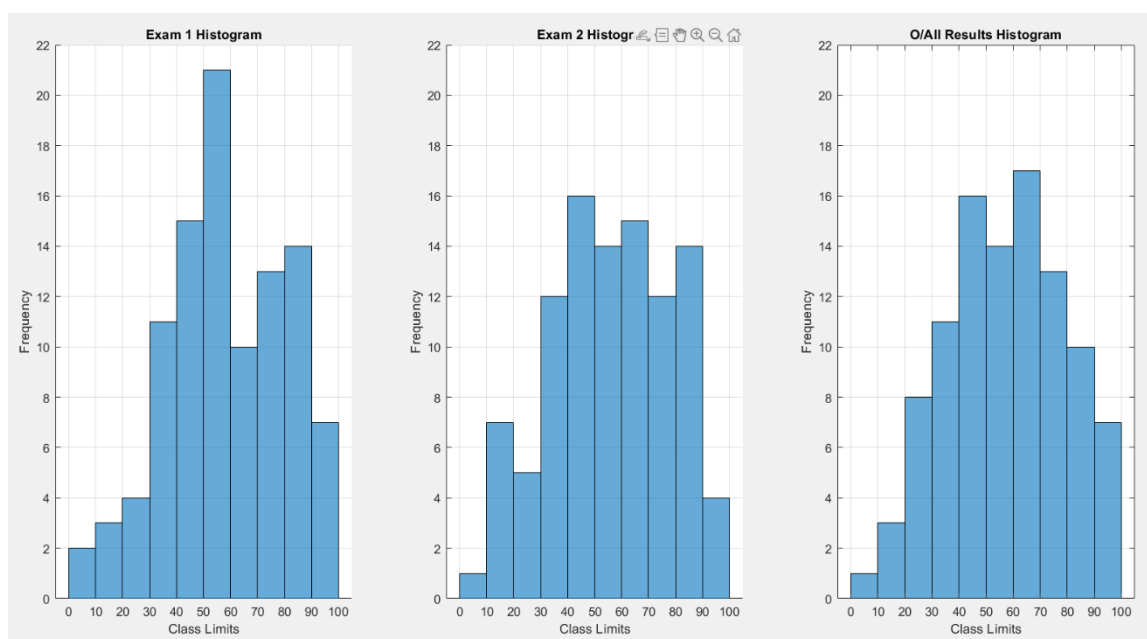


Figure 5 - Grouped data displayed in Histograms.

The advantages of histograms are that they give you a simple and clear visualization of the frequency helping identify quickly patterns and figures with the relation to the bars shown. We can have a visual for skewness or a symmetric distribution pattern, helping to identify and understand the overall shape and distribution of the data. Histograms can help in identifying unusual data points, this can be done

by identifying very few or very high frequencies which doesn't align with the pattern. Multiple histograms can be used to compare against each other with ease allowing the understanding of distribution across different databases or verifying data credibility against each other. (Javatpoint, 2024)

Figure 5 shows that the distribution of all three histograms is skewed slightly to the right with most of the student scoring are between 30 to 90. Exam 1 peak frequency is between 50-60 with a score of 21 but with the ability to compare with Exam 2 this may be an anomaly as its a lot higher and the frequency in class limits 60-70 is lower than the patterns in Exam 2.

4.3 Ogives

Ogives shows a clear representation of the cumulative (relative) distribution of frequencies. They provide a way of visualizing the cumulative totals and are good at identifying the number or percentage of observation at a particular value. Both types of ogives are used in *Figure 6*.

- Less than Ogive, (Represented as the red curve in *Figure 6*) – Plots the cumulative frequency against the upper class limit.
- More than Ogive, (Represented as the blue curve in *Figure 6*) – Plots the cumulative frequency against the lower class limit.

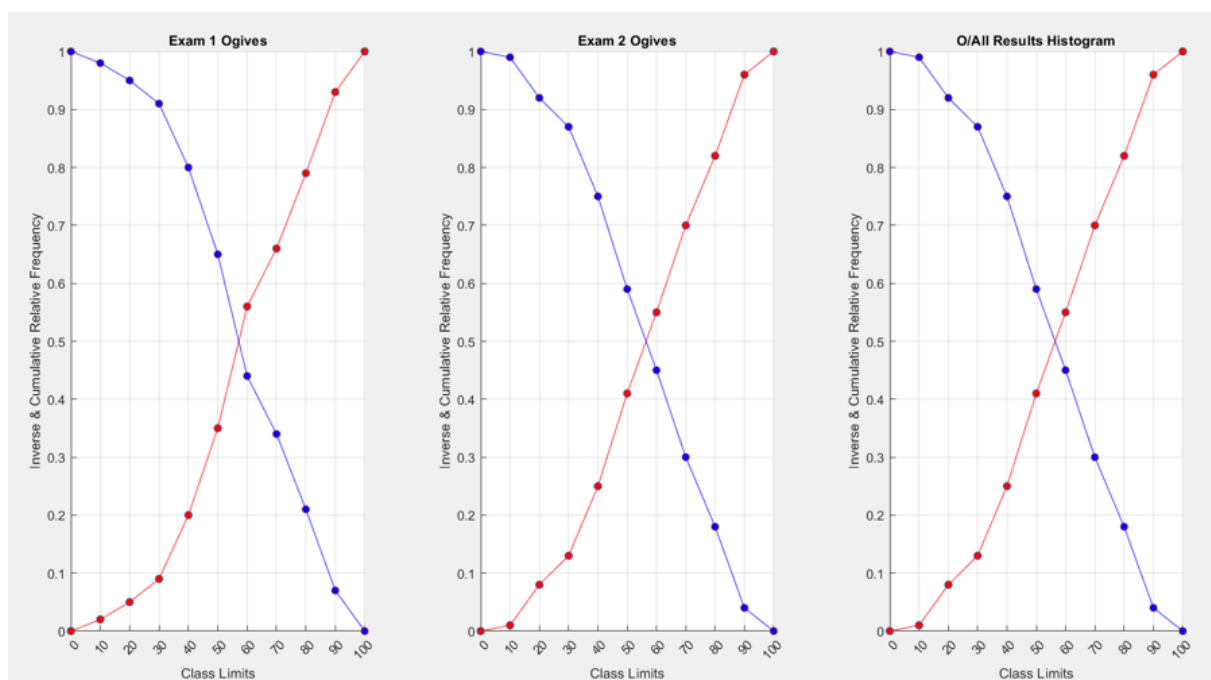


Figure 6- Ogives Results

The advantages of Ogives are that they provide a quick and clear representation of the relative frequency data, making it easy to see the number or percentage or a specific value. Seeing, setting or understand thresholds and cut-off point are easy to understand also allowing you to look for central tendencies. Median can be identified by identifying the crossover point of the lines. Multiple Ogives graphs as per above can be used to compare against each other helping to understand how performance varies across different tasks (exams in the *Figure 6* case).

Figure 6 shows that all three graphs are skewed slightly to the right indicating that the majority of students had scored less than 60, and all three graphs line intercepts between the 50-60 suggesting that all 3 have a similar median.

5 Discussion and Conclusion

5.1 Discussion of Findings and Implications

The analysis of the dataset revealed in the correlation matrix a strong positive correlation between the exam scores and attendance, suggesting a high attendance is important for achieving high grades. This finding has an important implication for educational strategies and planning, being able to emphasize the importance of attendance with data, concentrate on supporting students with low attendance and considering teaching skills and methods that could apply to be more flexible to allow students to attend when available. This could significantly improve student outcomes. Whereas age had very weak correlation indicating that age is not a factor in performance or attendance.

The Hypothesis test indicated that there was no significant difference between the female and male genders with regarding their overall exams scores suggesting gender doesn't play a role in performance.

5.2 Strengths and Limitations of Statistical Methods

The methods used have been positive in giving a clear visual understanding of trends and grouping in the form ability to identify the shape, spread and central tendency of the data. However, summary statistics condenses the data into few key outputs. This can potentially hide important details, patterns. Correlation analysis doesn't imply causation this means a high correlation between two variables doesn't mean that one variable causes the change to the other as stated by Australian Bureau of Statistics "The correlation coefficient should not be used to say anything about cause-and-effect relationship. By examining the value of 'r', we may conclude that two variables are related, but that 'r' value does not tell us if one variable was the cause of the change in the other." (Australian Bureau of Statistics, 2024). With all this considered the data provided doesn't provide any information regarding student background information or affairs that might factor their performance.

5.3 Conclusion

The data analysed of the students' performance against age, attendance and gender displayed important findings in the factors what may impede or improve students' performance. Strong correlation between exam scores and attendance and the Hypothesis test revealing that there are no mitigating factors between genders and performance. This insight can allow curriculum design to be targeted at improving attendance to improve overall performance of each student. However, it's

important to remember that further research to explore potential external influences on student performance would be required for a more accurate analysis.

6 References

Akoglu, H, 2018. User's guide to correlation coefficients. Turkish Journal of Emergency Medicine. 18. 10.1016/j.tjem.2018.08.001.

Australian Bureau of Statistics, 2024. Correlation and Causation

[Correlation and causation | Australian Bureau of Statistics \(abs.gov.au\)](#)

Javatpoint, 2024. Advantages and Disadvantages of Histogram

[Advantages and Disadvantages of Histogram - Javatpoint](#)

Khan Academy, 2024. Mean, median, and mode review (article)

[Mean, median, and mode review \(article\) | Khan Academy](#)

Kappal S, 2018. Limitation of the Measure of Central Tendency Statistics

[Limitations of the Measure of Central Tendency Statistics – DATA DOJO](#)

[\(wordpress.com\)](#)

Matlab, 2104. Matlab Documentation.

<https://www.mathworks.com/help/matlab/>

Walker, J., 2019. Hypothesis tests. *BJA Education*, 19(7), pp.227-231. DOI:

<https://doi.org/10.1016/j.bjae.2019.03.006>.

7 Appendix

```
clc
clear
close all

%***** Task I, Data Import and Summary Statistics *****

% a) Import the dataset into Matlab
SGD = readtable('student_grades.csv'); %Student Grades Dataset (CSV file)
% Moved from task C to add to table2array
SGD.("All") = (SGD.("Exam_1") + SGD.("Exam_2")) / 2; % Adding O/All score. Being
the average of both exam score.
ND=table2array(SGD(:,3:end));% Convert Numerical Data to Array

% b) Compute summary for each numerical variable
% i) Mean
Mean = mean(ND)';

% ii) Median
Median = median(ND)';

% iii) Mode
Mode = mode(ND)';

% iv) Range
Min_Range = min(ND)';
Max_Range = max(ND)';
R_O_B = [Min_Range, Max_Range]; %Range Of Both%    %Min and Max Range of All
Variables

% v) Standard Deviation
STD = std(ND)';

% c) Use function table to display the summary statistics in the command window.
Variable = ["Age";"Attendance";"1st Exam Score";"2nd Exam Score";"O/All Score"];
%variable Column for table
disp('Below is a table showing the summary statistics')
ST = table(Variable,Mean, Median, Mode, R_O_B, STD) %Summary statistics table

%***** Task II, Correlation Analysis *****

% a) Matrix of the pairwise linear correlation coefficients
disp('The table below shows the pairwise linear correlation coefficients')
CC = corr(ND(:,:)) %Correlation Coefficients

% b) Visualise the correlation matrix
VariNames = {'Age', 'Atdce', 'Exam1', 'Exam2', 'O/All'}; %Names of the variables in
the correlation matrix
figure;1;
CG = corrplot(ND, "varNames", VariNames, TestR = 'on'); %Correlation Graphs/Plots
with variable names and 'TestR'

% c) The relationship between the genders and overall module results
```

```

% i) Organise your data: contain gender, overall module results
GD_Names = {'Gender', 'O/All Score'}; %Headings for table
GD = table(SGD{:, "gender"}, SGD{:, "All"}, 'VariableNames', GD_Names); %Gender and
overall exam score data

% ii) Group the data: Gender specific with overall exam result.
FD = GD{GD.Gender == "Female", "O/All Score"}; %Overall exam scores by females
MD = GD{GD.Gender == "Male", "O/All Score"}; %Overall exam scores by males

% iii) Perform a hypothesis test
[H,P] = ttest2(FD,MD);

%iv) Display the test results
disp('The hypothesis test is used to assess whether there is a difference in
overall score between the females and the males')
if H==0
    disp('H = 0 is a Null Hypothesis, this states that there are no difference in
the overall score between the genders')
else
    disp('H = 1 is a Alternative, this states that there is a difference in the
overall score between the genders')
end
% Results of P is from 1 to 0
if P <= 0.05
    fprintf(['The p-value is statistically significant and is equal or less than
0.05. The result is %g \n. ' ...
'This shows that the obtained evidence does not provide enough evidence to
support Null Hypothesis'],P)
else
    fprintf(['The p-value is above statistically significant level of 0.05. ' ...
'The result is %g \n. This shows that the obtained evidence does provide
enough evidence to support Null Hypothesis'],P)
end

%***** Task III, Visualisation of grouped data for Exam 1, Exam 2 and O/All
results *****
% a) Using histcounts, determine the frequency of,
% i) Exam 1 results grouped in classes of width 10
Edges = 0:10:100; %width of 10 covering all exam score possibilities
HC_E1 = histcounts(SGD.Exam_1,Edges); %Frequency by histcounts for Exam 1

% ii) Exam 2 result grouped in classes of width 10
HC_E2 = histcounts(SGD.Exam_2,Edges); %Frequency by histcounts for Exam 2

% iii) O/All results grouped in classes of width 10
HC_ALL = histcounts(SGD.All,Edges); %Frequency by histcounts for Exam 2

maxCount = max([HC_E1, HC_E2, HC_ALL]); %Find the highest frequency
Ytick = (0:2:maxCount +1); %Yticks for all 3 graphs
Ylim = ([0 maxCount + 1]); %Ylim for all 3 graphs

% b) Using function histogram, plot in one figure the histograms for,
% i) The classified Exam 1 results,
figure; %New figure window for histograms
subplot(1,3,1) %Attaching histogram graphs next to each other in figure 2
hold on

```

```

histogram('BinEdges',Edges,'BinCounts',HC_E1)%Histogram
yticks(Ytick)%Display tick marks along the y-axis
xticks(Edges)
ylim(Ylim)%Y axis limits
ylabel('Frequency')
xlabel('Class Limits')
title('Exam 1 Histogram')
grid

% ii) The classified Exam 2 results,
subplot(1,3,2);%Attaching histogram graphs next to each other in figure 2
hold on
histogram('BinEdges',Edges,'BinCounts',HC_E2)%Histogram for Exam 2
yticks(Ytick)%Display tick marks along the y-axis
xticks(Edges)
ylim(Ylim)%Y axis limits
ylabel('Frequency')
xlabel('Class Limits')
title('Exam 2 Histogram')
grid

% iii) The classified O/All results,
subplot(1,3,3);%Attaching histogram graphs next to each other in figure 2
histogram('BinEdges',Edges,'BinCounts',HC_ALL)%Histogram for Exam 2
yticks(Ytick)%Display tick marks along the y-axis
xticks(Edges)
ylim(Ylim)%Y axis limits
ylabel('Frequency')
xlabel('Class Limits')
title('O/All Results Histogram')
grid

% c)Using function plot, plot in one figure the more (in blue) and less (in red)
than ogives for,
% i) the classified Exam 1 result
%Frequency Distribution
CW = 10; %Class Width
LCL=0: CW:90; %Lower Class Limit Vector
HCL=LCL+CW; %Higher Class Limit Vector
M=(LCL+HCL)/2; %Class Marks Vector
CF=cumsum(HC_E1); %Cumulative Frequency Vector
R=HC_E1/sum(HC_E1); %Relative Frequency Vector
RCF=cumsum(R); %Cumulative Relative Frequency Vector
ICRF=cumsum(R,'reverse'); %Inverse Cumulative Relative Frequency Vector

%Ogives
figure;3;%New figure window for ogives
subplot(1,3,1);%Attaching ogives graphs next to each other in figure 3
hold on
plot([LCL(1) HCL],[0 RCF],'LineStyle','none','Marker','o','MarkerFaceColor','r')
%Less Than Ogive
plot([LCL HCL(end)],[ICRF
0],'LineStyle','none','Marker','o','MarkerFaceColor','b') %More Than Ogive
xlim([LCL(1) HCL(end)])
xticks(Edges)

```

```

ylim([0 1])
yticks(0:0.1:1)
title('Exam 1 Ogives')
ylabel('Inverse & Cumulative Relative Frequency')
xlabel('Class Limits')
grid

%Interpolation
NQP=10; %Number of query points
xq=LCL(1):(HCL(end)-LCL(1))/NQP:HCL(end); %x Location of the query points
yq=interp1([LCL(1) HCL],[0 RCF], xq,'spline'); %Less Than Ogive
plot(xq,yq,'r')
yq2=interp1([LCL HCL(end)],[ICRF 0],xq,'Spline'); %More Than Ogive
plot(xq,yq2,'b')

% ii) the classified Exam 2 result
%Frequency Distribution
CF2=cumsum(HC_E2); %Cumulative Frequency Vector
R2=HC_E2/sum(HC_E2); %Relative Frequency Vector
RCF2=cumsum(R2); %Cumulative Relative Frequency Vector
ICRF2=cumsum(R2,'reverse'); %Inverse Cumulative Relative Frequency Vector

%Ogives
subplot(1,3,2)%Attaching ogives graphs next to each other in figure 3
hold on
plot([LCL(1) HCL],[0 RCF2],'LineStyle','none','Marker','o','MarkerFaceColor','r')
%Less Than Ogive
plot([LCL HCL(end)],[ICRF2
0],'LineStyle','none','Marker','o','MarkerFaceColor','b') %More Than Ogive
xlim([LCL(1) HCL(end)])
xticks(Edges)
ylim([0 1])
yticks(0:0.1:1)
title('Exam 2 Ogives')
ylabel('Inverse & Cumulative Relative Frequency')
xlabel('Class Limits')
grid

%Interpolation
NQP=10; %Number of query points
xq=LCL(1):(HCL(end)-LCL(1))/NQP:HCL(end); %x Location of the query points
yq=interp1([LCL(1) HCL],[0 RCF2], xq,'spline'); %Less Than Ogive
plot(xq,yq,'r')
yq2=interp1([LCL HCL(end)],[ICRF2 0],xq,'Spline'); %More Than Ogive
plot(xq,yq2,'b')

% iii) the classified O/All result
%Frequency Distribution
CF3=cumsum(HC_ALL); %Cumulative Frequency Vector
R3=HC_E2/sum(HC_ALL); %Relative Frequency Vector
RCF3=cumsum(R3); %Cumulative Relative Frequency Vector
ICRF3=cumsum(R3,'reverse'); %Inverse Cumulative Relative Frequency Vector

%Ogives
subplot(1,3,3)%Attaching ogives graphs next to each other in figure 3
hold on

```

```

plot([LCL(1) HCL],[0 RCF3],'LineStyle','none','Marker','o','MarkerFaceColor','r')
%Less Than Ogive
plot([LCL HCL(end)],[ICRF3
0],'LineStyle','none','Marker','o','MarkerFaceColor','b') %More Than Ogive
xlim([LCL(1) HCL(end)])
xticks(Edges)
ylim([0 1])
yticks(0:0.1:1)
title('O/All Results Histogram')
ylabel('Inverse & Cumulative Relative Frequency')
xlabel('Class Limits')
grid

%Interpolation
NQP=10; %Number of query points
xq=LCL(1):(HCL(end)-LCL(1))/NQP:HCL(end); %x Location of the query points
yq=interp1([LCL(1) HCL],[0 RCF3], xq,'spline'); %Less Than Ogive
plot(xq,yq,'r')
yq2=interp1([LCL HCL(end)],[ICRF3 0],xq,'Spline'); %More Than Ogive
plot(xq,yq2,'b')

```