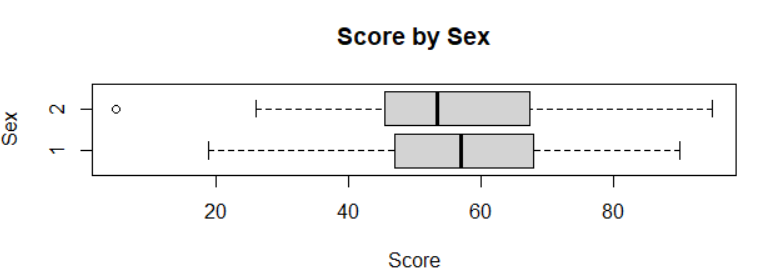
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Emran Sabbagh** | **ID** | **20088634** | **Programme** | **Computer Forensics and Sec** |

**Q1: Do females out-perform males in the final exam?**

To answer this question, first we have to make a summary of the mean, SD and size of the data.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Mean** | **SD** | **Size** |
| **1** | 56.2 | 15.7 | 59 |
| **2** | 55.1 | 17.9 | 59 |

Also we must represent the data in a plot. An appropriate plot at this situation is the **boxplot**.



From the plot above, we can see that the first and third quartiles for males are higher than the corresponding quartiles of females. We can also see that minimum and maximum for females is higher than the corresponding minimum and maximum of males.

An appropriate test for those two variables is the **independent sample t test**.

After performing the test, we get a p-value of **0.7077** which is higher than 0.05.

This means that we don’t need to check the assumption of normality and we don’t reject the null hypothesis.

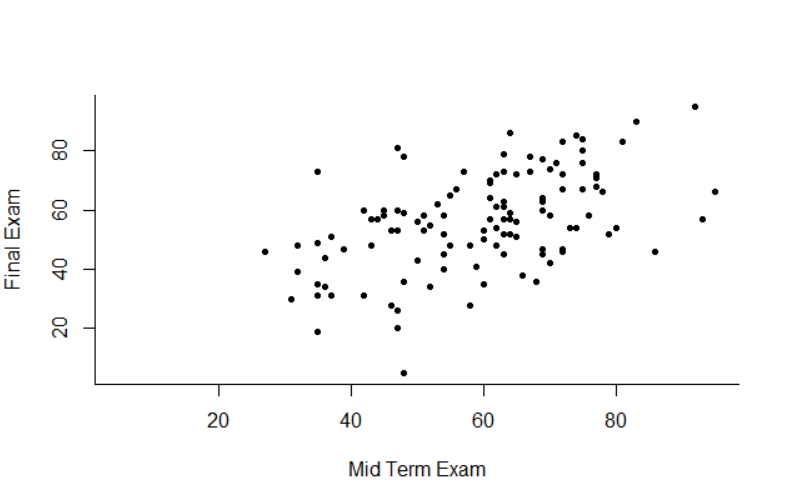
This also means that females don’t score more than males in the final exams just like what the plot indicates.

**Q2: On average, do students improve in the final exam compared to the mid-term exam?**

To answer this question we need first to summarize the data.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** | **NA's** |
| **Final Exam Score** | 5 | 46.5 | 56 | 55.89 | 67 | 95 | 7 |
| **Mid Term Score** | 5 | 48 | 61 | 58.78 | 69 | 95 | 2 |

For the following situation, we use a **scatter plot** to represent the data.



From the scatter plot we can see that there is a relation between the two variables. The relation is moderate, positive and linear. We can clearly see that the points go upwards from bottom left to top right. There are also some differences, for example, we can see that in Mid Term Exams, there were no students who scored less than 20, but in final exam, there were some students who scored below 20. However, that data in the plot are not enough to make an observation and we must perform an appropriate test first.

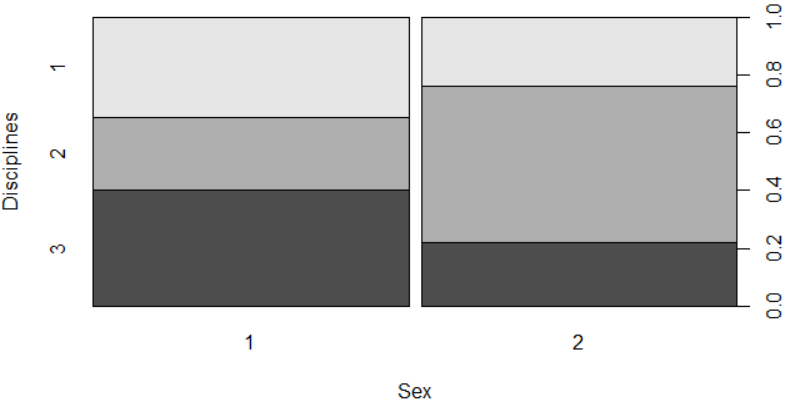
An appropriate test to perform in this situation is the **paired t-test**. When we perform that we end up with a p-value of **0.02** which is less than 0.05. This means that we must do the **Hezne Zirkler test** to check the assumption of joint normality. The p-value we get when we perform that test is **0.34** which is greater than 0.05 so we can say that the assumption of joint normality is good. This means that we reject the null hypothesis. Therefore, we say that students do improve in the final exam compared to the mid-term exam.

**Q3: Is the male-female ratio the same for all three disciplines?**

To answer this question we first summarize the data in a neat table as below.

|  |  |  |
| --- | --- | --- |
|  | **1** | **2** |
| **1** | 19 | 13 |
| **2** | 14 | 30 |
| **3** | 22 | 12 |

After that we need to visualize the data with an appropriate plot. So we use a **bar chart** for that.



From this plot we can see that there are more males in 3 and 1, while there are more females in 2.

An appropriate test in this case is the **Chi-Square test**.

After performing the Chi-Square test we end up with a p-value of **0.007** which is less than 0.05 and thus we reject the null hypothesis.

Because the Chi-Square is a non-parametric test we don’t need to check the assumption of normality in this case.

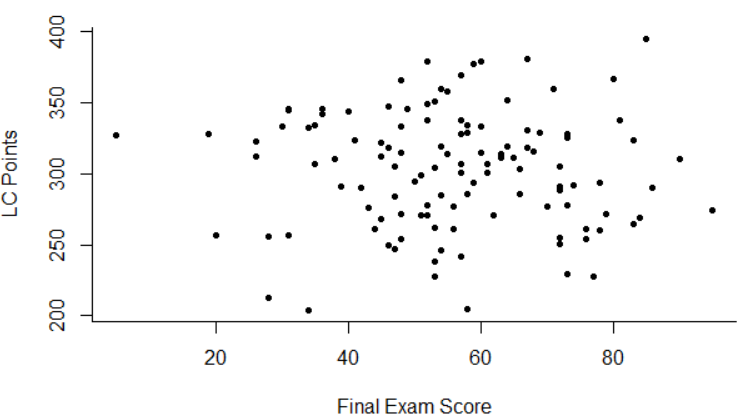
This means that the male female ratio is the same for all three disciplines.

**Q4: Do leaving certificate points give any indication of performance in the final exam?**

To answer this question we first need to summarize data.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** | **NA's** |
| **LC Points** | 204 | 271.8 | 307 | 302.8 | 332.2 | 395 | 2 |
| **Final Exam Score** | 5 | 46.5 | 56 | 55.89 | 67 | 95 | 7 |

Then we need to visualize data in an appropriate plot. In this case we will use the **scatter plot**.



As we can see in the graph, there is no relation at all between the two variables. We can see that in the locations of the points on the x and y axis. The points are spread everywhere in the graph, there are not gathered and they don’t go in a straight line.

We also need to perform an appropriate test on those variables. An appropriate test in this case is the **Correlation test**.

After doing the test, we end up with **0.76** which is greater than 0.05.

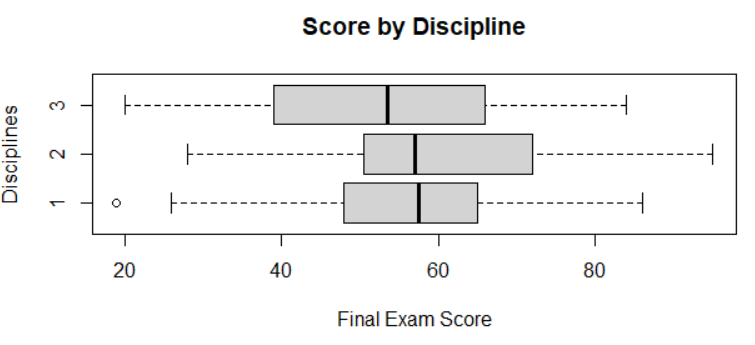
This means that we don’t reject the null hypothesis, which means that leaving cert points give no indication of performance in the final exam just like the graph shows.

**Q5: Do all disciplines perform about equally well in the final exam?**

To answer this question we need to summarize the data first.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Mean** | **SD** | **Size** |
| **1** | 55.76 | 15.79 | 35 |
| **2** | 59.95 | 15.45 | 47 |
| **3** | 52.29 | 16.35 | 36 |

We also need to visualize data in an appropriate plot. In this case we will use the **box plot**.



As we can see in the plot, discipline 2 has the highest minimum, maximum, first and third quartiles, however the median is slightly lower than discipline 1’s median. However, discipline 3 has the lowest minimum, maximum, first quartile and median but the third quartile is slightly greater than discipline 1’s. For discipline 1, the median is very close to but greater than discipline 2’s median, but it has minimum, maximum, first and third quartiles less than the corresponding ones in 2. This indicates that not all disciplines perform about equally well in the final exam, but this is not enough for us to make the observation.

We also need to perform an appropriate test on those variables. An appropriate test in this case is the **Anova test**.

After doing the test, we end up with a p-value of **0.107** which is greater than 0.05.

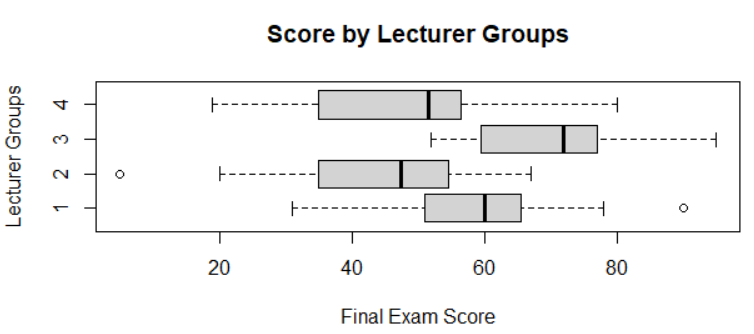
This means that we don’t reject the null hypothesis, which means that not all disciplines perform about equally well.

**Q6: Is there evidence that performance in the final exam varies by lecture group?**

To answer this question we need to summarize the data first.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Mean** | **SD** | **Size** |
| **1** | 58.58065 | 12.27131 | 35 |
| **2** | 44.125 | 13.37969 | 33 |
| **3** | 69.625 | 11.19548 | 32 |
| **4** | 48.6 | 16.74389 | 22 |

We also need to visualize data in an appropriate plot. In this case we will use the **box plot**.



As we can see in the plot, the variety can be noticed in group 3 which has the highest first, third quartiles and median. We can also see that in group 1 which has first, third quartiles and a median that varies from all the other corresponding values in the other groups. However, we can see that first and third quartiles in group 2 and 4 are very close. All the maximum and minimum values are different except for groups 2 and 4, which have close minimums. This indicates that performance varies depending on the lecturer groups, but this is not enough for us to make the observation.

We also need to perform an appropriate test on those variables. An appropriate test in this case is the **Anova test**.

After doing the test, we end up with a p-value of **1.78e-11** which is less than 0.05.

In this situation we need to perform a **Shapiro test** to check the assumption of normality to see if exam scores are normally distributed at each group.

When we do the test on group 1 we get p-value of **0.9802**, for group 2 we get 0.1468, for group 3 we get 0.3766, for group 4 we get 0.5328

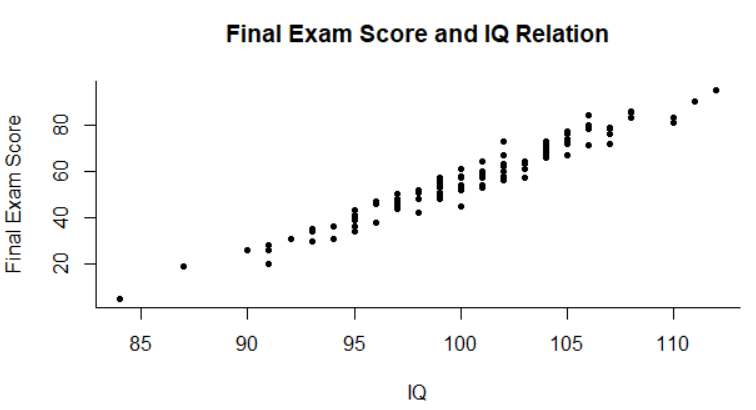
This means that the assumption of normality is good and we reject the null hypothesis, which means that the performance in the final exam varies by lecturer groups.

**Q7: Is final exam score independent of IQ?**

To answer this question we need to summarize the data first.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** | **NA's** |
| **Final Exam Score** | 5 | 46.5 | 56 | 55.89 | 67 | 95 | 7 |
| **IQ** | 84 | 97 | 100 | 100 | 104 | 112 | 7 |

Then we need to visualize data in an appropriate plot. In this case we will use the **scatter plot**.



As we can see in the plot, there is a relation between the two variables. We can see that in the locations of the points on the x and y axis. We can see that the points tend to go upwards to the right in a straight line. This means that the relation between those two variables is strong, positive and linear, but this is not enough for us to make the observation.

We also need to perform an appropriate test on those variables. An appropriate test in this case is the **Correlation test**.

After doing the test, we end up with a p-value of **2.2e-16** which is less than 0.05.

In this situation we need to perform a **Hezne Zirkler test** to check the assumption of joint normality.

When we do the test we get p-value of **0.4253441** which is greater than 0.05.

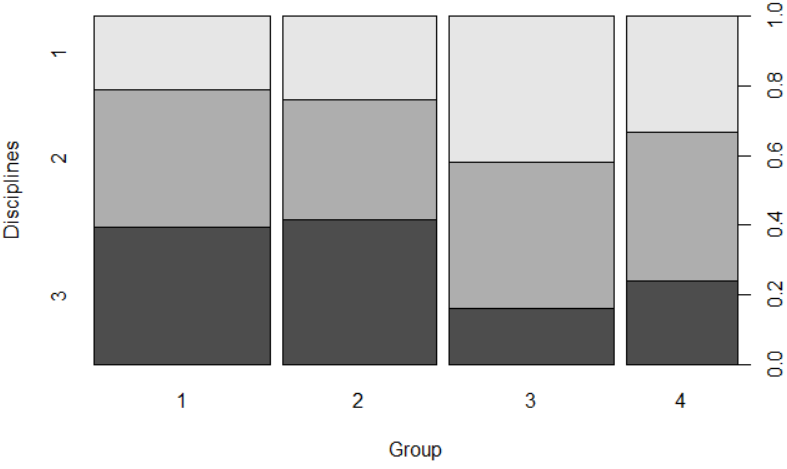
This means that the assumption of joint normality is good and we reject the null hypothesis.

**Q8: Is there reason to doubt that the allocation of students of different disciplines to classes is random?**

To answer this question we first summarize the data in a neat table as below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** |
| **1** | 7 | 7 | 13 | 7 |
| **2** | 13 | 10 | 13 | 9 |
| **3** | 13 | 12 | 5 | 5 |

After that we need to visualize the data with an appropriate plot. So we use a **bar chart** for that.



From this plot we can see that discipline 1 occurs the most in group 3, discipline 2 occurs the most in group 1, 3 and 4 and discipline 3 occurs the most in group 2.

An appropriate test in this case is the **Chi-Square test**.

After performing the Chi-Square test we end up with a p-value of **0.2858** which is more than 0.05 and thus we don’t reject the null hypothesis.

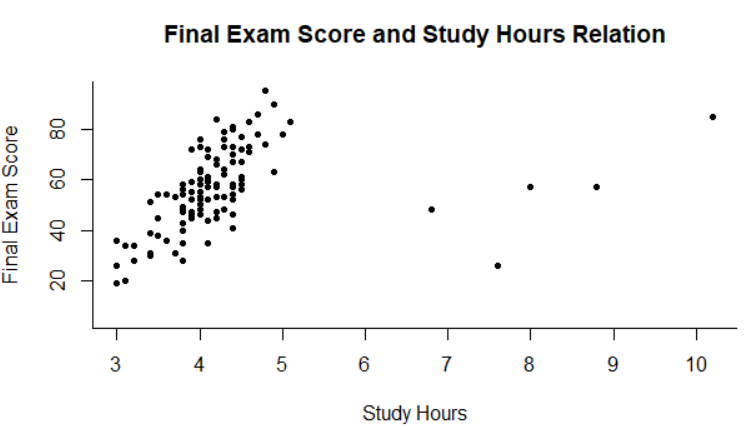
This means that there is no reason to doubt that the allocation of students of different disciplines to classes is random.

**Q9: Is hours of study an important factor in exam score?**

To answer this question we need to summarize the data first.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** | **NA's** |
| **Final Exam Score** | 5 | 46.5 | 56 | 55.89 | 67 | 95 | 7 |
| **Study Hours** | 3 | 3.8 | 4.1 | 4.26 | 4.4 | 10.2 | 8 |

Then we need to visualize data in an appropriate plot. In this case we will use the **scatter plot**.



As we can see in the plot, there is a relation between the two variables. We can see that in the locations of the points on the x and y axis. We can see that the points tend to go upwards to the right in a straight line. However, we can see some spread towards the right with few points that are on their own. This means that the relation between those two variables is moderate, positive and linear, but this is not enough for us to make the observation.

We also need to perform an appropriate test on those variables. An appropriate test in this case is the **Correlation test**.

After doing the test, we end up with a p-value of **0.00012** which is less than 0.05.

In this situation we need to perform a **Hezne Zirkler test** to check the assumption of joint normality.

When we do the test we get p-value of **3.330669e-16** which is less than 0.05.

This means that the assumption of joint normality is not good and we can’t reject the null hypothesis.

############################################################

#

# Statistics Assignment (Stats and Probability)

#

# Please run the commands seperately

#

# R commands

#

setwd("C://Users//USER//Desktop//Statistics CA/")

Emran = read.table("Sabbagh\_Emran\_20088634.csv", sep= ",", header = TRUE)

attach(Emran)

view(Emran)

head(Emran)

names(Emran)

############################################################

# Q1 #######################################################

tapply(examScore, stud.sex, mean, na.rm=T)

tapply(examScore, stud.sex, sd, na.rm=T)

tapply(examScore, stud.sex, length)

boxplot(examScore ~ as.factor(stud.sex), main="Score by Sex", ylab="Sex", xlab="Score", horizontal=T)

t.test(examScore ~ stud.sex)

# 0.7

# Q2 #######################################################

summary(examScore)

summary(midTermScore)

plot(examScore ~ midTermScore, xlab="Mid Term Exam", ylab="Final Exam", pch=20, bty="L")

t.test(examScore, midTermScore, paired=T)

#0.02

tmp = !is.na(examScore) & !is.na(midTermScore)

HZ.test(data.frame(examScore[tmp], midTermScore[tmp]))

#0.34

# Q3 #######################################################

plot(as.factor(Disp) ~ as.factor(stud.sex), xlab="Sex", ylab="Disciplines", pch=20, bty="L")

table(as.factor(Disp), as.factor(stud.sex))

chisq.test(as.factor(Disp), as.factor(stud.sex))

#0.007

# Q4 #######################################################

summary(examScore)

summary(lcPoints)

plot(lcPoints ~ examScore, xlab="Final Exam Score", ylab="LC Points", pch=20, bty="L")

cor.test(lcPoints, examScore)

# 0.7562

# Q5 #######################################################

tapply(examScore, Disp, mean, na.rm=T)

tapply(examScore, Disp, sd, na.rm=T)

tapply(examScore, Disp, length)

boxplot(examScore ~ as.factor(Disp), xlab="Final Exam Score", ylab="Disciplines", main="Score by Discipline", horizontal=T)

summary(aov(examScore ~ as.factor(Disp)))

# 0.107

# Q6 #######################################################

tapply(examScore, stud.group, mean, na.rm=T)

tapply(examScore, stud.group, sd, na.rm=T)

tapply(examScore, stud.group, length)

boxplot(examScore ~ as.factor(stud.group), xlab="Final Exam Score", ylab="Lecturer Groups", main="Score by Lecturer Groups", horizontal=T)

summary(aov(examScore ~ as.factor(stud.group)))

shapiro.test(examScore[stud.group==1])

# 0.9802

shapiro.test(examScore[stud.group==2])

# 0.1468

shapiro.test(examScore[stud.group==3])

# 0.3766

shapiro.test(examScore[stud.group==4])

# 0.5328

# Q7 #######################################################

summary(examScore)

summary(studIQ)

plot(examScore ~ studIQ, main="Final Exam Score and IQ Relation", xlab="IQ", ylab="Final Exam Score", pch=20, bty="L")

cor.test(examScore, studIQ)

# 2.2e-16

tmp = !is.na(examScore) & !is.na(studIQ)

HZ.test(data.frame(examScore[tmp], studIQ[tmp]))

# 0.4253441

# Q8 #######################################################

plot(as.factor(Disp) ~ as.factor(stud.group), xlab="Group", ylab="Disciplines", pch=20, bty="L")

table(as.factor(Disp), as.factor(stud.group))

chisq.test(as.factor(Disp), as.factor(stud.group))

#0.2858

# Q9 #######################################################

plot(examScore ~ studyHours, main="Final Exam Score and Study Hours Relation", xlab="Study Hours", ylab="Final Exam Score", pch=20, bty="L")

cor.test(examScore, studyHours)

# 0.00012

tmp = !is.na(examScore) & !is.na(studyHours)

HZ.test(data.frame(examScore[tmp], studyHours[tmp]))

# 0.4253441