**Scenario**

An analysis is carried out on the final exam mark attained by first year, college students in a statistics examination. Information is collected on each student includes their intelligence quotient (is), the number of points they attained on their leaving certificate, their schoolor discipline (computing, science engineering), their sex and the midterm mark they obtained in an assessment in the same module. Students are randomly assigned to one of four lecture groups, which are taught by different lecturers.

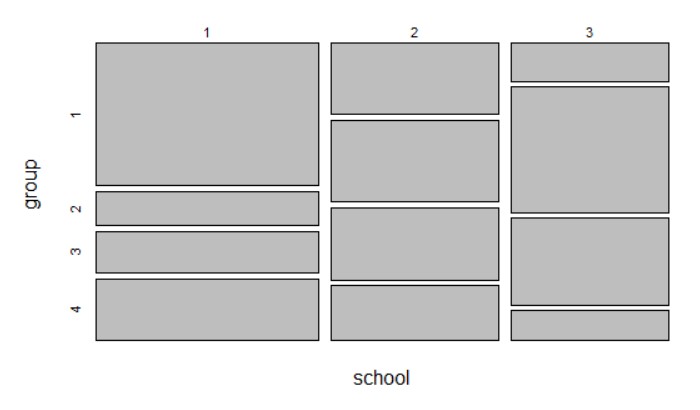
Eight questions are to be considered in relation to this data:

1. Is there reason to doubt that the allocation of students of different disciplines to groups is random?
2. Is final exam mark independent of IQ?
3. Is the male-female ratio the same for all three disciplines?
4. On average, do students improve in the final exam compared to the mid-term exam?
5. Do leaving certificate points give any indication of performance in the final exam?
6. Is there evidence that performance in the final exam varies by lecture group?
7. Do all disciplines perform about equally well in the final exam?
8. Do females out-perform males in the final exam?

**Solutions**

**Q1 -** Is there reason to doubt that the allocation of students of different disciplines to groups is random?

A 100% stacked bar chart is shown below comparing the “group” and “school” categorical data.



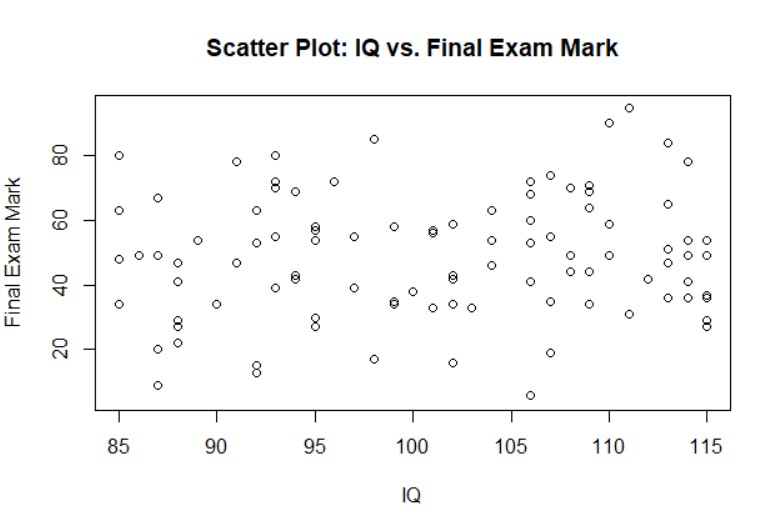
This plot does indicate that there may be reason to doubt there is no correlation between a student’s school and their allocated group as it can be seen that group 1 takes up a majority of school 1, and school 3 has a large majority of group 2 and 3 students. If it were a random allocated, the bars would not be so heavily weighted.

A suitable statistical test to determine if there is reason to doubt that the allocation of students of different disciplines to groups is random is the chi-squared test, because both sets of data are categorical.

The results from performing this test on these two sets of data yielded a p value of 0.005497964. As p is less than 0.05, it is suggested that there is reason to doubt that the allocation of students of different schools to their groups is random, in fact due to the very low p value it is likely that there is a significant correlation between the student’s school and the group they are allocated.

**Q2 -** Is final exam mark independent of IQ?

This Scatter Plot below portrays the correlation between IQ and Final Mark.



As seen in the scatter plot, there does not seem to be much correlation between IQ and the final mark, as the dots on the scatter plot seem to be spread out across the plot, however there may be a slight correlation as there are less poor results towards the higher end of “IQ”.

A Pearson’s Correlation Test was carried out using R to determine if the mark was independent of IQ, as both variables are numerical and continuous.

The test determined that the correlation coefficient was 0.1067172. This implies a weak correlation between student’s IQ and their final exam mark, which is in line with what was observed from the scatter plot graph.

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

The bar plot below displays the male-female ratio for each school.

Chart, bar chart

Description automatically generated

This bar plot clearly shows that the male-female ratio for the three disciplines/schools are not the same.

Using R we can calculate the exact ratios for each school. For school 1, there are approximately 46.5% males and 46.5% females. For school 2, there are approximately 54.8% males and 38.7% females, and for school 3 there are approximately 60% males and 36.7% females.

This clearly shows that the male-female ratios are not the same for any of the schools when compared to each other. School 1 has a 1:1 ratio of male to female, school 2 a 1.41:1 ratio and school 3 a 1.25:1 ratio.

**Q4 –** On average, do students improve in the final exam compared to the mid-term exam?

This clustered dot plot shows that the majority of students either did worse in their final exam compared to their midterm or did not have a significant change to their mark. The number of students who did significantly better in their final exam mark is small.Chart, scatter chart

Description automatically generated

A paired t-test is the best way to analyse this data as we are compared past test results to new test results. The paired t-test revealed that the mean difference between the results was -4.83. This means that on average the student’s marks were lower by approximately 4.83 marks. A 95% confidence interval estimate for marks lost is given by [-6.89, -2.76]. As this excludes zero, we can conclude there has been a significant change between the results.

**Q5 –** Do leaving certificate points give any indication of performance in the final exam?

The Scatter Plot below portrays the correlation between Leaving Certificate Points and the Final Mark. Chart, scatter chart

Description automatically generated

From the Scatter Plot, it can be interpreted that there is a correlation between Leaving Certificate Points and the final mark. There is a clear upward trend indicating that the higher the Leaving Cert points, the better the student did on the exam.

A Pearson’s Correlation Test was carried out using R to determine if the L.C. Points gave an indication of performance on the exam, as both variables are numerical and continuous.

The test determined that the correlation coefficient was 0.8667911. This implies a very strong positive correlation between student’s Leaving Certificate points and their final exam mark, which is in line with what was observed from the scatter plot graph. The correlation coefficient being close to 1 tells us that higher LC points correlate heavily to better exam performance, as we observed from the graph.

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

This boxplot below is a visualization of each disciplines performance in relation to their final exam mark using an ANOVA test.Chart, box and whisker chart

Description automatically generated

We use an ANOVA test for this as we are dealing with 4 different groups of data and comparing them. This graph indicates that performance in the exam does tend to vary between each group, for example we can clearly see that group 3 has a very high performance in the exam compared to the other groups, and that the other three groups are more varied but tend to average between 30 - 60 marks. The p-value received from the ANOVA test when run in R is 1.3e-06, which is very small. This indicates a large variances between the mean marks between the lecture groups.

**Q7 –** Do all disciplines perform about equally well in the final exam?

This boxplot shows us the averages of final exam marks in relation to the different schools or disciplines.

Chart, box and whisker chart

Description automatically generated

We can interpret from this plot that there is not a huge variance between the final marks between the schools, which is supported by the p-value of 0.874 produced by an ANOVA test ran in R. Because the p-value is greater than 0.05, we know that there is no significant difference between the performance of different disciplines in the exam.

**Q8 –** Do females out-perform males in the final exam?

This boxplot, where the top row is female performance in exams and the bottom is males, shows that there is not much variance between the sexes final exam marks.

Chart, box and whisker chart

Description automatically generated

A t-test was carried out in R as these are two numerical data samples. The p-value returned was 0.1268 and the corresponding t-value was 1.534. The p-value is greater than the significance level of 0.05 and thus we can determine that there is not a significant difference between the performance of males and females in the exam.

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

# Statistics Assignment (Stats and Probability)

#

# Liam Flynn / ID: 20098690 / Programme: Applied Computing / Game Dev.

#

# R commands to address eight questions pertaining to the

# exam results of leaving-cert students.

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

# Load the data from the clipboard

setwd("C:/Users/Liam/OneDrive - Waterford Institute of Technology/Statistics & Probability")

students = read.csv("54 Liam Flynn 20098690.csv")

attach(students)

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

#

# Q1 -Chi Squared Test#######################################################

#

# Create Chart

plot(table(school, group))

#calculate p value

q1\_table = table(students$school, students$group)

q1\_result = chisq.test(q1\_table)

q1\_result$p.value

#

# Q2 - Correlation Coefficent#######################################################

#

# Create a scatter plot of IQ vs. final exam mark

plot(iq, mark, xlab = "IQ", ylab = "Final Exam Mark", main = "Scatter Plot: IQ vs. Final Exam Mark")

#calculate

q2\_cor\_coef = cor(students$mark,students$iq,use = "complete.obs")

q2\_cor\_coef

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

#

# Q3 - Calculate & Compare Ratios#######################################################

#

# Create subsets for each discipline

school1 = subset(students, school == "1")

school2 = subset(students, school == "2")

school3 = subset(students, school == "3")

#calculate ratios

ratio1 = table(school1$sex) / nrow(school1)

ratio2 = table(school2$sex) / nrow(school2)

ratio3 = table(school3$sex) / nrow(school3)

#display ratios

ratio1

ratio2

ratio3

#create barchart

# Create a vector of ratios

ratios <- c(ratio1["1"], ratio2["1"], ratio3["1"], ratio1["2"], ratio2["2"], ratio3["2"])

# Specify the labels for the bars

bar\_labels <- c("School 1 (Male)", "School 2 (Male)", "School 3 (Male)",

"School 1 (Female)", "School 2 (Female)", "School 3 (Female)")

# Create the bar plot

barplot(ratios, names.arg = bar\_labels, xlab = "Schools", ylab = "Ratio",

main = "Male-Female Ratio between Schools")

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

# Q4 - paired ttest#######################################################

#

#Carry out paired t-test

q4 = t.test(mark,midterm,paired=TRUE)

q4

#make dot plot

change = mark - midterm

dotchart(change, labels = students$id, groups = students$discipline,

xlab = "Change in Results",color="black",pch=16, ylab = "Student", main = "Change from Midterm to Final Exam")

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

# Q5 - Correlation Coefficient#######################################################

#

# Create a scatter plot of IQ vs. final exam mark

plot(points, mark, xlab = "LC Points", ylab = "Final Exam Mark", main = "Scatter Plot: Leaving Certificate Points vs. Final Exam Mark")

#calculate

q5\_cor\_coef = cor(students$mark,students$points,use = "complete.obs")

q5\_cor\_coef

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

# Q6 - ANOVA#######################################################

#

#ANOVA test

q6 = aov(mark ~ group, data = students)

summary(q6)

#plot

boxplot(mark ~ group, data = students,

xlab = "Lecture Group", ylab = "Final Exam Mark",

main = "Boxplot: Final Exam Mark/Lecture Group Comparison")

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

# Q7 - ANOVA#######################################################

#

# ANOVA test

q7 <- aov(mark ~ school, data = students)

summary(q7)

#plot

boxplot(mark ~ school, data = students,

xlab = "School", ylab = "Final Exam Mark",

main = "Boxplot: Final Exam Mark/Discipline Comparison")

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

# Q8 - t-test#######################################################

#

#Subsets for m and f

f = subset(students, sex == "2")

m = subset(students, sex == "1")

#t-test

q8 = t.test(f$mark, m$mark)

# Print the t-test result

print(q8)

boxplot(f$mark, m$mark, xlab = "Sex", ylab = "Final Exam Mark",

main = "Boxplot: Final Exam Mark by Sex", horizontal = TRUE)