MA4413 Assignment **(20% of Module)**

Overview

- This assignment must be the work of your group (max. 4 students). Evidence of copying will result in a score of zero.
- **All** group members must contribute and this will be assessed through a random selection of groups being interviewed in Week 13. If any group member does not understand their project, the whole group will get zero.
- Please state your group members before **Wednesday, November 16th** by way of **one** group member emailing me (`kevin.burke@ul.ie`) the list of relevant names and student IDs.
- Project deadline: **Friday, November 25th (Week 12) before 4pm.**
  - Hard copy of full report to be submitted to the Maths Department office (room D2034) or in lecture before deadline.
  - Electronic copy of report and accompanying `R` script file containing all code to be submitted via the "Assignments" section on Sulis.
  - Only submit **one assignment per group**, i.e., only one student from the group submits the project Sulis; **make sure all group members are named on the report**.
- The report is a typed document (*not* your `R` script file!) saved as a pdf file. The main font may be arial, calibri or times new roman. However, the font for any included `R` output must be `courier`. Font size is 11pt.
- The accompanying `R` script file must be organised so that code can exactly replicate all results in the order in which they appear in your report.

Project **(90 marks)**

Important: in your written report and R script, present your work in the order shown below.
Note: "independent research" means the content was not covered in lectures. You must independently discover how to carry out these procedures in `R` and correctly interpret the results. Google is your friend here obviously!

1) **Analysis of Midterm Scores** **(60 marks)**
The (fictional) dataset "`midscores.csv`" contains midterm scores for 100 students with the following headings: `score` (the midterm score), `submit` (whether or not the student submitted "`early`" – on days 0-5 – or "`late`" – on days 6-7) and `day` (the day of submission for this student).

   You will analyse a random sample of data from `midscores.csv` (as we will pretend that this full class list is unavailable to you - as would usually be the case in practice).

   Based on your analysis, you will decide whether or not a significant different between early and late midterm submissions exists.

i) **Loading Data into `R` and Drawing Random Samples**

- Save the `midscores.csv` to your computer.
- Load the dataset into `R` using the following code:

```
setwd("C:\\Users\\John Smith\\Documents")
midscores = read.csv("midscores.csv", header=T)
```

  where you will replace `C:\\Users\\John Smith\\Documents` with the location of the `midscores.csv` file on your computer.
- In the `midscores` file, the first 50 students are early and the second 50 are late. Draw a random sample of 10 early submissions and 10 late submissions as follows:

```
set.seed(123456789)
rows = c(sample(1:50, 10), sample(51:100, 10))
midsample = midscores[rows,]
```

  where you will replace `123456789` with one of the student ID numbers of your group members. This `seed` will allow me to reproduce your results from your `R` script. **If I cannot reproduce your results, you will be heavily penalised**.

  Important: You will analyse data from `midsample` and **not** the full dataset (as we are pretending you do not have the full dataset).
- Separate the groups of midterm scores for analysis as follows:

```
midearly = midsample$score[midsample$submit=="early"]
midlate = midsample$score[midsample$submit=="late"]
```

ii) **Graphical and Numerical Summaries** (Lecs 1 and 2) **(10 marks)**

- Plot histograms for both groups.
- Plot the boxplots for both groups on the same graphical window.
- Calculate the mean, standard deviation, quartiles, IQR, minimum and maximum midterm score for both groups. In your report, present these summaries in a table with two columns (one for each group, rounding all numbers to two decimal places).
- Comment on all of the above output with reference to the shape of the distributions, centre and spread etc. **This must be concise and to the point.**

iii) **Check for Normality of Data** (Lec 10) **(10 marks)**

- Use Q-Q plots to determine whether or not the two data vectors are approximately normally distributed (also refer back to the histograms and boxplots).
- *(independent research)* There also exists a hypothesis test of normality called the Shapiro-Wilk test. Carry out this test (formally state hypotheses) in `R` and interpret the results.

iv) **Confidence Intervals and Hypothesis Testing** (Lecs 13,14,15,16) **(20 marks)**

While the summaries from part (ii) above are useful for describing a *sample* of data, we require confidence intervals and hypothesis tests to make make statements about the *whole population.*

For all of the hypothesis tests in this section you must do the following: copy the `R` output into your report using `courier` font, clearly state (mathematically) the null and alternative hypotheses and provide your conclusion based on the p-values and confidence intervals in both statistical and non-statistical language.

- Test the hypothesis that the overall mean midterm score is equal to 50, i.e., here you are not splitting up the two groups for this.
- Test the hypothesis that there is no difference between the variances in each group.
- Test the hypothesis that there is no difference between the means in each group. Note: decide whether or not equal variances may be assumed using the result of the previous hypothesis test.
- *(independent research)* When the samples are small, the t-test requires the assumption that data is normally distributed. If this is not reasonable, "non-parametric" methods can be used. For comparing two independent groups, the Wilcoxon rank-sum test (also known as the Mann-Whitney U test) exists. Carry out this test in R to compare the midterm scores in the two groups. Note: do not confuse with the Wilcoxon signed-rank test for paired data.

v) ***(independent research)* Simple Linear Regression** **(15 marks)**

In the above, you have analysed the effect of the time of submission by way of comparing `early` (day 0 - 5) and `late` (day 6 - 7) submissions, i.e., we have split the *numeric* variable `score` by the *categorical* variable `submit`.

However, we have the numeric variable "`day`" in the dataset too. So we should use this information rather than artificially grouping time as "`early`" and "`late`". Specifically, we can explore the *relationship* between the two *numeric* variables `score` and `day`.

Note: you are still using the sample of data, `midsample`, here.

- Plot the `score` against the `day` and comment briefly on the relationship between the variables (if any).
- Calculate the *correlation coefficient* using `R` and comment on its magnitude and sign.
- Fit the *linear regression model* using the `lm` function in `R`. Write down the equation of the regression line and interpret the regression coefficient of `day` (i.e., the *slope* of the line). Interpret the p-value associated to this coefficient (stating hypotheses formally and conclusion). Note: all of this information is contained in the output of `summary(model)` where "`model`" is your fitted regression model.
- Comment on how well the regression line fits the data by overlaying it on top of the plotted data (i.e., how close is the line to the data points?) and also refer to the R^2 value (which is contained in the `summary(model)` output).

vi) **Brief Summary of Analysis** **(5 marks)**

- Briefly summarise the main results of the above analysis analysis, i.e., parts (ii) - (v). Also, provide your final conclusion in non-statistical language. Be clear and concise. A few key sentences is sufficient - **no more than half a page**.

2) **Simulation Study (Central Limit Theorem)** (Lec 12) **(10 marks)**

We have seen in a simulation study at the end of Lecture 12, for exponential data, that the sample mean, $\bar{x}$, is approximately normally distributed when the sample size is large.

- You are required to carry out a similar simulation study but for Bernoulli data, i.e., to show that the sample proportion, $\hat{p}$, is also approximately normally distributed when the sample size is large (say $n = 50$) and the true proportion is $p = 0.5$. Note that Bernoulli data can be generated using `rbinom(n=50, size=1, prob=0.5)`.
- Carry out another study where $p = 0.05$ but keep the sample size at $n = 50$. You should find that $\hat{p}$ is *not* approximately normal in this case.

Note: try varying the sample size to investigate how large it needs to be so that $\hat{p}$ is approximately normally distributed.

For both scenarios (i.e., $p = 0.5$ and $p = 0.05$) you have to set the "simulation seed" **so that I can reproduce your results exactly**. Use `set.seed(123456789)` where you will replace `123456789` with one of the student ID numbers of your group members.

3) **Probability Questions** (Lecs 7,8,9,10) **(10 marks)**

Answer the questions below using `R` functions `dbinom`, `dpois` and `pnorm`:

Note: Round your answers to **four decimal places**:

i) $\Pr(X \geq 6)$ where $X \sim \text{Binomial}(n = 10,\ p = 0.65)$

ii) $\Pr(X < 30)$ where $X \sim \text{Binomial}(n = 100,\ p = 0.2)$

iii) $\Pr(15 \leq X \leq 30)$ where $X \sim \text{Binomial}(n = 50,\ p = 0.32)$

iv) $\Pr(X = 8)$ where $X \sim \text{Poisson}(\lambda = 6)$

v) $\Pr(X > 35)$ where $X \sim \text{Poisson}(\lambda = 41)$

vi) $\Pr(2 \leq X \leq 5)$ where $X \sim \text{Poisson}(\lambda = 1)$

vii) $\Pr(X > 12)$ where $X \sim N(\mu = 7,\ \sigma = 2.5)$

viii) $\Pr(X > 9.8)$ where $X \sim N(\mu = 10,\ \sigma = 1)$

ix) $\Pr(X < 38)$ where $X \sim N(\mu = 50,\ \sigma = 5)$

x) $\Pr(4 < X < 8)$ where $X \sim N(\mu = 5,\ \sigma = 3.6)$

4) **Report Layout** **(10 marks)**

- There are marks going for overall layout/presentation of the report which must take the form of a professional document.