

MA4413 Assignment

Project (complete in the order shown below)

1) Analysis of Midterm Scores

The (fictional) dataset, `midscores.csv`, contains midterm scores for students with high Sulis activity and students with low Sulis activity. You have been assigned the task of analysing the `midscores.csv` data. Based on the analysis, you will decide whether or not there is a difference between midterm scores of students with high / low Sulis activity.

i) Loading Data 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.

- Draw a random sample of 20 individuals with a high midterm score and 20 individuals with a low midterm score as follows:

```
set.seed(123456789)
midhigh = sample(midscores$high, size=20)
midlow = sample(midscores$low , size=20)
```

where you will replace `123456789` with one of the student ID numbers of your group members.

Important: You will analyse the random samples `midhigh` and `midlow`, **not** the full dataset.

ii) Graphical and Numerical Summaries

- Plot histograms for both groups.
- Plot the boxplots for both groups on the same graph.
- 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.

iii) Check for Normality of Data

- 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).

iv) Confidence Intervals and Hypothesis Testing

While the summaries from part (ii) above are useful for describing a *sample* of data, we cannot make statements about the *whole population* without constructing confidence intervals and performing hypothesis tests.

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 average midterm score is equal to 7.5. Note: this is a one-sample test, i.e., you must combine the two samples into one vector here.
- 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: your decision on whether or not equal variances can be assumed must be based on the previous hypothesis test.

v) Brief Summary of Analysis

- Briefly summarise the main results of your analysis. 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)

- 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. Note that Bernoulli data can be generated using `rbinom(n=50, size=1, prob=0.5)`, i.e., the sample size is $n = 50$ and the true proportion is $p = 0.5$.
- 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: you can try larger sample sizes to investigate how

large this 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”. Use `set.seed(123456789)` where you will replace 123456789 with your student ID

3) Probability Questions

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 $X \sim N(\mu = 5, \sigma = 3.6)$