STAT2401 Analysis of Experiments
Semester 1, 2025
Assignment 1 (15%, 50 marks)

Due date: Week 7 Friday 11th of April 2025 by 11:59pm

- Working in **pairs** of two students is strongly encouraged. If you work as a pair:
  - submit only one assignment per team;
  - each student must contribute towards ALL questions;
  - list the team members by name, student ID and **state what each team member has contributed to the assignment before answering the questions. Your assignment will not be marked without this statement.**
- You must submit **two files (the main document and Rmarkdown)** using **two separate LMS submission button**:
  1. The main document consists of your answers is to be submitted to the first **Turnitin LMS** button as a **pdf** file only. Unlimited versions can be submitted until the due date and must comply to the following requirements:
     - You are only required to submit your submission **in PDF** with the following **approximate layout and order**:
       Page 1-2 Question 1 (a)-(d)
       Page 2-3 Question 2 (a)-(c)
       Page 4-10 Question 3 (a)-(g)
     - Use font size **12** for answer or comment.
     - Use smaller font size of (eg 9) for R-code snippets, figures or `R`-outputs.
     - Your submission main document should have no more than **10** pages.
     - Do not include any cover sheet, cover page or title page since that should be counted as 1 page.
     - Do not include the scenario of the questions in your submission. See the Rmd template.
     - Resize the plots accordingly to fit the space or combine the plots whenever applicable, using par(mfrow) command.
     - **Your submission will not be marked if you don't follow the layout and order**.
     - Only electronic submission through LMS is acceptable.
  2. Rmarkdown (no need to be knitted) attachment that contains the complete R code as your working is to be submitted into the second LMS button within File Response.
- **Marking of late assignments will follow the university rules.**

Assignment Questions

The data required for this assignment can be found within the Assignment 1 folder on the LMS.

1. **Normal distribution. (8 marks)** Social networking sites such as Facebook and Instagram have grown in popularity. The website Hitwise reported that the mean time spent by a user at Facebook during April 2020 was 25 minutes.

   Suppose the distribution of time spent at Facebook per month is normally distributed, with a mean $\mu = 25$ minutes and the standard deviation $\sigma = 4.0$ minutes.

   If a Facebook user is selected at random:

   (a) (1 mark) Find the probability that the user spends more than 15 minutes per month at the site.

   (b) (2 marks) Find the probability that the user spends between 20 and 35 minutes per month at the site.

   (c) (2 marks) What is the amount of time per month a user spends on Facebook, if only 1% of users spend this time or longer on Facebook?

   (d) (3 marks) Between what values do the time spent of the middle 90% distribution of Facebook users fall?

   Hint. A snippet of R code should be provided as your working for (a)-(d). Your final answers must be presented in four decimal points x.xxxx.

2. **Blood fat concentration (11 marks)**

   Data were collected on the concentration of plasma cholesterol in mg/dl for 371 male patients evaluated for chest pain. For 51 patients there was no evidence of heart disease; for the remaining 320 there was evidence of narrowing of the arteries.

   Let y represent the concentration of plasma cholesterol. The summary statistics for each group as shown below :

	No disease	Disease
sample mean plasma cholesterol, $\bar{y}$	195.2745mg/dl	216.1906mg/dl
sample variance, s^2	1303.9231mg/dl	1850.2488mg/dl
sample size, n	51	320

   Assume unequal variances using the exact method to answer the following questions.

   (a) (6 marks) Conduct a two-independent sample t-test using R to determine whether the concentration of plasma cholesterol is significantly different between patients with no evidence of heart disease and those with narrowing of the arteries.

   (b) (3 marks) Determine a 99% confidence interval for the mean difference in concentration of plasma cholesterol between the two groups of patients.

   (c) (2 marks) Explain the correspondence between the confidence interval in (b) and a test of the hypotheses you listed in question (a).

   Hint. A snippet of R functions and the results must be provided as your working for (a)-(b). Use a significance level of 1%.

3. **Regression (31 marks)**

   A researcher considers an example involving the management at a Canadian port on the Great Lakes to estimate the relationship between the volume of a ship's

cargo and the time required to load and unload this cargo. It is envisaged that this relationship will be used for planning purposes as well as for making comparisons with the productivity of other ports.

Records of the tonnage loaded and unloaded as well as the time spent in port by 31 liquid-carrying vessels that used the port over the most recent summer are available. The data are available on the book website in the file **glakes.txt.**

Two models are proposed:

$$M_1 : \text{Time} = \beta_0 + \beta_1 \text{Tonnage} + \epsilon$$
$$M_2 : \log(\text{Time}) = \beta_0 + \beta_1 \text{Tonnage}^{0.5} + \epsilon$$

where log is to the base e or natural logarithms.

(a) (2 marks) Fit a simple linear model M_1 to these data. Present the appropriate scatterplot and plot the fitted line onto the scatterplot. Comment about the output in a few concise sentences.

Hint. Your answer must include 1 plot and comment only. Do not include the R code.

(b) (5 marks) Provide the model summary and diagnostics checking plots for model M_1. Does the straight line regression model M_1 seem to fit the data well? Comment about the output in a few concise sentences.

Hint. Use 'plot(file.lm)' and 'par(mfrow=c(2,2))' for diagnostics plots. Your answer must include the model summary, the plots and comment. Do not include the R code.

(c) (5 marks) Do you think there are outliers or influential points in the data? What influence do these points have on the model fit? Use leverage and Cook's distance for this investigation.

Hint. Your answer must include a snippet of R code, the results, 2 plots and comment. Use the interval of (-2,2) for standardised residuals.

(d) (4 marks) Fit a regression model to the transformed M_2 model. Present the appropriate scatterplot and plot the fitted line onto the scatterplot. Does the transformed line regression model M_2 seem to fit the data well? Comment about the output in a few concise sentences.

Hint. Your answer must include a snippet of R code for transformation, 1 plot and comment only.

(e) (5 marks) Provide the model summary and diagnostics checking plots for model M_2. Does the straight line regression model M_2 seem to fit the data well? Comment about the output in a few concise sentences.

Hint. Use 'plot(file.lm)' and 'par(mfrow=c(2,2))' for diagnostics plots. Your answer must include the model summary, the plots and comment. Do not include the R code.

(f) (4 marks) Perform a hypothesis testing for a positive slope at a significance level of 5% based on model M_2.

Hint. Do not include the R code.

(g) (6 marks) Compare a 95% confidence interval of the mean response and a 95% prediction interval for a new value when Tonnage = 10,000 using the untransformed model M_1 and transformed model M_2 respectively. Provide two scatterplots that consist the fitted model, the confidence and prediction

intervals for each of M_1 and M_2 respectively. Comment about the output in a few concise sentences.

Hint. Your answer must include a snippet of R code for intervals, the results, the two plots and comment.