DH 302 Spring 2025 Assignment 02

The assignment is based on Lecture 7, Lecture 8, Lecture 9, and Lecture 10. Due at 11:59PM (IST), Tuesday 18^{th} February, 2025 via Gradescope. Raw template is available here and an upto date version of this pdf is available here. Total points = 150

YOUR NAME HERE (YOUR ROLL NUMBER HERE)
NAME AND ROLL NUMBERS OF YOUR COLLABORATOR(S) WITH Question(
Example: Saket (09D02007): Q1,Q2,Q7

Instructions

Submit your solutions via gradescope by 11:59 PM (IST) Tuesday, 18^{th} February 2025. In-person submissions will not be entertained. Please upload a single PDF file. Late submissions are allowed with a 10% per day penalty (only upto 22nd February). You can raise your questions related to the assignment on Piazza - please tag these as `assignment_02`.

- For theory questions, you can either write your response for in latex or put a screenshot/picture of your handwritten response in the appropriate section. To embed scanned images, use this format: `![question1](/path/to/question1.png)` where `/path/to/question1.png` is the local path (on your laptop) to your scanned (handwritten) response to question1.
- If you are writing the solutions for theory questions by hand please use a pen. Pencil submissions are difficult to read when scanned. You will have to take a scan for each such answer and embed it in this document.
- Your final submission has to be the PDF that comes from this template - one single pdf. No Exceptions.
- Please mention the name(s) of people you collaborated with and what exactly you discussed.

Making your submission: Raw template is available here and an upto date version of this pdf is available here. Open the template in Rstudio (you will need to ensure Quarto is installed). Once you are done with your answers, use the "render" (arrow like) button on the toolbar to create a pdf. Only pdf submissions are allowed.

Problem 01 [25 points]

Quality of life: Improvement in quality of life was measured for a group of heart disease patients after 8 weeks in an exercise program. This experimental group was compared to a control group who were not in the exercise program. Quality of life was measured using a 21-item questionnaire, with scores of 1–5 on each item. The improvement data are as follows and are plotted below.

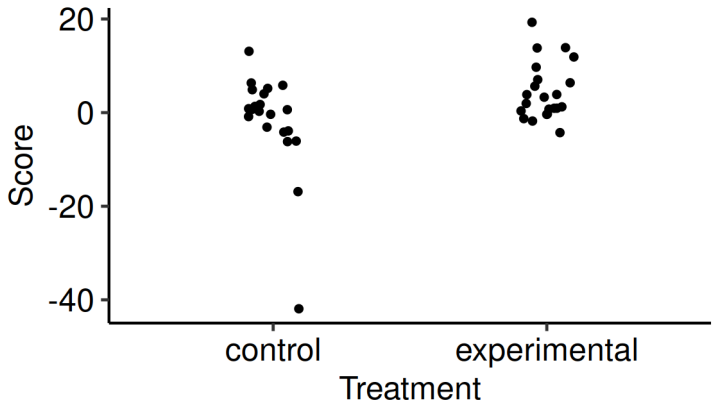


Figure 1: Problem 1 image is available here

1a. What is the null hypothesis? [2.5 points]

YOUR RESPONSE HERE

1b. What conclusion could you draw from the dotplot? [2.5 points]

YOUR RESPONSE HERE

1c. Here is computer output for a t test. Explain what the P-value means in the context of this study. [5 points]

$t = 2.505$, $df = 33.23$, p-value = 0.00866 alternative hypothesis: true difference in means is greater than 0

YOUR RESPONSE HERE

1d. If type-1 error $\alpha = 0.01$, what is your conclusion regarding H_0? State your conclusion in the specific context of this problem. [5 points]

YOUR RESPONSE HERE

1e. The computer output in part (c) is for the directional test. What is the P-value for the nondirectional test? [5 points]

YOUR RESPONSE HERE

1f. If the test were nondirectional, and α= 0.01, what conclusions would we make? [5 points]

YOUR RESPONSE HERE

Problem 02 [10 points]

Normality goes for a toss: Researchers took skin samples from 10 patients who had breast implants and from a control group of 6 patients. They recorded the level of interleukin-6 or IL6 in picogram/ml/10 g of tissue, a measure of tissue inflammation, after each tissue sample was cultured for 24 hours. The dataset is available below (in R)

```
il6.breast.implant.patients <- c(231, 308287, 33291, 124550, 17075,
                                 22955, 95102, 5649, 840585, 58924)
il6.control.patients <- c(35324, 12457, 8276, 44, 278, 840)
df.breast <- data.frame(value = il6.breast.implant.patients)
df.contorl <- data.frame(value = il6.control.patients)
```

2a. Draw a boxplot, violin and ridgeline plot [5 points]

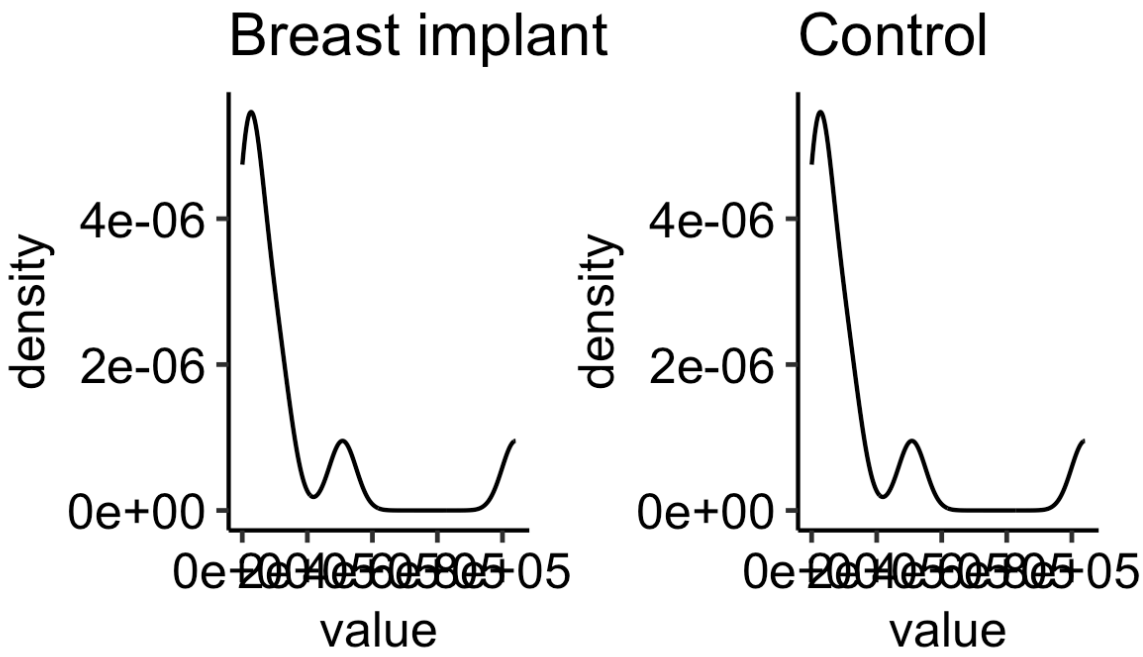
```
# YOUR RESPONSE HERE
# OUTPUT should be 3 plots
# While making it with the defaults but correctly which fetch you full points
# you can go the extra mile and show someone the creativ side of you
# good colors, good scaling, good lines, showing all data points, adding a legend
library(ggplot2)
library(ggridges)

# YOUR CODE HERE
```

2b. Draw a Q-Q plot for both the measurements [5 points]

You can use the geom_qq function

```
# YOUR CODE HERE FOR GENERATING A QQ PLOT
# FIX ME - replace this with something that will give you a qqplot
implant.qqplot <- ggplot(df.breast, aes(x = value)) +
  geom_density() +
  ggtitle("Breast implant")
# FIX ME  - replace this with something that will give you a qqplot
contro.qqplot <- ggplot(df.breast, aes(x = value)) +
  geom_density() +
  ggtitle("Control")
# !! DO NOT EDIT/REMOVE !!
library(patchwork)
implant.qqplot | contro.qqplot
```



Problem 03 [10 points]

Sneak peek: I perform a t test of the null hypothesis that two means are equal. I decided to calculate the means and then choose an alternative hypothesis H_A $\mu_1 > \mu_2$ because I observed $\bar{y}_1 > \bar{y}_2$.

3a. Explain what is wrong (if anything) with this procedure and why it is wrong (if anything). [5 points]

YOUR RESPONSE HERE

3b. Suppose I reported t = 1.97 on 25 degrees of freedom and a P-value of 0.03. What is the proper P-value? [5 points]

YOUR RESPONSE HERE

Problem 04 [25 points]

"What I cannot build, I cannot understand – Feynman". We studied different t-tests and summarised it here. This was like a recipe - do this when you have this ingredient, do that if not. The goal of this problem is to flip this and figure out what happens if we try breaking this assumption

4a. The defaults. Simulate two normal samples (n=200) with each mean = 10 and sd = 1, 10000 times. Apply a t-test for each iteration and calculate the p-value. With $\alpha = 0.05$, how many times do you reject the null hypothesis that the mean of two samples is equal? What do you conclude? [2.5 points]

```
# # !! DO NOT EDIT/REMOVE !!
# OUTPUT: n_rejects -- Number of rejections for alpha=0.05

set.seed(42)
alpha <- 0.05
n_rejections <- 0
N_tries <- 10000
n_sample_size <- 200

# YOUR RESPONSE HERE
n_rejections / N_tries
```

```
[1] 0
```

n_rejections/N_tries: YOUR RESPONSE HERE

Conclusion: YOUR RESPONSE HERE

4b. Unequal variance. Simulate two normal samples (n=100) with each mean = 10 and sd1 = 1 and sd2=2.5, 10000 times. Apply a t-test for each iteration and calculate the p-value. With $\alpha = 0.05$, how many times do you reject the null hypothesis that the mean of two samples is equal? What do you conclude and is that unusual? Use the default `t.test` [2.5 points]

```
# OUTPUT: n_rejects -- Number of rejections for alpha=0.05

set.seed(42)
alpha <- 0.05
n_rejections <- 0
N_tries <- 10000
n_sample_size <- 200
mean1 <- 10
mean2 <- mean1

sd1 <- 1
sd2 <- 2.5

# YOUR RESPONSE HERE

n_rejections / N_tries
```

```
[1] 0
```

n_rejections/N_tries: YOUR RESPONSE HERE

Conclusion: YOUR RESPONSE HERE

4c. Unequal variance revisited: Repeat the example in 4b with now `var.equal=FALSE`:`t.test(var.equal=TRUE)`. With $\alpha = 0.05$, how many times do you reject the null hypothesis that the mean of two samples is equal? What do you conclude? [2.5 points]

```
# OUTPUT: n_rejects -- Number of rejections for alpha=0.05

set.seed(42)
alpha <- 0.05
n_rejections <- 0
```

```
N_tries <- 10000
n_sample_size <- 200
mean1 <- 10
mean2 <- mean1

sd1 <- 1
sd2 <- 2.5
# YOUR RESPONSE HERE
n_rejections / N_tries
```

```
[1] 0
```

n_rejections/N_tries: YOUR RESPONSE HERE

Conclusion: YOUR RESPONSE HERE

4d. Severe violation: Following 4c, now set sd1=10, sd2=1, and use a non-welch t-test to tabulate the number of times you reject the null? With $\alpha = 0.05$, how many times do you reject the null hypothesis that the mean of two samples is equal? What do you conclude? [2.5 points]

```
# OUTPUT: n_rejects -- Number of rejections for alpha=0.05

set.seed(42)
alpha <- 0.05
n_rejections <- 0
N_tries <- 10000
n_sample_size <- 200

mean1 <- 10
mean2 <- mean1

sd1 <- 10
sd2 <- 1

# YOUR RESPONSE HERE

n_rejections / N_tries
```

```
[1] 0
```

n_rejections/N_tries: YOUR RESPONSE HERE

Conclusion: YOUR RESPONSE HERE

4e. Severe violation2: Following 4d, now simulate different sample sizes with n_sample_size1 =30 and n_sample_size2=70, sd1=10, sd2=1, and use a non=welch t-test to tabulate the number of times you reject the null? With $\alpha = 0.05$, how many times do you reject the null hypothesis that the mean of two samples is equal? What do you conclude? [2.5 points]

```
# OUTPUT: n_rejects -- Number of rejections for alpha=0.05

set.seed(42)
alpha <- 0.05
n_rejections <- 0
N_tries <- 10000
n_sample_size1 <- 30
n_sample_size2 <- 70
mean1 <- 10
mean2 <- mean1

sd1 <- 10
sd2 <- 1

# YOUR RESPONSE HERE
n_rejections / N_tries
```

```
[1] 0
```

n_rejections/N_tries: YOUR RESPONSE HERE

Conclusion: YOUR RESPONSE HERE

4f. Severe violation3: Repeat 4e with n_sample_size1 = 70 and n_sample_size1=70, sd1=10, sd2=1, and use a non-welch t-test to tabulate the number of times you reject the null? With $\alpha = 0.05$, how many times do you reject the null hypothesis that the mean of two samples is equal? What do you conclude? [2.5 points]

```
# OUTPUT: n_rejects -- Number of rejections for alpha=0.05

set.seed(42)
alpha <- 0.05
n_rejections <- 0
N_tries <- 10000
n_sample_size1 <- 70
n_sample_size2 <- 30
mean1 <- 10
mean2 <- mean1

sd1 <- 10
sd2 <- 1

# YOUR RESPONSE HERE
n_rejections / N_tries
```

```
[1] 0
```

4g. Severe violation4: Repeat 4f with n_sample_size1 = 70 and n_sample_size1=70, sd1=10, sd2=1, and use a Welch t-test to tabulate the number of times you reject the null? With $\alpha = 0.05$, how many times do you reject the null hypothesis that the mean of two samples is equal? What do you conclude? [2.5 points]

```
# OUTPUT: n_rejects -- Number of rejections for alpha=0.05

set.seed(42)
alpha <- 0.05
n_rejections <- 0
N_tries <- 10000
n_sample_size1 <- 70
n_sample_size2 <- 30
mean1 <- 10
mean2 <- mean1

sd1 <- 10
sd2 <- 1
# YOUR RESPONSE HERE
n_rejections / N_tries
```

```
[1] 0
```

n_rejections/N_tries: YOUR RESPONSE HERE

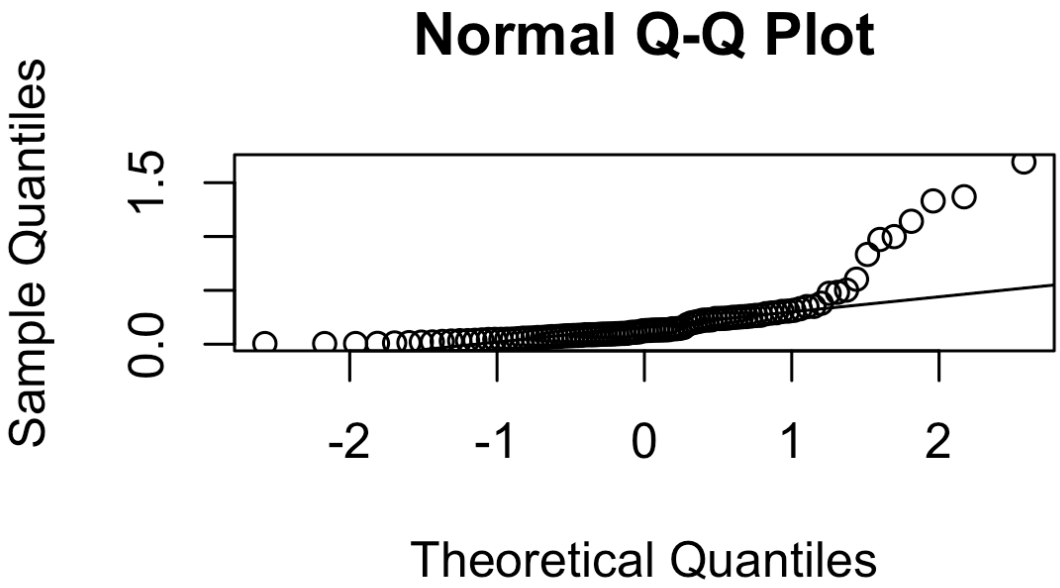
Conclusion: YOUR RESPONSE HERE

4h. Toss in exponential: Hopefully you have got a feeling of what is happening. Now we take normality for a toss. I have mentioned multiple times in the class, it is a relaxable assumption, but is it really? [2.5 points]

For example if we simulate an exponential distribution this is what it looks like

```
set.seed(42)

exp_data <- rexp(100, rate = 5)
qqnorm(exp_data)
qqline(exp_data)
```



For a n_sample_size1 = 70 and n_sample_size1=n_sample_size2=100, rate parameters r1=5 and r2=5 use a t-test to tabulate the number of times you reject the null? With $\alpha = 0.05$, how many times do you reject the null hypothesis that the mean of two samples is equal? What do you conclude?

```
# OUTPUT: n_rejects -- Number of rejections for alpha=0.05
set.seed(42)

# YOUR RESPONSE HERE

n_rejections / N_tries
```

```
[1] 0
```

n_rejections/N_tries: YOUR RESPONSE HERE

Conclusion: YOUR RESPONSE HERE

4i. Toss in exponential2: Following 4h, repeat the experment with sample_size1=n_sample_size2=100, rate parameters r1=5 and r2=10 use a t-test to tabulate the number of times you reject the null? With $\alpha = 0.05$, how many times do you reject the null hypothesis that the mean of two samples is equal? What do you conclude? [5 points]

```
# OUTPUT: n_rejects -- Number of rejections for alpha=0.05

set.seed(42)
# YOUR RESPONSE HERE
n_rejections / N_tries
```

```
[1] 0
```

Problem 05 [30 points]

Standard error: Adults who broke their wrists were tested to see if their grip strength (kg) decreased over 6 weeks. The data is provided as a tsv here.

5a. What is the null and alternative hypothesis? [2.5 points]

YOUR RESPONSE HERE

5b. Which test would you employ to test your hypothesis and why? [2.5 points]

YOUR RESPONSE HERE

5c. If there are conditions to be satisfied for applying your test, write R code to output checks for validity of your test. [5 points]

```
# CODE for validity of your test of choice
# INPUT: data frame of subject based grip strengths
# OUTPUT: plots or metrics for assessing the validity of your test
```

5d. Perform a test (using R) to test your hypothesis in a) reporting p-value and the conclusion. [5 points]

```
# CODE
```

p-value: YOUR RESPONSE HERE

Conclusion: YOUR RESPONSE HERE

5e. Non parameteric test [5 points]

A class of tests we did not discuss in detail in class are non-parameteric tests. We did discuss them in this slide. Using hints from the slide, perform a relevant non-parameteric test (using R) for testing your hypothesis in a) and report the pvalue and conclusion

```
# YOUR RESPONSE HERE
```

5f. Transform the original variable by a log() transformation and repeat your analysis in 5c and 5d [10 points]

```
# YOUR RESPONSE HERE
```

```
# YOUR CODE HERE
```

p-value: YOUR RESPONSE HERE

Conclusion: YOUR RESPONSE HERE

Question: how does the result of your analysis compare with your conclusion in d?

YOUR RESPONSE HERE

Problem 06 [50 points]

PPT - Principal proteins test: This is a data heavy question and designed to give you another exposure to real world data which is messy, hard to parse and often not so well documented. It is also to tell you how a random (or not so random) twitter/X thread can be turned into an assignment problem - which is one of the reasons the assignment was delayed (the other one being figuring out how to combat ChatGPT usage for direct copy pasting answers (not code)). It is also to applaud the government of India for the wealth of data it collects - you just need to find(parse) it.

The tsv here has the amino acid content of different food items. The data was programmatically extracted from this PDF and contains the amino acid profile of various food items - some familar and some not so familiar ones. Very few people (in the world) have taken a look at this data in the way you are going to.

Using dimensionality reduction techniques we studied in the class, perform an exploratory analysis and build a story around your plot. While the question is broad, there are full points only for specific answers. The questions are broad, but your responses should be specific.

6a. The plot. Plot the output of PCA to demonstrate what the lower dimensional representation of the data looks like. [10 points]

Remember the dataset is messy, if you want to do dimensionality reduction, you want to only retain numeric columns. While the original tsv also has uncertainity, to make the task easier I have generated a cleaned version of the tsv here removing the uncertainity values. There are 21 columns in total:

1. `food_code` = a short code for different food items,
2. `food_name` = long description of the food
3. `number_of_regions` = Number of sampling regions (this can be IGNORED for this question)
4. 18 columns corresponding to the 18 out of 20 amino acids with absolute quantities for each

```
# YOUR CODE GOES HERE
# OUTPUT: A reduced dimensional representation of your data
# You can choose to color your data points on a specific variable of choice
# the variable can either exist in data or can be extracted(appended) by
# processing the data frame
df <- read_tsv("Table8_amino_acid_profile_no_uncertainity.tsv")
```

```
Rows: 612 Columns: 39
-- Column specification --------------------------------------------------------
Delimiter: "\t"
chr  (2): food_code, food_name
dbl (37): number_of_regions, Alanine, Arginine, Aspartic Acid, Glutamic Acid...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df <- df %>%
  drop_na() %>%
  unique() %>%
  as.data.frame()
```

6b. The story Do you notice any discernible pattern in your data? What are some factor(s) that explain your PC1 and PC2? [10 points]

```
# WRITE ADDITIONAL CODE TO EXPLAIN YOUR STORY
#
```

6c. The factors Identify the top 2 factors (features) that are associated with your PC1 and that with PC2. [15 points]

```
# WRITE ADDITIONAL CODE TO SHOW FACTORS ASSOCIATED WITH PC1 and PC2
#
```

6d. Is it statistically significant? Based on the topmost factor that you identified in 6c for PC1, perform a statistical test to test if this factor is statistically different between the two groups you identified in c. To define the two groups you can make use of the `code_desc` column (HINT: Individual food codes are not so useful but categories probably are) [15 points]

```
# YOUR CODE HERE
```