# STM1001: Assignment 4

## Science/Health Stream Students Only

* This assignment is due **Thursday 13th October at 11.59pm**
* This assignment is worth 15% of your final mark and is out of 50 marks. The available marks for each question are displayed in the question.
* You must submit your assignment electronically and as a single Word or pdf file via the LMS page for this subject. To avoid incurring late penalties, please ensure your assignment is correctly submitted and that the orientation of your file is correct (i.e. not sideways or upside down).
* Where questions require the use of R, you must attach the relevant code and computer output and/or plots.
* Where your answers to any question refer to a -value, you must explicitly state the -value you in your answer. If the -value is less than 0.001, express the -value as “.”
* Unless otherwise specified, assume a significance level of where relevant
* Round to four decimal places where relevant

*In submitting your work, you are consenting that it may be copied and transmitted by the University for the detection of plagiarism. Please start with the following statement of originality, which must be signed and dated by you: “This is my own work. I have not copied any of it from anyone else.”*

**Statement of Originality:**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Student ID | Signature | Date |
|  |  |  |  |

# 1 Question 1

*(13 marks total)*

## 1.1

Using R, perform Bonferroni correction and FDR correction on the following set of p-values, which have been computed as part of a multiple hypothesis testing procedure.

Summarise your results in a short statement, clearly noting the number of significant -values for each method (use an threshold).

*Note: Make sure to include any R code and output in your answer.*

*(5 marks)*

## 1.2

A recent study (Song et al. 2020) assessed gene expression differences between the leaves of healthy watermelon plants and watermelon plants that had undergone drought stress. In this question we will analyse this data [GSE144814](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE144814), with the aim of identifying statistically signicantly differentially expressed genes between the healthy and drought-stressed leaves.

The data for this question is in the watermelon\_data.RDS file on LMS. Download this data set, load it into R, and store it in an object called watermelon\_data.

*Hint: Make sure you are using the correct function to read in your RDS data.*

*(1 mark)*

## 1.3

Using an appropriate function from the edgeR R package, carry out FDR correction on the -values in the watermelon\_data data set. Specify that only results for the top 5,000 genes should be retained, and use an adjusted -value threshold of 0.01.

*Hint: Check Question 3 of Computer Lab 8B if you are not sure how to proceed.*

*(5 marks)*

## 1.4

How many of the 5000 top genes are identified as significant, following this FDR correction process?

*(2 marks)*

# 2 Question 2

*(24 marks total)*

For this question, you will need to analyse data found in the data file depression.csv ([Kassambara 2019](#ref-datarium)). Download the data set from LMS and then save it in a relevant location on your PC.

You will then need to import your data into R. The following steps may be useful:

1. Set the working directory to the folder in which you have saved the depression.csv file
2. Load your data into RStudio using the below code:

* depression <- read.csv("depression.csv")

1. You should now be ready to carry out the data analysis in R using the data frame called depression

The data set provides the depression score for each individual in the study, measured at four different time points. The variables included in depression.csv are:

* id: The ID of the individual
* Time: Time-point at which the depression score was measured. Possible values are t0 (time-point 0), t1 (time-point 1), t2 (time-point 2) and t3 (time-point 3)
* Score: depression score

When answering the following questions, you may assume that all necessary hypothesis test assumptions have been met.

## 2.1

Create a boxplot of the depression scores, separated by Time. For this question, marks will be awarded for:

* Accuracy of the plot
* Presentation of the plot (title, axis labels, colours)
* Provision of R code

*(5 marks)*

## 2.2

Based on your boxplot from the previous question, do you believe that, on average, the depression scores changed significantly over time? Explain.

*(2 marks)*

## 2.3

Using an appropriate ANOVA analysis, we wish to test for a difference in average depression scores across time-points. Write down the null and alternative hypotheses, ensuring you define any parameters mentioned (e.g.  etc.).

*(2 marks)*

## 2.4

In this example, what is the *dependent* (response) variable?

*(1 mark)*

## 2.5

Carry out an appropriate ANOVA analysis to test for a difference in average depression scores across time-points. Provide your code and a summary of the results.

*(2 marks)*

## 2.6

Do we have enough evidence to conclude that there is a statistically significant difference in depression scores across time? Justify your answer.

*(2 marks)*

## 2.7

Write a one-sentence summary of your ANOVA results using the format taught in this subject.

*(5 marks)*

## 2.8

Comparing each pair of time-points with each other, which time-points are significantly different from each other in terms of average depression? Justify your answer with appropriate R code, output, and -values.

*(5 marks)*

# 3 Question 3

*(13 marks total)*

## 3.1

In this question, we will be examining a research paper. You can access the research paper [here](https://www.sciencedirect.com/science/article/pii/S0196655317301414/pdfft?casa_token=KrDCQ7ClCuwAAAAA:7YccqifLwztiRgVG1SidvSUKsTlshUAd5KFrwHaPwedNjNPg0bnm4PGkw_QBushkf9gVfH3JW9U&md5=0dd45172e7b9a2847cf51711a1f32c21&pid=1-s2.0-S0196655317301414-main.pdf). Note that you may need to use your La Trobe credentials to gain full access to the paper. Alternatively, you may access the paper from La Trobe library by searching for *“Self-reported behaviors and perceptions of Australian paramedics in relation to hand hygiene and gloving practices in paramedic-led health care”*.

Save a PDF copy of the paper on your device for your own reference. Using Harvard referencing style, provide the details of the paper here. For example:

Johnson, J. (2004) Flexible working: changing the manager’s role. *Management Decision*, [Online]. Available from:<http://www.emerald-library.com>. [Accessed 20 August 2004].

*(1 mark)*

Barr, N. et al. (2017) "Self-reported behaviors and perceptions of Australian paramedics in relation to hand hygiene and gloving practices in paramedic-led health care", *American Journal of Infection Control*, 45(7), pp. 771-778. doi: 10.1016/j.ajic.2017.02.020.

## 3.2

Read the **Abstract**, and answer the following questions.

### 3.2.1

In Topic 1B, we learnt about *Types of research*. What type of research has been used in this study? Explain.

*(2 marks)*

### 3.2.2

In terms of the ***Hierarchies of Evidence*** framework discussed in this subject, what type of evidence is provided by the article? Name one type of evidence that would be stronger than the type provided in the article, and explain the difference between the two.

*(3 marks)*

## 3.3

Read the **Methods** > *Research design* section, and answer the following question:

For the first part of the study, what type of sampling method is most likely to have been used? Is this a random or non-random sampling method?

*(2 marks)*

## 3.4

In the **Results** > *Confidence with IPC practices*, the result of a hypothesis test is provided. Answer the following questions with regard to this test.

### 3.4.1

What type of test has been used?

*(1 mark)*

### 3.4.2

What was the -value provided? Explain what this result means.

*(2 marks)*

## 3.5

What type of statistics (descriptive or inferential) are provided in Figures 1 and 2? Explain.

*(2 marks)*

# References

Song, Q., Joshi, M., DiPiazza, J. and Joshi, V. 2020. *Functional Relevance of Citrulline in the Vegetative Tissues of Watermelon During Abiotic Stresses*. Front Plant Sci, 11:512. PMID: 32431723

Kassambara, Alboukadel. 2019. *Datarium: Data Bank for Statistical Analysis and Visualization*. <https://CRAN.R-project.org/package=datarium>.