# STM1001: Assignment 3

## Science/Health Stream Students Only

* This assignment is due **Thursday 30th May (Week 12) 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 as a single **Word or pdf file** and that the orientation of your file is correct (i.e. not sideways or upside down).
* Where questions require the use of R/jamovi, 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
* Unless otherwise specified, round to three 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

*(8 marks total)*

In this question, you will perform a -means clustering cluster analysis.

A study by ([Fischer et al. 2018](#ref-Fischer)) compared various physical characteristics of South Georgian Diving Petrels living in different locations in the Southern Hemisphere. The petrels were grouped into three main petrel populations: South Atlantic Ocean (*SAO*), South Indian Ocean (*SIO*) and New Zealand (*SPO*).

## 1.1

A subset of the Fischer et al. (2018) data set has been prepared for you, and is available on the LMS.

1. Navigate to the Assignment 3 section of the Assignments tile on LMS, and download the *A3\_petrel\_data.csv* file
2. Load this data into jamovi

The petrel variables in this data set include:

* Live: Whether the bird was alive (1) or a museum specimen (0)
* Origin: The geographic origin location of the bird
* Pool: The population (SAO, SIO or SPO)
* WingLength: The wing length (mm)
* ForkDepth: The distance between the longest and shortest tail feathers (mm)
* BillLength: The bill length (mm)
* BillWidth: The bill width (mm)
* BillDepth: The bill depth (mm)
* SkullLength: The skull length (mm)

*Please note that the continuous numeric variables in your data object have been* **scaled**, *to help with subsequent analyses*.

Before you proceed, make sure that you have the snowCluster module installed in jamovi. *If you encounter errors using this module, double-check you are using the latest version of jamovi.*

*(No marks)*

## 1.2

Perform **-means clustering** on your A3\_petrel\_data.csv data, using the numeric variables and a value of 3.

Use the default algorithm setting. Do not worry if either of the following error messages appear

* more cluster centers than distinct data points
* number of cluster centers must lie between 1 and nrow(x)

It is safe to ignore these messages.

After carrying out the -means clustering analysis, complete the following:

* Include the Clustering Table in your answers
* List the number of birds in each cluster
* Create a plot of the means across clusters and include this in your answers

*(4 marks)*

## 1.3

Create a cluster plot to visualise the -means clusters identified in 1.2.

Include your plot in your answer and provide a comment on the plot, noting key features. Do you notice anything interesting or strange?

*(4 marks)*

# 2 Question 2

*(10 marks total)*

Suppose you have computed a set of -values as part of a multiple hypothesis testing procedure for genomic data. In your chosen statistical software (jamovi or R) run the code below to store these -values in the object base\_p\_values, before answering the following questions.

base\_p\_values <- c(0.0024, 0.2713, 0.0089, 0.0080, 0.0037, 0.0037, 0.0071, 0.0028, 0.0029, 0.0216)

## 2.1

Perform Bonferroni correction on these -values to control the FWER. Make sure to include any R code and output in your answer.

*(1 mark)*

## 2.2

Using an threshold, summarise your results in a short statement, clearly noting the number of statistically significant -values before adjustment, and then after adjustment.

*(2 marks)*

## 2.3

A recent study ([Sharma et al. 2021](#ref-Sharma)) assessed gene expression differences between Annapurna (heat tolerant variety) and IR64 (heat sensitive variety) rice cultivar seedlings when exposed to heat stress (42 degrees Celsius for 2 hours).

In this question we will analyse their data, with the aim of identifying statistically significantly differentially expressed genes between the two varieties of rice, under heat stress. (If you are interested, the original data are available here: [GSE41648](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41648)).

A selection of this data is stored in the A3\_rice\_data.csv file on LMS. Due to its size, this data can be considered Big Data. Download this file now.

Open the A3\_rice\_data.csv data file in jamovi. Then run the following code in the *Rj Editor*. This code will load the data (line 1) and then store the base -values for you (line 2) in the rice\_base\_p\_values object:

rice\_data <- data  
rice\_base\_p\_values <- rice\_data$P.Value

Of the methods we have learnt about in this subject, which method of -value adjustment is most appropriate for this data (Bonferroni, Holm, Hochberg, Hommel, or FDR)? Explain.

*(2 marks)*

## 2.4

Using your chosen method from 2.3, carry out a correction on the -values in rice\_base\_p\_values. Store the adjusted -values in a new object called rice\_p\_adjusted. Provide your R code in your answer (you do not need to provide any other output).

*(1 mark)*

## 2.5

Using a -value threshold of , how many statistically significant -values are present in rice\_base\_p\_values?

*Hint: Check Question 4 of* [Computer Lab 8B](https://rpubs.com/LTU_STM1001/SMDSMCL8) *if you are not sure how to proceed.*

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

*(2 marks)*

## 2.6

Using an adjusted -value threshold of , how many statistically significant -values are obtained for your chosen correction method, in rice\_p\_adjusted?

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

*(2 marks)*

# 3 Question 3

*(16 marks total)*

For this question, you will need to analyse data found in the data file anxiety\_data.csv ([Kassambara 2019](#ref-datarium)). Download the data set from LMS and open it in jamovi.

The data set contains the following variables:

* id : the ID of the individual (ranges from 1 to 45: we are considering a random sample of 30 individuals from the original 45 in the data set)
* t1 : anxiety score at time-point 1
* t2 : anxiety score at time-point 2
* t3 : anxiety score at time-point 3

When answering the following questions, you may assume that all necessary hypothesis test assumptions have been met. Remember to round to three decimal places where relevant.

## 3.1

What is the mean anxiety score for each of the three time-points?

*(2 marks)*

## 3.2

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

*(4 marks)*

## 3.3

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

*(1 mark)*

## 3.4

Carry out an appropriate ANOVA analysis to test for a difference in average anxiety scores across time-points. Provide the ANOVA table obtained from jamovi.

*(3 marks)*

## 3.5

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

*(2 marks)*

## 3.6

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

*(4 marks)*

# 4 Question 4

*(11 marks total)*

## 4.1

In this question, we will be examining a research paper. The article outlines a protocol for a future research study that will be carried out. You can access the article [here](https://bmjopen.bmj.com/content/13/4/e070533).

Save a PDF copy of the paper on your device for your own reference. Using Harvard or APA7 referencing style, provide the details of the paper here. For example:

Wattal, S, Schuff, D, Mandviwalla, M & Williams, CB 2010, ‘Web 2.0 and Politics: The 2008 U.S. Presidential Election and an e-politics Research Agenda’, *MIS Quarterly*, vol. 34, no. 13, viewed 20 June 2019, <http://immi.se/intercultural/nr13/petraki.htm>.

*(1 mark)*

## 4.2

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

### 4.2.1

In terms of the ***Hierarchies of Evidence*** framework discussed in this subject, what type of evidence will be provided in the study? Name one type of evidence that would be stronger than this type of evidence.

*(2 marks)*

### 4.2.2

What is the primary aim of the study?

*(2 marks)*

### 4.2.3

Is the study design Descriptive, Observational, or Experimental? Justify your answer.

*(2 marks)*

### 4.2.4

The **Methods** section mentions **double-blinding**. Who would be “blinded” in this study?

*(2 marks)*

### 4.2.5

The authors state that,

“*A small randomised trial showed an approximately one-third lower incidence of AD in goat milk formula-fed compared with cow milk formula-fed infants. However, due to limited statistical power, AD incidence difference was not found to be significant.*”

The authors propose to increase the sample size in the future study to rectify this issue.

* What level of statistical power are the authors aiming for in the future study? (You will need to refer to the body of the article to answer this question.)
* By increasing the statistical power, what Type of error is now less likely to occur?

*(2 marks)*

# 5 Question 5

*(5 marks total)*

For this question, your task is to reflect upon what you have learnt in STM1001, and to write a self-reflective piece (a short paragraph).

While your response must relate to STM1001, what you choose to write about here is to a large extent up to you, and will be specific to your chosen tertiary qualification(s). Some examples (not exhaustive) are provided below:

* A reflection on your most favourite/least favourite STM1001 content covered
* How what you have learnt in STM1001 could tie in with your current/future career
* An area of statistics you would like to learn more about, based on what you have learnt in STM1001
* A difficulty you encountered in learning the content, and what you did to address this issue

Regardless of your selected area of discussion, your response must demonstrate that you have reflected upon and thought about your learning. Marks will be awarded based on this criterion.

# 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

Fischer, J. H., I. Debski, C. M. Miskelly, C. A. Bost, A. Fromant, A. J. D. Tennyson, and et al. 2018. “Analyses of Phenotypic Differentiations Among South Georgian Diving Petrel (Pelecanoides Georgicus) Populations Reveal an Undescribed and Highly Endangered Species from New Zealand.” *PLoS ONE* 13 (6): e0197766. <https://doi.org/10.1371/journal.pone.0197766>.

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

Sharma, E., P. Borah, A. Kaur, A. Bhatnagar, and et al. 2021. “A Comprehensive Transcriptome Analysis of Contrasting Rice Cultivars Highlights the Role of Auxin and ABA Responsive Genes in Heat Stress Response.” *Genomics* 113 (3): 1247–61. <https://doi.org/10.1016/j.ygeno.2021.03.007>.