ICA SPECIFICATION

Module Title: **Statistical Methods for Data Analytics**	Module Leader: **Prof Claudio Angione**
	Module Code: **CIS4066-N**
Assignment Title: **Statistical Methods for Data Analytics ICA**	Deadline Date: **10 January 2024**
	Deadline Time: **4:00pm**
	Submission Method: Middlesbrough Tower ☐ Online (Blackboard) ☑

Online Submission Notes:

Please follow carefully the instructions given on the Assignment Specification

When an extension has been granted, a fully completed and signed Extension form must be submitted to the SCEDT Reception.

Library Support for Academic Skills

Did you know you can book an individual 30 minute tutorial in the Learning Hub with an adviser to help you with your academic skills, writing or numeracy? Or that there are loads of really useful workshops available to help you with your studies and assessments? Have a look at the Succeed@Tees workshops for more details.

FULL DETAILS OF THE ASSIGNMENT ARE ATTACHED INCLUDING MARKING & GRADING CRITERIA

Marking criteria

Each question has individual marks. There are 100 marks available in total.
The individual mark for each question is reported in square brackets at the end of the question. **The pass mark is 50%.**

The grade will be the sum of the individual marks, and the following scheme indicates the quality of submission associated with the overall grade. Grade boundaries are identified in accordance with the School's standards.

A (≥70%) Excellent piece of work of a professional standard across all parts. Demonstrates learning and other input beyond the taught programme.

B (60-69%) Good piece of work. Satisfies all requirements to a high standard. A carefully designed solution to the questions.

C (50-59%) Satisfactory. Meets the majority of the assignment's objectives. Demonstrates broad understanding and basic ability to use the key ideas introduced in the taught programme. A few gaps or omissions, but overall satisfactory solutions.

D (40-49%) Relatively weak. Some assignment objectives met but generally unconvincing solutions to the questions.

E (30-39%) Poor. Few objectives met. Little evidence of understanding or use of taught material. No attempt to extend learning, poor solutions to the questions.

F (<30%) Inadequate. Unsatisfactory. Objectives not met.

Learning outcomes

This assessment constitutes 100% of the overall module mark. It covers all the module learning outcomes as detailed below.

Personal and Transferable Skills

1. Use own judgement to select a valid statistical method in the context of the research project.
2. Manage complex data within the application of data analysis software in order to run tests.

Research, Knowledge and Cognitive Skills

3. Analyse complex related and unrelated data sets using a range of methods.
4. Critically analyse and interpret outcomes of statistical tests in order to identify patterns and significance levels.
5. Critically appraise the validity and reliability of the methods available.

Professional Skills

6. Demonstrate an ethical understanding of data analysis and its effect in a wider social context.

Statistical Methods for Data Analytics CIS4066-N

In-Course Assessment

Submission Method

The work should be submitted as a single Word/PDF document to Blackboard.

Include a title page with:

- Name of the module
- Module code
- Your name
- Your student number
- The submission date

Your submission should also include an appendix containing your R code.

Important

It is very important that you include explanations and step-by-step justifications for your answers.

Please note this is an individual piece of work. Group work is not permitted.

Qualitative statistics: thematic analysis

Using the Guardian website (https://www.theguardian.com/uk), select an article with at least 20 comments. Select a subset of comments (around 20) for the following analysis.

1. Provide a link to the article you have selected. List the comments you have chosen, the key themes you identify and the number of times they occur from the comments.

   [4]

2. Describe the themes you identified and the key features that make the theme recognisable (words, metaphors, emotional language, etc.). Discuss your findings - what conclusions can you make?

   [5]

Probability and statistics fundamentals

3. Given the event E: “The Paris Olympic Games will take place in 2024”, define the following events F_1 and F_2, both in general (using the definitions of dependent/independent events), and giving concrete examples of what they might be:

   - an event F_1 such that E and F_1 are independent
   - an event F_2 such that E and F_2 are dependent

   [6]

4. Derive with mathematical steps the final formula of Bayes’ theorem. Describe in general terms in what cases the theorem is useful, and propose a concrete scenario where it can be used.

   [6]

5. Discuss the different scales of measurement. Provide two examples for each scale.

   [6]

6. There are four medals (Gold, Silver, Bronze and Wood) on a table, but they are all wrapped with dark wrapping paper, such that it is impossible to distinguish them. You would like to find the gold medal.

   The game starts as follows. You pick one medal without unwrapping it, and then the game host unwraps one of the remaining medals and reveals that it is a silver medal. (Assume here that the host unwraps a medal with equal probability, but knowing where the gold medal was and avoiding unwrapping the gold medal if still on the table, to keep the game interesting to watch until the end.)

   You have now three medals left to unwrap (one in your hand, two on the table). At this point, the host gives you the option to change your mind and swap your medal for one of the two left on the table.

   What would you do at this point? Would you keep your medal, or swap it with one of the two medals left on the table? If so, which one?

   Hints: Find the solution by using Bayes’ theorem, calculating all the conditional probabilities involved. Start calculating the probability of having Gold in our hands given that we know that the host unwraps Silver, $P(G|Hs) = \ldots$.
   Then compare with the probability of having Bronze or Wood in our hands given

that we know that the host unwraps Silver, $P(B|Hs) = \ldots$, $P(W|Hs) = \ldots$

[12]

7. In June 2021, during the vaccine rollout for the Covid-19 emergency, it was estimated that 90% of the population over 50 years old were fully vaccinated, while only 6% were completely unvaccinated. (The remaining 4% had only one dose or had an unknown vaccination status, and therefore will not be considered here.)

   A Public Health England report on cases and hospitalisation from the "delta" variant (originally sequenced in India) was published at the end of June 2021. The report showed that, between February and June 2021, among the 418 people admitted to the hospital with the "delta" variant:

   - 163 were fully vaccinated
   - 136 were not vaccinated
   - The remaining people had only one dose or an unknown vaccination status and will not be considered here

   One may therefore wrongly conclude that a fully vaccinated person is surprisingly more likely to be hospitalised than an unvaccinated person. Using Bayes' theorem to calculate the relevant probabilities from the data above, prove that this claim is wrong. Show that this data actually proves that vaccines are extremely effective at reducing the risk of hospitalisation after contracting the "delta" variant.

   [10]

Central tendency and variability

8. The 2010 salaries of the White House staff are provided in the table *"2010_White_House_Staff.xlsx"*

   Carry out a pipeline of descriptive statistical methods in R, including central tendency and variability measures, to describe the dataset. Interpret and discuss the results, showing also some plots.

   [12]

Statistical tests

9. A variable X follows a normal distribution with mean 1.2 and standard deviation 2. Calculate the probability P(X < 0.5), computationally in R or usual manual tables.

   [4]

10. You would like to test whether a herb works for the treatment of insomnia. 80 people volunteered to take part in the study.

    - Design how you would carry out the experiment, what tasks the participants should perform, and define what could be a null and alternative hypothesis in this case.
    - Describe what could be an error of type I or type II, how they are defined theoretically and what they represent in this case.

    [10]

11. A company has produced a batch of 1000 CPUs whose clock speeds follow a normal distribution centred around 2.2 GHz, with a standard deviation of 0.5 GHz. The company is trying new approaches to manufacturing CPUs, therefore 20 of these CPUs were produced with an additional new experimental feature. The clock speed of these 20 experimental CPUs is as follows (in GHz):

    2.5, 1.7, 2.9, 2.7, 1.4,
    1.9, 1.3, 1.9, 2.8, 2.2,
    2.4, 1.7, 1.8, 2.2, 2.2,
    1.9, 2.7, 3.3, 1.9, 2.1.

    Design and perform a statistical test to check if the difference in clock speed for this new experimental technology is statistically significant, or if the difference is just due to chance.

    Hints: use a one-sample t-test, specifying the assumptions and finding the p-value. What additional assumption would be needed to use the z-test instead?

    [10]

Regression

12. Using the file boxOffice.csv, perform a logistic regression analysis to find out whether the budget spent on a movie affects its chances of winning an Oscar. Discuss the general role of logistic regression, the logit scores, z-values and p-values from the summary of the model output. Can we reject the null hypothesis?

[10]

13. What needs to change in the overall problem and/or assumptions if one wants to use linear regression?

[2]

14. What other methods could one try if linear regression does not perform well?

[3]