COMP6235 Stats Coursework Instructions

Module:	*Foundations of Data Science*	**Lecturers:**	*JS*		
Assignment:	*Statistics coursework*	**Weight:**	*30%*		
Deadline:	*4pm 23/11/2023*	**Feedback:**	*11/01/2024*		

Instructions

The following coursework is worth 30**%** of the assessment of the module. Please note that you are expected to use Python to carry out this coursework (hence, e.g., plots are expected to have been generated using Python), and then create a technical report.

Download the data set fish1.txt about the catch of a hypothetical fishing fleet from

http://edshare.soton.ac.uk/19466/

and import it into Python. The data set records data for a time period of one day during which one fisherman has fished in a lake. The fisherman uses three types of fishing rods, labeled A, B, and C, each using different bait. The fisherman has recorded every catch he has made during this time. The data set consists of three columns with X values giving the times at which the fisherman has made a catch, the Y values indicate the size of that catch (i.e. its weight in kg), and the Z values give a letter A, B, or C which indicates which fishing rod was used to make that catch. Using Python, your task is to analyse this data set.

In a first step, generate a plot that illustrates the distributions of X values (times of catch, the format is hours, fraction of hours on a 24h schedule for the day). Then also plot the distribution of Y values (size of catch), and finally generate a plot which analyses the effectiveness of each type of bait. Characterise and describe these distributions by measures of centrality, spread, and suitable additional measures introduced in the introduction to statistics lectures that you think shed light on the shape of the respective distributions. Assuming that the data are a sample from a larger population, give mean values with 95% confidence intervals for both distributions.

In a second step, it is of interest to analyse the dependence between time of catch (X value) , size of catch (Y value), and the type of bait used (Z). Generate suitable plots to analyze these relationships and characterize them by statistical measures. What is the correlation between X and Y? Analyse the amount of information about Y that is given by knowledge of X.

More generally, address the following questions and give support for your answers:

What is the best time to go fishing at this lake?
Which bait is most effective?
What is the best type of bait to use at 3pm in the afternoon?

Write up your finding in a short technical report of no more than 4 pages (and font size no smaller than 10pts) and submit it electronically via handin as one pdf file.

Submission

You must submit the following documents

One pdf document that contains your written technical report and includes the figures you produced

before **4pm Thursday 23th November 2023**.

Marking Scheme

	Distinction (above 70%)	Pass (50% – 69%)	Fail (below 50%)
Quality of the figures (5 marks)	Very good figures. They meet professional standards, including all the necessary information. Relationships discussed in the text clearly are visible.	Figures explain the data analysis but could be improved. Some information was missing or not clear from the report text.	Figures exist but not in professional standards. They may show some information but quite general, and some data is not correct. They might be difficult to read as well. Some key components like axes and captions are missing. Or maybe there are no figures at all.
Technical content (15 marks)	Excellent content. All the questions have been addressed. Theories have been chosen properly and the answers are correct. Extensive review of related work as in a typical technical paper.	Good work. Some related work reviewed. Most of questions have been addressed properly, although there might be some mistakes for the answers. More detailed data analysis is preferred.	There are no or some attempts to solve the questions but not in the correct way. Few and/or incorrect answers provided. More data analysis is needed to support the provided findings.
Quality of the writing (10 marks)	Very well written, structured and formatted. No or few spelling grammatical errors. Abstract can be read independently. Clear description of findings about the dataset. Good references.	Well written, structured and formatted, references to material used. Some spelling and grammar mistakes. Description of findings are clear enough to understand.	Poorly or adequately written and structured report. Many spelling and grammar mistakes. Some of the description of findings is confusing. Poor or missing references/abstract.
Extra	If the submission is not a technical report, 15-mark penalty will apply (till mark of 0).		