COMP6235 Stats Coursework Instructions

Module:	*Foundations of Data Science*	**Lecturers:**	*JS*		
Assignment:	*Statistics coursework*	**Weight:**	*30%*		
Deadline:	*20/11/20*	**Feedback:**	*11/12/2020*		

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).

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 report of no more than 4 pages (and font size no smaller than 10pts) and submit it electronically via handin as one pdf file before 4pm Friday 20th November 2020.

Submission

You must submit the following documents

- **One** pdf document that contains your written report and includes the figures you produced.

The deadline for the submission is 4pm Friday 20th November 2020.

Marking Scheme

Quality of the figures (do they meet professional standards, are relationships discussed in the text clearly visible? Are axes labeled properly? Captions?): 20%
Technical content (do you address all questions? Are the answers correct?) : 50%
Quality of the writing (formatting, correct spelling and grammar, description of your findings about the data set): 30%