Assignment Q4 2019

The Nature of Data (301114)

Due 17th of November 2019

This assignment consists of four questions, each of equal value, giving a total mark of 40 for this assignment. The beginning of each question provides a breakdown of marks for each part of the question. For example, a breakdown of (1 + 3 + 6 = 10) implies a question of three parts, where the first, second and third parts are worth 1, 3 and 6 marks respectively. R and RStudio are *the only tools* to be used for generating answers used by this assignment, this includes such things as statistical results and graphs.

Submission is due Sunday the 17th of November 2019 via vUWS. Late submissions will receive a 10% reduction in marks for each day late. Submission consists of two parts:

- The primary document, in the form of a PDF file, which contains your textual answers and any other supporting material, such as graphs. It should also include relevant output from any R code / functions used to answer the question. No R code should appear in this document, except for perhaps mentioning the name(s) of relevant variables used in your R code.
- A single R script file that contains **all** the code used to produce answers for the assignment. All R code should only appear in this file. Include comments as appropriate to describe the purpose of a block of code. Relevant output of your code should be placed in the primary document.

Both of the above files must also contain the following:

Your name

Student number

Unit name

Unit number

in the case of the R script file, prefix each of the above four lines with "# ", also incorporate your name in the names of the above files, e.g. 301114_francoUbaudi.R

You *must* also include in your primary document the declaration shown on the next page. You *must* also sign that declaration.

Assignment Q4 2019

The Nature of Data (301114)

By including this statement, I the author of this work, verifies that:

- I hold a copy of this assignment and can produce it if requested by the unit co-ordinator
- I hereby certify that no part of this assignment has been copied from any other student's work or from any other source, except where due acknowledgement was made in this assignment
- No part of this assignment has been written for me by another person except where such collaboration has been authorised by the unit co-ordinator
- I am aware that this work may be reproduced and submitted to plagiarism detection software systems for the purpose of detecting possible plagiarism. *NB that such systems may retain a copy for future plagiarism checking*
- I hereby certify that I have read and understand what the School of Computing, Engineering and Mathematics, defines as minor and substantial breaches of conduct as outlined in the learning guide for this unit

Note: That the unit co-ordinator has the right not to mark this assignment, if the above declaration has not been added to the assignment.

Assignment Q4 2019

The Nature of Data (301114)

1. Answer the following questions (2 + 1 + 2 + 2 + 3 = 10)

(i)

In a certain part of a city, the need to buy drugs is known to be responsible for 68% of all thefts. What is the probability that three of the next four thefts are due to the need for drugs?

(ii)

Use the information from above (i) to answer this question. Provide a bar plot that shows the probabilities for zero, one, two, three and four thefts being due to drugs? The plot should be worthy of use in a report.

(iii)

The probability that a patient recovers from a lymphoma is 0.4. If 14 people are known to have this disease, what is the probability that either 4, 5, 6, 7, 8 or 9 of these people will survive?

(iv)

Use the information from above (iii) to answer this question. Provide a bar plot that shows the probabilities for each number of possible survivors? The plot should be worthy of use in a report.

(v)

Generate a Poisson approximation to (iv). Provide a comparative plot for this approximation? Make use of a legend in your comparative plot and make this graph worthy of use in a report. Describe the closeness of the distributions and provide an explanation of the result? Use no more than a total of five sentences.

2. Flu medication and blood pressure side-effect (1 + 2 + 4 + 2 + 1 = 10)

A new medication had been designed to reduce the chance of obtaining the flu. However it is suspected that the medication has a side effect of increasing the blood pressure of its consumer. To examine this suspicion an experiment was designed and necessary data collected. Your task is to analyse this data.

Use the question_2.csv file to answer the above question. This file contains records of before and after blood pressure and the type of medication used. The variables contained in the file are:

- medicationType: the type of medication taken by the participant:
  - either the active drug (the flu medication being examined)
  - or a Placebo (a control, containing no medication / drug)
- bloodPressure_before: blood pressure of the participant before receiving the medication
- bloodPressure_after: blood pressure of the participant after receiving the medication

Use the following steps to determine the effect of the medication on the participants' heart rate.

(i)

Compute the "mean after medication blood pressure" for those that received the medication and the "mean after medication blood pressure" for those that received the placebo. Also produce a boxplot for "after medication blood pressure", with respect to the two medications used? Make sure the boxplot is worthy of use in a report.

(ii)

Using the above results (i), does the active drug appear to increase blood pressure? Is there a flaw in the approach used to make this judgement, if so, how could it have been avoided? Your response must consist of no more than six simple sentences.

(iii)

Answer this question in light of any potential flaws identified above (ii). Using an appropriate set of hypotheses, determine if the active drug results in an increase in blood pressure, when compared to using no drug?

In answering the above question, make sure you cover the following aspects:-

- declare the hypotheses used and briefly explain them
- briefly describe the key elements of the test performed, e.g. data used, test parameters used
- briefly describe any key assumptions
- declare the conclusion drawn and the basis on which you drew that conclusion
- include a boxplot that reflects the hypothesis test performed and briefly comment on it

Use no more than a total of ten simple sentences for answering this question.

(iv)

What is the 90% confidence interval for the measured difference, according to the hypothesis test conducted above (iii)? Provide a brief interpretation of this interval and what it could mean with respect to the change in blood pressure affected by the two medications?

(v)

Briefly describe the role of a placebo in experiments such as the above?

Assignment Q4 2019

The Nature of Data (301114)

3. Religious denominations and attendance patterns (2 + 2 + 4 + 2 = 10)

This question investigates whether a connection exists between attendance patterns and religious denomination? Consider the following table:

Attendance Pattern	Protestant	Catholic	Jewish
Regular	182	213	203
Irregular	154	138	110

which shows three denominations and the attendance pattern of one thousand church members.

(i)

Produce a bar plot that clearly shows the distribution of religion versus attendance pattern. Make sure the bar plot is worthy of use in a report. Provide a brief interpretation of what the plot shows?

(ii)

Create suitable hypotheses to test whether a difference exists with respect to religion and attendance patterns. The hypotheses themselves should be brief and to the point. In addition, also provide a sentence or two explanation for each of them.

(iii)

Using the above hypotheses (ii), execute two versions of a hypothesis test:

- using a test that exploits a theoretical distribution
- using a simulation that involves many repetitions

Briefly describe the details of these two tests? You don't need to interpret the results here, rather, just how the tests were performed and the high-level purpose of the steps used.

(iv)

What conclusion did you draw from the hypothesis tests performed? Briefly explain your conclusion and basis for drawing that conclusion. Did both versions of the hypothesis tests agree with each other? Briefly elaborate on the differences of the two sets of results.

Assignment Q4 2019

The Nature of Data (301114)

4. Relationship between blood pressure and age for Pima females (2 + 2 + 3 + 3 = 10)

The Pima people of North America have one of the highest rates of type 2 diabetes in the world. It appears that a number of social and environmental factors have contributed to the incidence of diabetes for the Pima's. The data set to be used for this question is found in the file called PIMA.csv (use the version provided in vUWS), which contains information for 500 females and consists of the following four variables or features:

- age in years
- diastolic blood pressure
- Body Mass Index (BMI)
- whether the individual as ever been pregnant

an extract of the dataset is shown below

Age	Diastolic	BMI	Ever.pregnant
35	84	35	yes
23	90	44.1	no
27	80	44.2	no
41	84	39.9	yes
22	64	33.6	no

This question evaluates the relationship between age and blood pressure.

(i)

Produce a plot that shows the relationship between age and blood pressure, showing the independent variable on the x-axis. Include a simple linear regression model for this data in the above plot. Make this plot worthy of use in a report; not a work of art, just something intelligible.

(ii)

Using an appropriate method, calculate the correlation between the plotted variables above (i) and briefly provide an interpretation. Include a brief statement why you considered the chosen method was appropriate.

(iii)

Using the above (i) linear regression model, predict the diastolic blood pressure of a 39 year old female. Although this is a point estimate for the population, *briefly* consider the limitations of this estimate for 39 year old females in general (Pima population)?

(iv)

Determine a 90% confidence interval for the diastolic blood pressure of 39 year old females. Considering the results of (iii), why does this estimate provide better accuracy with respect to 39 year old females in general?