Office Use Only

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Monash_2-RGB |  |  |  |  |

**Semester One 2020**

**Final Assessment Period**

**Faculty of Information Technology**

**EXAM CODES: FIT5197**

**TITLE OF EXAM: Statistical Data Modelling**

**EXAM DURATION:** 2 hours 10 minutes or 130 minutes

***THIS PAPER IS FOR STUDENTS STUDYING AT: (tick where applicable)***

**🞏** Caulfield **🞏** Clayton **🞏** Parkville **🞏** Peninsula

**🞏** Monash Extension **🞏** Off Campus Learning **🞏** Malaysia **🞏** Sth Africa

**🗹** Suzhou (China)

During an exam, you must not have in your possession any item/material that has not been authorised for your exam. Any authorised items are listed below.

You must not retain, copy, memorise or note down any exam content for personal use or to share with any other person by any means following your exam.

As a student, and under Monash University’s Student Academic Integrity procedure, you must undertake your in-semester tasks, and end-of-semester tasks, including exams, with honesty and integrity. In exams, you must not allow anyone else to do work for you and you must not do any work for others. You must not contact, or attempt to contact, another person in an attempt to gain unfair advantage during your exam session. Assessors may take reasonable steps to check that your work displays the expected standards of academic integrity. e.g. perform similarity checking on your submission. If required, you may be contacted after your exam to discuss any concerns.

Failure to comply with the above instructions, or attempting to cheat or cheating in an exam is a discipline offence under Part 7 of the Monash University (Council) Regulations, or a breach of instructions under Part 3 of the Monash University (Academic Board) Regulations.

**AUTHORISED MATERIALS**

**OPEN BOOK 🗹 YES 🞏 NO**

**CALCULATORS 🗹 YES 🞏 NO**

**SPECIFICALLY PERMITTED ITEMS 🗹 YES 🞏 NO**

**if yes, items permitted are:** A ruler is allowed and the calculator should be a non-programmable calculator. FIT5197, 2020 Semester 1, Formula Sheet, mobile phone/camera/tablet for uploading images of hand written work into this word document.

***Candidates must complete this section if required to write answers within this paper***

STUDENT ID: \_\_ \_\_ \_\_ \_\_ \_\_ \_\_ \_\_ \_\_

|  |  |  |
| --- | --- | --- |
| Question | Points | Score |
| [Descriptive Statistics](#_bookmark0) | 5 |  |
| [Probability](#_bookmark1) | 7 |  |
| [Expectation](#_bookmark2) | 6 |  |
| [Distributions](#_bookmark3) | 8 |  |
| [Inference](#_bookmark4) | 6 |  |
| [Simulation](#_bookmark5) | 10 |  |
| [Regression](#_bookmark7) | 8 |  |
| [Modelling](#_bookmark8) | 10 |  |
| Total: | 60 |  |

**Instructions**

* The formulae sheet for the unit is provided in a separate booklet. This supports answers for many of the questions.
* For all questions, part marks are given for working. So it is best to show working. Then you can still receive part marks if your final answer is wrong.

**ASSESSMENT MARKS**

Total marks for the Exam is **60**.

This exam accounts for **50% of the total assessment** in FIT5197.

**Instructions**

* This exam consists of eight topics listed in the table above and you should complete all topics.
* At the exam start time you will download the paper, then during the exam you will **hand-write your answers** on your own paper, **take photos** of your answers using a phone, tablet or camera, **upload them** to your computer and **embed them** **in this word document**.
* You should embed images of your answers to a given topic at the end of each topic in the document since you will likely write answers to different questions on the same topic on the same page of paper. This means you should take images of entire pages, rather than individual questions to save time when embedding images.
* Please answer concisely with clear working for each question and make sure you label your answer to each question/sub-question with the question label (e.g. “Part 1, Q1(a)”).
* We strongly recommend you set an alarm for ten minutes before the end time.
* Remember, at the end of the time allowed, you must upload the exam paper (with embedded images of your answers on it) to be marked. It will take time to embed the images and submit the document.

1. **Descriptive Statistics: ………………………………………………….*Total 5 marks***
2. Consider the following set of numbers: 12, -41, -7,3,9,21,8,17,-12,2,11. For each of the questions below, state your answer, showing working if necessary.

(2 marks) (a). What is the range and the inter-quartile range?

(2 marks) (b). Given the sum (∑*x*) is 23 and the sum of squares (∑ *x*2) is 3027, what is the sample

standard deviation to 3 decimal places?

(1 marks) (c). Determine if there are any outliers for this sample.

**EMBED IMAGES OF YOUR HAND-WRITTEN ANSWERS TO THIS TOPIC HERE**

**2. Probability:** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .***Total: 7 marks***

(2 marks) (1) You want to determine if too much exercise leads to an increase in cardiovascular mortality (i.e.

risk of dying from cardiovascular disease). Prior research suggests that cardiovascular mortality as a function of volume of exercise follows a single exponential function where an increase in the volume of exercise leads to a reduction in cardiovascular mortality, but you are not sure if they studied high enough volumes of exercise. You enroll 10000 people and study them over term period of 10 years. Participants work with your research team to record the volume of exercise completed each week and any deaths. Over the study period 278 people drop out of the study. If the number of people in the study willing to do extreme amounts of exercise was small relative to those willing to do high, moderate or light exercise, what type of bias would we have to deal with and why?

(2 marks) (2) The joint probability distribution of two discrete random variables *X* and *Y* , i.e. *P* (*X, Y* ) is given by:

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Y* =yellow | *Y* =blue | *Y* =orange |
| *X* =smooth  *X* =rough | 0.25  0.25 | 0.1  0.2 | 0.15  0.05 |

* 1. What are the marginal distributions *P* (*Y* ) and *P* (*X*)?

(1 mark) b. Are X and Y independent?

(2 marks) (3) A marketing company believes that people can be divided into two classes — those that believe in climate change and those that do not. The company’s statistics show that a climate change believer will buy solar panels for their home at sometime within a fixed 6 month period with probability 0.7, whereas this probability decreases to 0.3 for a person that doesn’t believe in climate change. If we assume that 75 percent of the population believe in climate change, what is the probability that a person will buy solar panels within the next 6 months? What is the probability that this person is a climate change believer?

**EMBED IMAGES OF YOUR HAND-WRITTEN ANSWERS TO THIS TOPIC HERE**

**3. Expectation:** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .***Total: 6 marks***

(1 mark) (1) The wealth of an individual is a random variable with probability density function

Furthermore, you are given the following integral

What is the mean of *x*?

(2 marks) (2) Let *E*[*Z*] = 4 and *E*[*Z*2] = 5, *E*[*Y* ] = −1 and *E*[*Y* 2] = 21, and *Z* and *Y* are independent, then

what is *V* [(4*Z* + 3*Y* )]?

(3 marks) (3) A small-time loan shark has 400 loan customers. If the total annual loan repayments made by an individual customer is a random variable with mean $725 and standard deviation $950, approximate the probability that the average total annual repayments made across all customers is greater than $730.

**EMBED IMAGES OF YOUR HAND-WRITTEN ANSWERS TO THIS TOPIC HERE**

**4. Distributions:** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .***Total: 8 marks***

(2 marks) (1) You toss 10 unfair coins where the chance of tossing a head is 0.55. What is the combined

probability of getting exactly 7 or exactly 8 heads?

(2 marks) (2) Suppose that a binary message—either 0 or 1—must be transmitted by electromagnetic trans- mission from location A to location B. However, the data sent over the transmission channel are subject to a channel noise disturbance and so to reduce the possibility of error, the value 2 is sent over the channel when the message is 1 and the value -2 is sent when the message is 0. If *x* is the value sent at location A then R, the value received at location B, is given by *R* = *x* + *N*, where *N* is the channel noise disturbance. When the message is received at location B, the receiver decodes it according to the following rule:

if *R*

if *R*< .5, then 0 is concluded

Assuming the channel noise *N* follows a standard normal distribution, determine the error probabilities of incorrectly determining a 0 if a 1 was sent and incorrectly determining a 1 if a 0 was sent, i.e. *P* (error*|*1 was sent) and *P* (error*|*0 was sent).

(4 marks) (3) Naval engineers believe that *W*, the amount of passenger weight (in units of 100 kgs) that a

cruise ship can tolerate without sinking is normally distributed with mean 500 and standard

deviation 45. Suppose that the weight (again, in units of 100 kgs) of a person is a random

variable with mean 1 and variance 0.001. How many people would have to be on the ship for the probability of sinking to be to be greater than .1587?

Hint: The problem can be converted to +…..+

Note: To help solve this you will need to use the fact that *x* = ( ) is the solution to the general quadratic equation

**EMBED IMAGES OF YOUR HAND-WRITTEN ANSWERS TO THIS TOPIC HERE**

**5. Inference:** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ….***Total: 6 marks***

(2 marks) (1) For sample *x* of size *n* distributed as the sum of squared errors (SSE) of mean estimate

Demonstrate using differentiation that the SSE point estimate , corresponding to the value of that minimises the SSE, is equivalent to the sample mean

(2) A company self-insures its large fleet of cars and trucks against collision. They think their trucks cost more to repair than their cars but aren’t sure, so take a random sample of 30 car and 30 truck accidents. Suppose that for cars and trucks the average repair costs for these accidents is

$4,300 and $4,900 and the sample standard deviations are $1,100 and $1,300, respectively.

(2 marks) (a). Give a 95% two-sided confidence interval for the difference between the means for cars and trucks. State your assumptions clearly.

(2 marks) (b). Form a hypothesis test to determine if the mean repair costs for cars and trucks are equal at the significance level of *α* = 0*.*05. State your assumptions clearly. Is there a statistically significant difference in the means for cars and trucks?

**EMBED IMAGES OF YOUR HAND-WRITTEN ANSWERS TO THIS TOPIC HERE**

**6. Simulation:** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ….***Total: 10 marks***

(1) Suppose we have a probability distribution function (PDF) defined as follows:

We can readily calculate the area under this curve knowing that Recall also that the inverse of the *sin* and *cos* functions are *arcsin* and *arccos* functions.

(3 marks) (a). Using appropriate equations and/or pseudocode, show how Inverse Transform Sampling could be used to re-sample from this distribution. Please sketch/draw all relevant functions.

(3 marks) (b). Using this same example, now show how Rejection Sampling could be used to re-sample from this distribution, again using appropriate equations or pseudocode. Please use a diagram to show how random samples are rejected or accepted.

1. Consider the simple Bayesian network (causal graph) below. Assume this network correctly models the relationships between A, B, C, D, and E



Figure 1: Simple Bayesian network

(2 marks) (a). Write a factorised probability expression for the full joint distribution *p*(*A, B, C, D, E*)*.* Please order the terms so that a sequential sampler could be used to sample each conditional probability distribution in turn reading from left to right.

(2 marks) (b). Sequential sampling of the full joint probability distribution in Part (a) would need to be performed differently if nodes are observed or known, since this changes the conditional probability relationships in the network. In this case Gibbs Sampling can be used. If the network described above was conditioned on A = what four (4) conditional Probability distribution functions would one need to be able to calculate in order to perform this type of sampling? Please write out how you would compute them.

**Hint: you don’t need to show how to compute the normaliser. Only proportional relationships are required for this.**

**EMBED IMAGES OF YOUR HAND-WRITTEN ANSWERS TO THIS TOPIC HERE**

**7. Regression:** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ….***Total: 8 marks***

Below is a table of the number of cricket chirps as a function of outdoor temperature



(4 marks) (1) Build a simple linear regression model to predict the number of cricket chirps as a function of outdoor temperature, and complete your solution by giving the linear prediction formula. The key statistics for data in this table are



(4 marks) (2) What is the co-efficient of determination, *R*2, of this model? Based on this *R*2 value, is this a good model of the data? What is the reference model used in the calculation of *R*2 in this case?

**EMBED IMAGES OF YOUR HAND-WRITTEN ANSWERS TO THIS TOPIC HERE**

**8. Modelling*………………………………………………………………………Total: 10 marks***

1. You have been hired to develop a machine learning classifier system for a hospital emergency room. The hospital would like to predict the likelihood of success of surgical procedures based on a large array of medical and other relevant variables. For the purposes of an initial prototype, a smaller subset of just 4 summary variables will be considered. You know that different techniques might be used for such a problem, each with its relative strengths and weaknesses. The following *target* and *predictor* variables will be used:

**Outcome** - A 3-valued target variable, where S = Success, F = Failure and I = Indeterminate.

**Difficulty** - A 2-valued predictor variable summarising the difficulty of the procedure, where S

= Standard and C = Complex.

**Qualification** - A 2-valued predictor variable indicating the skill and expertise of the surgeon, where N = Novice and E = Expert.

**Constitution** - A 3-valued predictor variable summarising the patient’s overall health, where S

= Satisfactory, F = Fair and C = Critical.

(3 marks) (a). For this question you may choose **one** of the following data modelling and classification

techniques:

* + 1. Full Bayes classifier
    2. Naïve Bayes classifier
    3. Logistic Regression

Given the **three** possible Outcome levels, write down the required formulae for your selected method for the case where Outcome = S.

# Don’t forget to show how to calculate any normalising constants or discretised predictor variables in your answer.

# (3 marks) (b). Now provide a brief justification for your choice of machine learning classifier for this

application. In your answer, please compare the approach which you recommend to the other two methods, identifying key advantages and disadvantages of each.

**THERE ARE MORE QUESTIONS ON THE NEXT PAGE**

(2) The four panels in the figure below show some noisy 2-D data clusters which we wish to model with the K-means algorithm. Initial proposed centroids or cluster centers are marked with *a*, *b*,

*c*. (Note that the *K* value is different in each panel.) Assume that a Euclidean distance measure is used to determine distance between points.



(2 marks) (a). Redraw the numbered panels above (not including the individual sample points) and use the operation of the K-means algorithm to draw the *first step* in the movements of the centroids *a*, *b* and *c*. Also draw the approximate outline or boundaries of the *final data clusters* in each case.

(2 marks) (b). Which of the four panels do you think has the best starting configuration or position and which do you think has the worst? Provide brief reasons for your observations.

**EMBED IMAGES OF YOUR HAND-WRITTEN ANSWERS TO THIS TOPIC HERE**