CITY UNIVERSITY OF HONG KONG

Course code & title : EE3001 Foundations of Data Engineering

Session : Semester A 2020/21

Time allowed : 2 hours

This paper has **SIX** pages (including this cover page).

Please make sure you follow all instructions from the University, ARRO, and EE. Please note the following:

1. Write your name and your student ID on the top of every answer sheet.
2. This paper consists of **4** questions. The questions are ALL compulsory. Make sure that you attempt all of them. The total score is 100.
3. This is an **open-book exam.** Students can read the lecture notes and/or other materials available online.
4. You are responsible for receiving the questions on Canvas, hand-write all answers on blank answer sheets, compile the answers into a single PDF file, and **upload the file before the deadline** of the exam.
5. **Stay on Zoom after the deadline** until the examiner allows you to leave.

Answering this exam paper implies your acknowledgment of the Pledge for following the Rules on Academic Honesty:

"I pledge that the answers in this examination are my own and that I will not seek or obtain an unfair advantage in producing these answers. Specifically,

1. I will not plagiarize (copy without citation) from any source;
2. I will not communicate or attempt to communicate with any other person during the examination; neither will I give or attempt to give assistance to another student taking the examination; and
3. I will use only approved devices (e.g., calculators) and/or approved device models.
4. I understand that any act of academic dishonesty can lead to disciplinary action."

On the first page of your answer sheets, copy the following sentence and sign it: *I pledge to follow the Rules on Academic Honesty and understand that violations may lead to severe penalties.*

Signature _____ Date _____

Name _____ Student ID _____

Contact Information

- Should you have any technical problem during the exam, contact your course leader or invigilator via Zoom private chat, email: kelviny.ee@cityu.edu.hk or by phone call at 3442 7717.
- If you are not able to contact course leader/invigilator, you can reach the department via:
  (a) Departmental hotline at (+852) 3442-7740
  (b) Department Whatsapp phone: 9269-4066
  (c) Department WeChat ID: wxid_lly7yf5fzoj722 or scan the following QR Code

Do this examination paper if the last digit of your student ID is <u>odd</u>

Qn 1 (25 marks)

a) It is important to develop reliable tests for diseases. A test for a disease has a false negative rate of 5%. It is known that 0.1% of the population has the disease. If it is desired that when the test gives a positive result, the probability that the patient has the disease is at least 90%, what should be its false positive rate in % ? (5 marks)

b) Assume the distribution to be independent and identically distributed. Let the event be that the sample mean deviates from the true mean by less than 1 standard deviation. Let the probability of this event occurring to be larger than 0.99.

Calculate the number of samples required. (5 marks)

c) The time in years for which a server will function before breaking down is modelled by an exponential distribution. The expected value is 10 years. 3 years have passed since the server starts operation and it is still working normally. What is the probability that the server is still working at the end of the 7^{th} year? (5 marks)

d) A player has a probability of winning of 0.2. How many matches is he expected to play before winning his 7^{th} match? What implicit assumption(s) have you made in your calculations? (5 marks)

e) It is suspected that in a box of 50 components, there are 2 faulty components. Suppose we draw without replacement 2 components out of the box, what is the probability that the fault is discovered? (5 marks)

Qn 2 (25 marks)

The price of two stocks A and B in six consecutive days are as follows:

Day	**1**	**2**	**3**	**4**	**5**	**6**
A	2.0	2.5	2.7	2.3	2.1	1.9
B	3.7	3.3	3.0	3.2	3.5	5.8

a) Compute the Pearson coefficient. (4 marks)

b) Test the hypothesis that the two stocks are linearly related. Compute the p-value. Only a best estimate (e.g. an inequality) is needed.

State the two underlying assumptions.

(6 marks)

c) Compute the Spearman coefficient. (5 marks)

d) If you wish to test the hypothesis that the two stocks are monotonically related, compute the T statistic. (4 marks)

e) Predict the price of stock A at the 7^{th} day using linear regression. (3 marks)

f) Explain why it is not a good idea to fit a curve by making the coefficient of determination zero. Then suggest a better method. (3 marks)

Qn 3 (25 marks)

Two stochastic algorithms *A* and *B* have the following performance when finding the minimum of engineering problems. Each algorithm is run 5 times. Since the algorithm is stochastic, each run gives a different minimum.

Performance of *A* in 5 runs: 15 18 20 10 5

Performance of *B* in 5 runs: 20 14 24 15 25

As the minimum is required, the lower the number, the better is the performance of the algorithm.

You may assume that the 10 runs are independent from each other and identically distributed.

In the following, you are required to choose the correct hypothesis test to use:

a) Assume that the performance of both algorithms is Gaussian distributed, test the hypothesis that comparing average performance, *A* is no different to *B* at level of significance 0.1. (10 marks)

b) If we do not assume that the algorithms are Gaussian distributed and instead assume that the distribution is unknown, test the hypothesis that A is better than B by computing the p-value.

Hint: You may use normal approximation for the test you are using. (10 marks)

c) Explain the theoretical support behind the normal approximation in b). (5 marks)

Qn 4 (25 marks)

a) The expected age of death of a population is 85 years. Compute the upper bound of the probability that a person lives to 120 years? Write down the name of the inequality you are using. (5 marks)

b) If it is additionally known that the standard deviation is 6 years, compute the upper bound of the probability that a person lives to 120 years. Write down the name of the inequality you are using. (5 marks)

c) Two candidates run for an election. Which candidate wins is determined by majority vote (i.e., who gets more votes wins). 13 polls are made about the support rating of a candidate. The results in % is as follows:

52 50 48 40 60 51 52 49 54 53 50 51 49

Set up a hypothesis test to compute the p-value that the candidate will win the election.

This question requires you to choose a test that makes the fewest assumptions possible.

State any assumption(s) that you must nevertheless make.

(10 marks)

d) We have four brands of gasoline A, B, C, D. We wish to test whether their performance is identical. We use four different cars. Each car is loaded with a different brand of gasoline. For each car, we record the distance covered from full tank to empty tank and repeat the experiment 10 times. Assume the probability distribution is unknown.

What statistical test would you suggest? In this test, what is the assumption made for the null hypothesis? (5 marks)

--- END ---