UNIVERSITY OF BRISTOL

January 2019 Examination Period

FACULTY OF ENGINEERING

First Year Examination for the Degree of
Bachelor of Science / Master of Engineering

COMS10011
Probability and Statistics

TIME ALLOWED:
2 hours

This paper contains *two* parts.
The first section contains *20* short questions.
Each question is worth *two marks* and all should be attempted.
The second section contains *three* long questions.
Each long question is worth *25 marks*.
The best *two* long question answers will be used for assessment.
The maximum for this paper is *90 marks*.

Other Instructions:

Calculators must have the Faculty of Engineering Seal of Approval.

TURN OVER ONLY WHEN TOLD TO START WRITING

Section A: short questions - answer all questions

Q1. What is the sample space for the result of rolling a three-sided dice twice?

Q2. What is an event?

Q3. A sample space consists of the words {and,the,fruit,of,that,forbidden,tree}. If all words are equally likely what is the event that the word has three or fewer letters.

Q4. A sample space consists of the words {and,of,the,fruit,of,that,forbidden,tree}. The event *A* consists of all words with three or fewer letters. The event *B* consists of all words ending in a vowel. What is $p(B)$ and $p(B|A)$?

Q5. What is the 'naïve' aspect of the naïve Bayes estimator?

Q6. What is a random variable?

Q7. There are two factories that make widgets. Factory A makes an equal number of red and green widgets. Factory B makes only green widgets. Factory A makes three times as many widgets as factory B. You select a widget at random. it is green. How likely is it your widget came from factory A?

Q8. Define the moment generating function.

Q9. What distribution is satisfied by the sum of two Gaußian variables?

Q10. The radius of sunflowers *R* has an approximately Gaußian distribution with $\mu = 3$ cm and variance $\sigma^2 = 1$ cm^2. Write down an expression in terms of the error function for the probability of finding a flower bigger than 5 cm.

Q11. An Analysis of Variance (ANOVA) is a statistic test for comparing non-parametric data, true or false? Explain yourself in one sentence.

Q12. A study aims at comparing the performance of four computer graphic cards (A B C and D). A log of performance (number of frames per second) is gathered during a week for each of the four graphic cards. The researcher wants to use T-test in order to compare all the graphic cards. Using Bonferroni corrections, what is the new significance level that the researcher should use when looking for significant results when comparing each pair?

Q13. A politician claims that the dropout rate for schools is less than 25%. Last year, 190 out of 603 students dropped out. A researcher is aiming to looking for an evidence to reject the politician's claim, should he use a one tail or a two-tail statistical test and why?

Q14. Why is it important to check that the data is following a normal distribution before running statistical tests?

Q15. What are the two tests we can use to check if some data is following a normal distribution. In which cases should we use one or the other?

Q16. Explain in two to three sentences what is the different between causality and correlation

Q17. Below is a list of variables that might be measured in a research study:

1. Whether a person has a sibling, recorded as 'Yes' or 'No'.
2. A person's weight, recorded in kilograms.
3. How long a person was in school for, recorded as the number of years.
4. A person's income, recorded as 'under $10 000', '$10 000 - $50 000', '$50 000 - $100 000', 'over $100 000'.
5. The change in concentration of an enzyme in a person's blood, recorded as a percentage of the original.
6. The treatment group a person was in, recorded as 'Group 1', 'Group 2' and 'Group 3'

Write down whether each variable is categorical or numerical.

Q18. A study attempted to find out if the length of a person's legs had any relationship to their ability to play the popular mallet-based ball game croquet. The researchers took measurements of the legs of 104 professional croquet players, calculating the length of each player's legs. They also recorded the number of victories for each player in the last two croquet seasons. Choose the most appropriate procedure to decide if the number of victories has any relationship with the length of the legs:

A. Chi-squared test
B. Paired T-test
C. Unpaired T-test
D. Linear regression

Q19. In a weight loss program, 50 participants were randomized to two groups. One group were instructed to eat lunch before 2pm, and the other group were instructed to have lunch after 2pm. At the end of the program, the percentage weight loss for each patient was recorded. Choose the most appropriate procedure to decide if there is a relationship between the time of day a person eats lunch and their percentage weight loss:

A. Chi-squared test
B. Paired T-test
C. Unpaired T-test

D. Linear regression

Q20. In a study, 20 participants were sent to two rooms in which they were interviewed. Before the first interview, they were asked to assume closed posture such as crossing their arms and hunching their shoulders. Then, before the second interview, they were asked to assume open posture such as placing their hands behind the head or their feet on the table. The concentration of the stress hormone cortisol was measured for each patient after each interview. Choose the most appropriate procedure to decide if there is a relationship between posture and cortisol concentration:

 A. Chi-squared test

 B. Paired T-test

 C. Unpaired T-test

 D. Linear regression

Section B: long questions - answer two questions

Q1. This question is about calculating probability and about the binomial and Poisson distributions.

(a) A poetry magazine publishes 5% of the submissions it receives, unfortunately one of the editors for the magazine is very lazy and selects the successful submissions randomly without reading them, in other words, each poem this editor reviews has a one in twenty chance of being accepted, regardless of its merit. A poet submits eight poems and all are sent to the lazy editor. How likely it is that two poems from this poet are accepted. *[6 marks]*

(b) A poet writes an excellent poem and sends it to the same magazine. If the poem is read by a diligent editor it will have a 50% chance of being accepted, but, of course, if it is reviewed by the lazy editor it will only have a 5% chance. Three editors are diligent and one is lazy, they are each equally likely to review the poem. What chance is there that it will be accepted? *[4 marks]*

(c) If N satisfies a Poisson distribution with mean λ, what is $p_N(n)$? *[3 marks]*

(d) Derive the Poisson distribution for the probability of r events in a time T if the average number of events in T is λ. Remember it is important to show that λ is the mean. You may want to use the limit of infinitely frequent compounding:

$$\lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n = e^x$$

[12 marks]

Q2. This question has two sections, one about the central limit theorem and the other about experimental design.

(a) State the central limit theorem. *[7 marks]*

(b) You wish to design an experiment to investigate if taking caffeine impact memorization skills. You set out to make participants drink a coffee cup or not and then make them perform a test of memorization. Describe your experimental design, that is,

   (a) What are your independent and dependent variables *[4 marks]*

   (b) Are you doing a within or between experiment and why? *[6 marks]*

   (c) I you are using counterbalancing or not and why? *[4 marks]*

   (d) What is the task that the participants are going to do. *[4 marks]*

For the task, you can take inspiration from the task done in class with the memorization game.

Q3. This question is about Pearson's chi-square test. You are planning to buy a restaurant and the current owner claims having a good model of his clients and promises you that you will get the following visit: on Monday 10% of the clients, Tuesday 10%, Wednesday 15%, Thursday 20%, Friday 30%, Saturday 15% (they are close on Sunday). You come for a week and gather the following observations: on Monday you see 30 clients, Tuesday 14, Wednesday 34, Thursday 45, Friday 57, Saturday 20. Using the Pearson's chi-square test you want to check if the current owner is telling the truth, that is, if his model fits with your observations.

(a) Compute, for each day, the expected numbers of visits using the owner model *[5 marks]*

(b) What is the chi square formula? *[5 marks]*

(c) Compute the chi square value using the observed and expected data *[5 marks]*

(d) How many degrees of freedom do we have in this experimental setup? *[5 marks]*

(e) Using the table below and a significance value of 0.05, conclude the analysis, that is. can you trust the owner model and why? *[5 marks]*

(cont.)

Degrees of freedom (df)	x^2 value										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.63	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

END OF PAPER