UNIVERSITY OF BRISTOL

September 2019 Examination Period

FACULTY OF ENGINEERING

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

COMS10011-R
Probability and Statistics

TIME ALLOWED:
2 hours

Answers to COMS10011-R: Probability and Statistics

Intended Learning Outcomes:

Section A: short questions - answer all questions

Q1. A bag contains one red ball, 1000 blue balls and 1000 yellow balls; two balls are taken from the bag. What is the sample space?

Solution:

$$\{YY, BB, YB, BY, YR, BR, RY, RB\}$$

Q2. What is an event?

Solution: It is a subset of the sample space.

Q3. Ireland is divided into four provinces: Munster, Leinster, Ulster and Connaught. Two provinces are selected at random; what is the event that both end in 'ster' and what is the probability of this event?

Solution:

$$\{ML, LM, MU, UM, UL, LU\}$$

and

$$1/2$$

Q4. A sample space consists of the words {stately, plump, buck, mulligan, came, from, the, stairhead}. The event A consists of all words with four or fewer letters. The event B consists of all words ending in a vowel. What is $p(B)$ and $p(B|A)$?

Solution:

$$p(B) = 1/4$$

and

$$p(B|A) = 1/2$$

also allow answer that included "y" as a vowel.

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

Solution: It assumes conditional independence of the evidence

$$p(\mathbf{w}|s) = p(w_1|s)p(w_2|s)\ldots p(w_n|s)$$

Q6. If X is a random variable define the expected value of X?

Solution:

$$\langle X \rangle = \sum_i x_i p(x_i)$$

Q7. A shopkeeper says she has two new baby beagles to show you, but she doesn't know whether they're male, female, or one of each. You tell her that you want only a male, and she telephones the fellow who's giving them a bath. "Is at least one a male?" she asks him. "Yes!" she informs you with a smile. What is the probability that the other one is a male? (This question is taken from Marilyn von Savant's Ask Marilyn column).

Solution:

$$1/3$$

Q8. A six sided dice is rolled ten times, what is the probability of getting exact three sixes?

Solution:

$$p = \begin{pmatrix} 10 \\ 3 \end{pmatrix} \frac{1}{6^3} \frac{5^7}{6^7}$$

Q9. Define the moment generating function.

Solution: For a random variable X

$$m(t) = \langle e^{tX} \rangle$$

Q10. The weight of the eggs laid by a particular chicken, W, has an approximately Gaußian distribution with $\mu = 50$ g and variance $\sigma^2 = 25$ g^2. Write down an expression in terms of the error function for the probability of finding a egg bigger than 65 g.

Solution:

$$z = \frac{x - 50}{5\sqrt{2}}$$

and

$$2P(W > 65) = \text{erf}\,(\infty) - \text{erf}\,(3/\sqrt{2}) = 1 - \text{erf}\,(3/\sqrt{2})$$

[no need to know erf of infinity is one for full marks]

Q11. A Friedman test is a statistic test for comparing parametric data, true or false? Explain yourself in one sentence.

Solution: (Answer: False this is used for comparing non-parametric data, i.e. with a skewed distribution)

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 a 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?

Solution: (Answer: 0.005/6 because there are 6 comparisons made)

Q13. Five-point Likert scales (strongly disagree, disagree, neutral, agree, strongly agree) are frequently used to measure motivations and attitudes. A Likert scale is a:

A Independent variable.

B Ordinal variable.

C Continuous variable.

D All of the above options (A, B and C)

Solution: D

Q14. Which statistical test is used to identify whether there is a relationship between two categorical variables?

A Student's t-test.

B Spearman's correlation test.

C Pearson's Chi-square test.

D Mann-Whitney test.

Solution: Pearson's Chi-square test

Q15. Explain why establishing correlation does not suffice to demonstrate causation with one example (e.g. the example seen in class).

Solution: Example seen where we saw a correlation of the increase of murder and the consumption of ice cream in summer month. One does not prove the other and there are many confounding variables, e.g. the weather temperature increasing which might be another reason why ice cream consumption increase.

Q16. In a survey about people's relationship with their pets, one of the questions asked if the person felt guilty when they left their pets alone when they went on holidays. The response for each person was recorded as 'Yes - feels guilty' or 'No - does not feel guilty'. The researchers were interested to see if men and women felt differently in this situation. Choose the most appropriate procedure to decide if the feeling of guilt about leaving pets alone is different for men and women:

A Chi-squared test for independence

B Paired T-test

C Unpaired T-test

D Linear regression

Solution: A

Q17. In weight loss program, 50 participants were randomised to two groups. One group were instructed to eat lunch before 2pm, and the other group were instructed to have lunch after that time. 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 for independence

B Paired T-test

C Unpaired T-test

D Linear regression

Solution: C

Q18. Researchers were interested in the care of head trauma patients while in hospital. From the hospital records, they recorded information about 400 patients across several hospitals. Choose the most appropriate procedure to decide if the data is following a normal distribution.

A Chi-squared test

B T-test

C Shapiro-Wilk test

D Kolmogorov-Smirnov test

Solution: D - Kolmogorov-Smirnov test is more appropriate for large sample size N ¿50 and Shapiro-Wilk for small sample size N ¡=50)

Q19. Five students take the machine learning unit in one year and the robotic unit the next year. Their overall course grades (max being 100) are listed below for both courses. Which of the following statistical procedures would be most appropriate to test the claim that student overall course grades are the same in both courses? Assume that any necessary normality requirements hold.

student	1	2	3	4	5
ML	80	72	99	91	75
R	85	71	93	93	75

A Two-tailed two-sample paired t-test of means

B Two-tailed two-sample unpaired t-test of means

C One-tailed two-sample paired t-test of means

D One-tailed two-sample unpaired t-test of means

Solution: A, two-tails because we want to compare in both directions, paired because the students did both course

Q20. A study attempted to find out if the length of an animal genome had any relationship to their life expectancy. The researchers took the data of 300 species of animal, calculated their age in days and ran a test to measure the length of their genome. Choose the most appropriate procedure to decide if the age has any relationship with the run speed:

A Chi-squared test

B Paired T-test

C Unpaired T-test

D Linear regression

Solution: D

Section B: long questions - answer two questions

Q1. This question is about calculating probabilities and conditional probabilities in the context of the dice game craps.

(a) A pair of fair six-sided dice is thrown. What is the probability their values will sum to seven? *[4 marks]*

(b) The game of craps involves repeated rolls of a pair of dice; it is a gambling game and participants can bet on different outcomes. For this question we will concentrate on the outcome called 'pass'. There are two phases to the pass game, the first is a single roll called a 'come out' roll. In the come out roll the player wins if the values sum to seven or 11, they lose if the values sum to two, three or 12; otherwise they progress on to the next phase of the game. What is the probability of these three possibilities? *[6 marks]*

(c) From now on we will call the summed value of the two dice 'the value of the roll'. If the value in the come out roll is not two, three, seven, 11 or 12, that is if the game progresses on to the next phase, the come out value is called the 'point'. The pair of dice will now be rolled repeatedly until the roll either gives the value seven, or its value is equal to the point. If it gives the point the player wins, if it gives seven the player loses. As an example, say in the come out round the value is six; six is now the point and the dice are rolled again, say in the next round the value is five, this is equal to neither the point nor seven so the game continues, in the next round a two is rolled and, again, the dice are rolled again, at which point the value is seven, so the game ends with a loss for the player. If the value of the come out roll is four, what is the probability the player will win? It might be useful to think of this as $P(\text{value} = 4 | \text{value is 4 or 7})$ *[6 marks]*

(d) If

- Q_i represents the chance of winning if the point is i, so, for example, the probability calculated above if Q_4
- P_i is the probability the point is i, so, for example, $P_4 = 1/12$

what is the probability of winning written in terms of Q_i and P_i? *[4 marks]*

(e) What is the probability of winning? *[5 marks]*

Solution: a) So there are 36 possible pairs of face values, each equally likes, of these 6+1, 5+2, 4+3, 3+4, 2+5 and 1+6 add to seven, so the probability of rolling a six is 1/6.
b) win is 1/6+1/18 since there are two ways to may 11: 5+6 and 6+5, this is 4/18=2/9. lose is 1/36+1/18+1/36 which is 1/9 and by subraction this leave 2/3 as the chance of going on to the next phase of the game. c) So using Bayes rule

$$P(4|4\|7) = \frac{P(4\|7|4)P(4)}{P(4\|7)}$$

with $P(4) = 3/36$ and $P(7) = 6/36$ and $P(4\|7|4) = 1$ hence

$$P(4|4\|7) = 3/(3+6) = 1/3$$

d) There are two ways of winning, either from the come out roll or a subsequent roll, which must be summed over the possible point values.

$$P(\text{first roll is 11 or seven}) + \sum_{4,5,6,8,9,10} P_i Q_i = \frac{2}{9} + \sum_{4,5,6,8,9,10} P_i Q_i$$

e) So

$$\frac{2}{9} + \frac{1}{3}\frac{1}{12} + \frac{2}{5}\frac{1}{9} + \frac{5}{11}\frac{5}{36} + \frac{5}{11}\frac{5}{36}\frac{2}{5}\frac{1}{9} + \frac{1}{3}\frac{1}{12} \approx 0.49$$

Q2. This question has two sections, one about Poisson processes and the other about experimental design.

(a) This is the section about Poisson processes.

(a) Write down the formula $p(n)$, the probability there will be exactly n buses in an hour. *[3 marks]*

(b) Buses arrive at a bus stop at a rate of five every hour. What is the chance of waiting for an hour without a bus arriving? *[3 marks]*

(c) What is the probability two or fewer buses arrive? *[3 marks]*

(d) Show $\sum_{n=0}^{\infty} p_n = 1$. *[3 marks]*

(e) You have designed a study to figure out if the type of movie children is watching make a difference in the number of snacks they will eat. A group of 50 children were randomly assigned to watch either a cartoon or a live action musical (25

to each). Crackers were available in a bowl, and the investigators compared the number of crackers eaten by children while watching the different kinds of movies. Please answer the following questions.

(a) What are your independent and dependent variables? *[3 marks]*

(b) Is this study a within or between experiment and why? *[3 marks]*

(c) In the study described above, one kind of movie was shown at 8 AM (right after the children had breakfast) and another at 11 AM (right before the children had lunch). It was found that during the movie shown at 11 AM, more crackers were eaten than during the movie shown at 8 AM. The investigators concluded that the different types of movies had an effect on appetite. The results cannot be trusted because:

- the study was not double blind. Neither the investigators nor the children should have been aware of which movie was being shown.
- the investigators were biased. They knew beforehand what they hoped the study would show.
- the investigators should have used several bowls, with crackers randomly placed in each.
- the time the movie was shown is a confounding variable.

Pick one possibility and explain. *[3 marks]*

(d) What you would do to adapt the study to avoid the problem exposed in the previous question while still keeping the time of the day as a factor of the experiment? Describe your experimental design? *[4 marks]*

Solution: a)
a) $p_n = (5^n/n!)e^{-5}$
b) e^{-5}
c) $e^{-5} + 5e^{-5} + 25/2e^{-5}$
d) So this follows from $e^x = \sum x^n/n!$.
b)
a) dependant = number of crackers eaten / independent = the type of movie
b) it is a between-subjects experiment as 25 children do one condition and the other 25 another one.
c) the time the movie was shown is a confounding variable. The change in number of crackers could have been both a reason of the type of movie or the time of the day.
d) the time of the day can be used as a new independent variable. There are multiple answers possible. The experiment would thus have 2 factors: one is the type of movies; one is the time of the day. These factors need to receive the same amount of sample data. One way would be to keep the study as between-subject for the type of movies and to invite the children another day at a different time of the day to do this experiment again. In such case the investigator would have to make sure the movies shown are different but are still comparable

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 according to the owner's 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]*

Degrees of freedom (df)	**χ^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

Solution: a) We have 200 observations, so the expected values are:
Monday 200*10% = 20 Tuesday 20 Wednesday 30 Thursday 40 Friday 60 Saturday 30
b) $\chi^2 = \sigma[(\text{observed} - \text{expected})^2/\text{expected}]$
c) $\chi^2 = 11.44$
d) $Df = 5.(6\text{observations minus} - 1)$
e) By looking in the table at the intersection of df = 5 and p = 0.05, we find the value

11.07. $\chi^2 > 11.07$, so we reject the null hypothesis, i.e. the model from the owner is not a good fit.

END OF PAPER