UNIVERSITY OF BRISTOL

January 2019 Examination Period

FACULTY OF ENGINEERING

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

COMS10011-J
Probability and Statistics

TIME ALLOWED:
2 hours

Answers to COMS10011-J: Probability and Statistics

Intended Learning Outcomes:

Section A: short questions - answer all questions

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

Solution:

$$\{11, 12, 13, 21, 22, 23, 31, 32, 33\}$$

Q2. What is an event?

Solution: It is a subset of the sample space.

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.

Solution:

$$3/7$$

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 $p(B|A)$?

Solution:

$$p(B) = 2/7$$

and

$$p(B|A) = 1/3$$

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. What is a random variable?

Solution: It is a map from the sample space to the non-negative real numbers such that

$$X(A \cup B) = X(A) + X(B)$$

if $A \cap B = \emptyset$ and $X(S) = 1$ where S is the whole sample space.

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?

Solution: We want $p(A|G)$ with

$$p(A|G) = \frac{P(G|A)P(A)}{P(G)}$$

Now $P(G) = P(G|A)P(A) + P(G|B)P(B) = 0.5 * 0.75 + 0.25 = 0.875$ and $P(G|A)P(A) = 0.375$ hence

$$p(A|G) = \frac{0.375}{0.875} \approx 0.43$$

Q8. Define the moment generating function.

Solution: For a random variable X

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

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

Solution: Another Gaußian.

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.

Solution:

$$z = \frac{x - 3}{\sqrt{2}}$$

and

$$P(R > 5) = \text{erf}\,(\infty) - \text{erf}\,(\sqrt{2}) = 1 - \text{erf}\,(\sqrt{2})$$

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

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

Solution: False ANOVA only works on parametric data, i.e. following a normal 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 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: 0.005/6 because there are 6 comparisons made = 0.0083

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?

Solution: one-tail is enough because the researcher is only interested in one direction. i.e. the dropout rate below 25%. If the question was 'equal to' we would use a two-tail test

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

Solution: because certain statistical tests assume the normality of data to compute p_values, e.g. t-test and Anova. If the assumption of normality is rejected, one should use non-parametric 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?

Solution: Kolmogorov Smirnov test for large sample size $N > 50$ and Shapiro-Wilk for small sample size $N <= 50$

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

Solution: Correlation is a measure that describes the relationship between two or more variables. A correlation does not mean that the change in one variable is the cause of the change in the values of the other variable. Causation indicates that one event is the result of the occurrence of the other event.

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.

Solution: 1, 4, 6 are categorical; 2, 3, 5 are 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

Solution: D

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

Solution: C

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

Solution: B

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]

Solution: a) So this is

$$\left(8//2 \right) 0.05^2 0.95^6$$

no need for the numerical value. [6 for correct, 2 for garbled version]
b) This is $0.75 * .5 + 0.25 * 0.05 = 0.3875$ [4 for correct, 1 for attempt]
c)

$$p_N(n) = \frac{\lambda^n}{n!} e^{-\lambda}$$

d)
Now if we are interested in the probability distribution for the number of events in an interval $T = n\delta t$, then, by the binomial distribution the probability of r events is

$$p(r) = \left(\begin{array}{c} n \\ r \end{array} \right) p^r (1-p)^{n-r} \tag{1}$$

Now write $\delta t = T/n$ and consider the $n \to \infty$ limit:

$$p(r) = \lim_{n\to\infty} \left(\begin{array}{c} n \\ r \end{array} \right) p^r (1-p)^{n-r} \tag{2}$$

Since p is the probability of an event in the small interval, it will become tiny as n becomes large, so, to deal with quantities that remain useful in the limit, let $\lambda = np$. Subsituting this in, and expanding out the binomial coefficient:

$$p(r) = \lim_{n\to\infty} \frac{n(n-1)(n-2)\ldots(n-r+1)}{r!} \left(\frac{\lambda}{n}\right)^r \left(1 - \frac{\lambda}{n}\right)^{n-r} \tag{3}$$

As n gets large the numerator of the first fraction just looks like n^r and cancels with the denominator of the second fraction. Recall that

$$\lim_{n\to\infty} \left(1 - \frac{x}{n}\right)^n = e^{-x} \tag{4}$$

The $(1 - \lambda/n)^{n-r}$ term has an extra $-r$ but

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-r} = 1 \tag{5}$$

Putting all this together we get

$$p(r) = \frac{\lambda^r}{r!} e^{-\lambda} \tag{6}$$

First, it is easy to check that the probabilities add to one; but notice that this range of the random variable is infinite! Using the Taylor expansion of the exponential:

$$e^x = \sum_{r=0}^{\infty} \frac{x^r}{r!} \tag{7}$$

we have

$$\sum_{r=0}^{\infty}\frac{\lambda^r}{r!}e^{-\lambda}=e^{-\lambda}\sum_{r=0}^{\infty}\frac{\lambda^r}{r!}=e^{-\lambda}e^{\lambda}=1 \tag{8}$$

Next consider the mean

$$\mu=\sum_{r=0}^{\infty}r\frac{\lambda^r}{r!}e^{-\lambda} \tag{9}$$

Because of the r in the summand, the $r=0$ term is zero, so

$$\mu=\sum_{r=1}^{\infty}r\frac{\lambda^n}{r!}e^{-\lambda} \tag{10}$$

Now, use $r! = r\times(r-1)!$:

$$\mu=\sum_{r=1}^{\infty}\frac{\lambda^r}{(r-1)!}e^{-\lambda} \tag{11}$$

and then pull a λ out the front

$$\mu=\lambda\sum_{r=1}^{\infty}\frac{\lambda^{r-1}}{(r-1)!}e^{-\lambda} \tag{12}$$

Finally set $s=r-1$ and

$$\mu=\lambda\sum_{s=0}^{\infty}\frac{\lambda^s}{s!}e^{-\lambda}=\lambda \tag{13}$$

so λ is the average event count!

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) If 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.

Solution: a) For X_1 to X_n i.i.d. [1] then if

$$S_n = \frac{1}{n}(X_1 + X_2 + ... + X_n)$$

[1] and

$$U_n = \frac{S_n - \mu}{\sqrt{n}\sigma}[2]$$

where μ and σ are the mean and variance of X [1]

$$U_n \sim \mathcal{N}(0, 1)$$

[2] b)

a) Dependent variable = memorisation score, i.e. number of numbers participants can remember Independent variable = has drink coffee or has not drink coffee

b) Between-subject. In a within-subject experiment, participants starting with the 'drink coffee' condition and then doing the 'no drink' condition will lead to biased data as they will have caffein in their system when doing the 2nd condition. We thus need to use a between subject experiment in which we take two distinct groups of participants, one group will drink coffee, the other one will not.

c) No counterbalance needed if you are doing a between subject-experiment with only one Independent variable.

d) The participants are asked to first drink (or not) a coffee. An audio recording of a series of number is played. At the end the participants are asked to write the series of number they can remember on a piece of paper and in the correct order. The length of the series of number is increasing in length (starting at one). If the participants make a wrong guess the experiment stop. The length of the last correct series guessed is recorded as memorisation score (dependant variable).

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]*

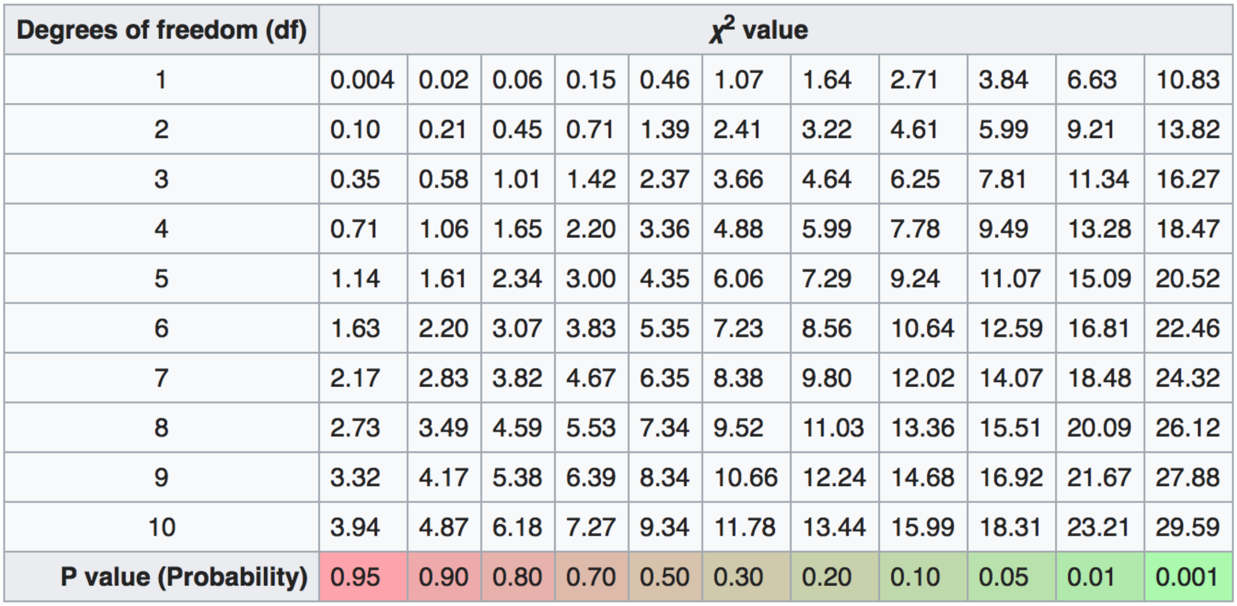
(b) What is the chi square formula? *[5 marks]*

(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]*

(cont.)

(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) χ^2 = 11.44
d) *Df* = 5.(6observations 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