The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS

Spring Semester 2019-2020

MAS5050 Mathematical Methods for Statistics
MAS5051 Probability and Probability Distributions
MAS5052 Basic Statistics

Nominal 2 hours in a 24 hour period

This is an open book exam.
Answer ***all*** *questions.*
The submission deadline is 10 am (BST), twenty-four hours after it is released. Late submission will not be considered without extenuating circumstances. It is expected that you will be able to complete this exam in approximately two hours and it is recommended that you submit the work within four hours. You will not be penalised for taking longer, however.
Unless it is explicitly stated otherwise, it is intended that calculations are performed by hand (possibly with the aid of a calculator). To gain full marks, you will need to show your working. You will not get full marks if you simply write down output from a computer package.
By uploading your solutions you declare that your submission consists entirely of your own work, that any use of sources or tools other than material provided for this module is cited and acknowledged and that no unfair means have been used. Open book examination. Candidates may use the lecture notes and associated lecture material (including set textbooks), plus a calculator. Candidates should attempt ***all*** *questions and are required to show necessary computation which allowed them to deduce their solution. The maximum mark for the various parts of each question is indicated.*

Section A: MAS5050

1 The numbers p, 10 and q are three consecutive terms of an arithmetic series. The numbers p, 6 and q are three consecutive terms of a geometric series. Show that $p^2 - 20\text{p}{+}36 = 0$ and hence find the values of p and q for which the geometric series converges. ***(5 marks)***

2 (i) Compute the derivative of $r(x) = \tan\dfrac{1}{x}$ with respect to x. ***(2 marks)***

(ii) Find a unit vector perpendicular to the plane that contains the vectors $\mathbf{a} = \mathbf{i} + 3\mathbf{j} + 4\mathbf{k}$ and $\mathbf{b} = 2\mathbf{i} + \mathbf{j} + 3\mathbf{k}$. ***(3 marks)***

3 Let

$$A = \begin{pmatrix} 5 & -3 \\ -6 & 2 \end{pmatrix}.$$

Find a matrix, P, and a diagonal matrix, D, such that $P^{-1}AP = D$. ***(5 marks)***

4 The gamma function is defined as

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx,$$

where z is positive.

You are given that $\Gamma\left(\dfrac{1}{2}\right) = \sqrt{\pi}$ and that $\Gamma(z) = (z-1)\Gamma(z-1)$.

By making use of symmetry and by means of a suitable substitution, evaluate the integral

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx,$$

expressing your answer as a multiple of $\sqrt{\pi}$. ***(5 marks)***

5 Let R be the region $\{(x, y) : 1 \le x \le 2, \;\; 0 \le y \le 1\}$.
Evaluate the double integral

$$\int\int_R \frac{1}{x+y}\,dx\,dy,$$

giving your final answer correct to 3 decimal places. ***(5 marks)***

Section B: MAS5051

1 A drug is claimed to be 95% successful in treating a particular illness. It is known that the probability a patient recovers without aid of the drug is 0.1. A trial is conducted on 100 patients with the illness. A randomly chosen subset of 25 patients is given the drug, while the remaining patients take a placebo (dummy tablet). If the claim is correct, what is the probability that a patient received the drug, given that they recovered? ***(3 marks)***

2 The frequency of buses arriving at a bus stop in an hour period can be modelled using a Poisson distribution with rate $\lambda = 3$. In exact form, calculate the probability that at most three buses arrive in an hour interval. ***(2 marks)***

3 Let X be a random variable with probability density function

$$f_X(x) = \begin{cases} \dfrac{\sin(x)}{2} & \text{for } x \in [0, \pi], \\ 0 & \text{otherwise.} \end{cases}$$

Calculate the moment generating function $M_X(t)$. ***(6 marks)***

4 (i) An exponentially distributed random variable X with rate λ has cumulative distribution function

$$\begin{cases} 1 - e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{x}<0. \end{cases}$$

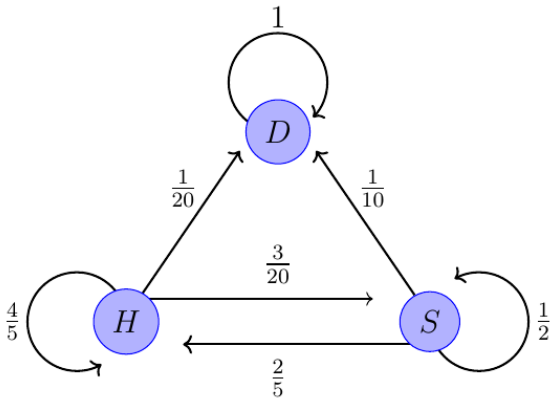
and obeys the Markov property given by

$$\mathbb{P}(X > s + t | X > s) = \mathbb{P}(X > t).$$

Prove X has this property. ***(5 marks)***

(ii) Explain briefly in your own words what it means for a random variable to possess the Markov property. ***(1 mark)***

5 Consider the following transition diagram representing a three state birth-death process with states {Healthy, Sick, Dead}.



Let $X_{n\in\mathbb{N}_0}$ be a random variable representing the state this process is in after n steps.

(i) Write down the transition matrix P for this diagram. ***(2 marks)***

(ii) Find $\mathbb{P}(X_1 = D|X_0 = S)$ ***(1 mark)***

(iii) Find $\mathbb{P}(X_6 = S|X_4 = H)$ ***(3 marks)***

(iv) Without calculation, write down the limit $\rho = (\rho_1, \rho_2, \rho_3)$ expressed by

$$\rho = \lim_{n\to\infty} P^n.$$

Explain why your answer makes sense in practical terms. ***(2 marks)***

Section C: MAS5052

1 Let $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ be two independent random samples with $\mathrm{E}[X_i] = \mu$, $\mathrm{Var}[X_i] = \sigma^2$, $\mathrm{E}[Y_j] = 3\mu$, $\mathrm{Var}[Y_j] = \sigma^2/2$.

(i) Prove that

$$T_1(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\bar{X} + \bar{Y}}{4}, \quad T_2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{3\bar{X} + \bar{Y}}{6} \quad \text{and} \quad T_3(\boldsymbol{X}, \boldsymbol{Y}) = \frac{n\bar{X} + 2m\bar{Y}}{n + 6m}$$

are unbiased estimators of μ. ***(6 marks)***

(ii) Explain (with justification) which of these estimators is best if $n = 2$ and $m = 22$. ***(7 marks)***

2 The relationship between annual average temperature over 10 years in various towns & cities and the area of the UK in which the town or city is located was investigated. Area is described as being one of five ordered categories A, B, C, D, E from north to south, with A being the most northerly (i.e. Scotland) and E being the most southerly.

(i) One investigator codes A, B, C, D, E as 1,2,3,4,5 respectively, and fits the regression $E(Y_i) = \mu_i = \beta_0 + \beta_1 x_i$, where Y_i is the annual average temperature and x_i the coded as area. Explain the interpretation of β_1. ***(2 marks)***

(ii) Another investigator thinks that A, B, C are more different than C, D, E and suggests using the codes 1,3,5,6,7 in the above regression. Interpret β_1 in this case. ***(2 marks)***

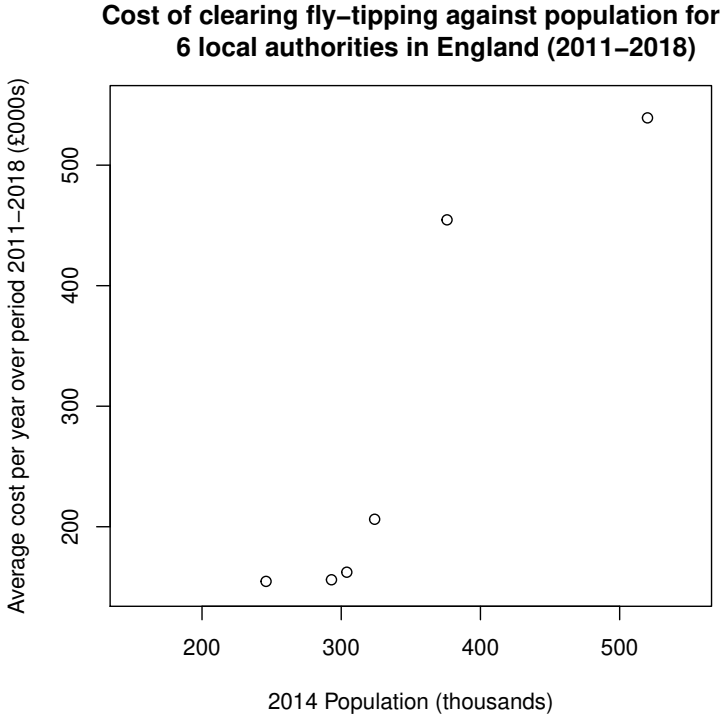
(iii) A third investigator says that as area is a categorical variable, one should use indicator variables, and fits the regression $\mu_i = \alpha_0 + \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + \alpha_3 x_{i,3} + \alpha_4 x_{i,4}$ where $x_{i,1}$ is an indicator variable for category A (i.e. $x_{i,1} = 1$ if observation i is from category A, and 0 otherwise), $x_{i,2}$ for category B, $x_{i,3}$ for category C, and $x_{i,4}$ for category D. Interpret the αs in this case. ***(3 marks)***

(iv) Which model would you use and why? ***(2 marks)***

3 In February 2020 the BBC reported that "Organised criminal gangs are dumping lorry-loads of rubbish across the UK as part of an illegal waste clearing service. The incidents are costly to clear and our analysis over has found councils have spent more than £59m on their removal since 2012." Taken from:

`https://www.bbc.co.uk/news/uk-england-50660138`

Local authorities are interested in what the cost is per person in their area. An hypothesis is that is may cost around £1 per person per year. The graph below illustrates data from six local authorities showing average yearly cost of removal over an eight year period (in thousands of pounds) against population (in thousands).



C3 (continued)

(i) Output from R obtained by fitting a least squares regression model to these data is given below. Draw (by hand) the relationship between the least squares regression line and the line $y = x$, where y represents cost and x represents population. *(4 marks)*

```
Call:
lm(formula = cost ~ population)

Residuals:
     1      2      3      4      5      6
-29.38 122.88 -40.00 -51.03 -39.22  36.76

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -286.7382   125.3360  -2.288  0.08406 .
population     1.6449     0.3532   4.657  0.00961 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 75.92 on 4 degrees of freedom
Multiple R-squared:  0.8443,Adjusted R-squared:  0.8054
F-statistic: 21.69 on 1 and 4 DF,  p-value: 0.00961
```

(ii) What does this tell you about the hypothesis of the cost to the authority being £1 per person per year? *(2 marks)*

(iii) Why might a regression line with zero intercept provide a more reliable way to estimate cost from population? *(2 marks)*

End of Question Paper