The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2016–2017**

MAS113 Introduction to Probability and Statistics **2 hours**

*Attempt **ALL** questions. The allocation of marks is shown in brackets. Total marks 60.*

1 Let S be the set $\{1, 2, 3, 4, 5\}$, and let P be a probability measure on S.

(i) You are given the information that $P(\{1, 2\}) = 0.3$, $P(\{2, 3\}) = 0.3$ and $P(\{3, 4\}) = 0.4$. For each of the following subsets of S, either give the measure under P of the set or state that there is not enough information given to calculate it.

(a) $A_1 = \{1, 2, 3, 4\}$; ***(1 mark)***

(b) $A_2 = \{1, 2, 3\}$; ***(1 mark)***

(c) $A_3 = \{5\}$; ***(1 mark)***

(d) $A_4 = \varnothing$; ***(1 mark)***

(e) $A_5 = \{1, 4\}$. ***(1 mark)***

(ii) You are given the same information about P as in (i) and the additional information that the events $\{1, 2\}$ and $\{2, 3\}$ are independent.

(a) Find the value of $P(\{2\})$. ***(1 mark)***

(b) Are $\{2, 3\}$ and $\{2, 5\}$ independent? ***(2 marks)***

2 A test for a disease will certainly be positive if the patient has the disease, but will be positive with probability 0.01 if the patient does not have the disease. Based on a patient's symptoms, but without using the test, a doctor believes that the probability the patient has the disease is 0.1.

(i) Based on the doctor's beliefs, find the probability that the patient has the disease if the test is positive. You should define your notation carefully and explain your method. ***(3 marks)***

(ii) What is the probability that the patient has the disease if the test is not positive? ***(1 mark)***

3 In each of the following situations, state which out of Binomial, Poisson and Geometric distributions you would expect to be the most suitable model. (You do not have to give the parameters.)

(i) T is the number of thunderstorms to hit Sheffield in June 2017. ***(1 mark)***

(ii) In a game where a player rolls a pair of standard dice on each turn, D is the number of turns until the player rolls a double. ***(1 mark)***

(iii) A commuter travels home on a train at the same time every Monday for 5 weeks, and L is the number of times the train is late. ***(1 mark)***

4 A discrete random variable X has $P(X = 0) = 1/2$ and $P(X = 1) = P(X = 3) = 1/4$.

(i) Find the mean and variance of X. ***(3 marks)***

(ii) Find the moment generating function of X. ***(2 marks)***

(iii) If Y is another random variable with the same distribution as X but which is independent of X,

(a) find the variance of $X + Y$; ***(1 mark)***

(b) show that the moment generating function of $X + Y$, $M_{X+Y}(t)$, is given by

$$M_{X+Y}(t) = \frac{1}{16}(2 + e^t + e^{3t})^2.$$

(2 marks)

(You may use results from the course.)

5 A continuous random variable X has probability density function given by

$$f_X(x) = \begin{cases} \frac{3}{4}(1 - x^2) & -1 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

(i) Find $P\left(X > \frac{1}{2}\right)$. ***(1 mark)***

(ii) Find the cumulative distribution function, $F_X(x)$, for $-1 \le x \le 1$. ***(2 marks)***

(iii) Find $E(X)$ and $\text{Var}(X)$. ***(3 marks)***

6 A colony of birds contains 100 pairs. In a particular year, each pair of birds will raise a number of chicks which is a discrete random variable with mean 3 and variance 1. You may assume that the numbers of chicks raised by different pairs are independent of each other.

(i) Explain why you would expect the distribution of the total number of chicks raised in the colony in the year to be approximately a Normal distribution, and give its mean and variance. ***(4 marks)***

(ii) To get an approximate probability that the number of chicks raised in the year is less than 285, an R command of the form `pnorm(x,mu,sigma)` can be used. In this command, what values would you use for `x`, `mu` and `sigma`? ***(3 marks)***

7 A new genetically-modified variety of wheat has been developed, and has been tested in a small trial. The wheat variety is planted in twelve separate plots of land, each of the same size, and the yield per plot is recorded. If y_i is the yield per plot (in tonnes) for $i = 1, \ldots, 12$, then summary statistics for the twelve yields are

$$\sum_{i=1}^{12} y_i = 36.66, \qquad \sum_{i=1}^{12} y_i^2 = 112.2872.$$

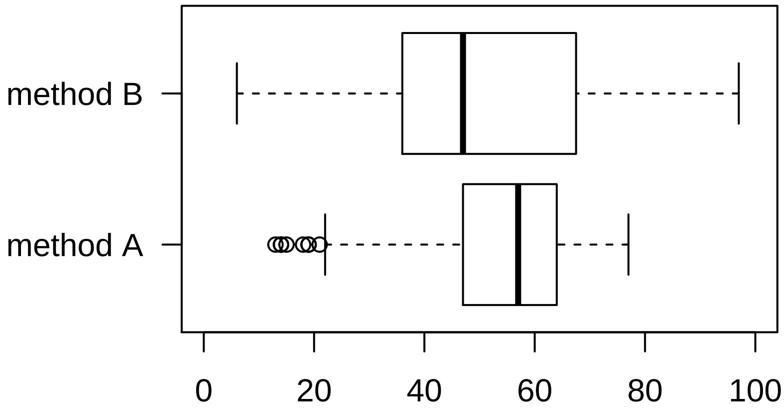
(i) Calculate a 95% confidence interval for the population mean yield per plot. You may use the following R output to help you (not all of which is relevant.) ***(3 marks)***

```
qnorm(c(0.95, 0.975), mean = 0, sd = 1)
## [1] 1.645 1.960
qt(c(0.95, 0.975), df = 11)
## [1] 1.796 2.201
qt(c(0.95, 0.975), df = 12)
## [1] 1.782 2.179
```

(ii) Suppose a hypothesis test is to be conducted at the 5% level of significance of the null hypothesis that the population mean yield is 3 tonnes per plot. *Without doing any further calculations*, would this null hypothesis be rejected or not? Briefly explain your answer (Solutions relying on additional calculations will not get any credit). ***(1 mark)***

(iii) Suppose we want the width of the 95% confidence interval to be no more than 0.1 tonnes. Suggest how many additional observations would be needed to achieve this. You may ignore changes to the resulting degrees of freedom. ***(2 marks)***

8 Two methods of teaching arithmetic to primary school children (aged 7 years) have been compared in an experiment. Children in two schools are taught with method A, and children in another two schools are taught with method B. A total of 200 children were taught using each method. After one month of teaching, the children took a test. Box plots of the children's test scores are plotted below.



(i) Describe the shape of the distribution of test scores for method A.

(1 mark)

(ii) What do the box plots suggest about the effectiveness of each method?

(2 marks)

(iii) Let a_i be the observed score of child i taught with method A, and let b_i be the observed score of child i taught with method B. Sample means and variances for the two sets of marks are as follows.

$$\bar{a} = \frac{1}{200}\sum_{i=1}^{200} a_i = 53.42, \qquad \bar{b} = \frac{1}{200}\sum_{i=1}^{200} b_i = 50.18,$$

$$s_a^2 = \frac{1}{199}\sum_{i=1}^{200}(a_i - \bar{a})^2 = 214.506,$$

$$s_b^2 = \frac{1}{199}\sum_{i=1}^{200}(b_i - \bar{b})^2 = 472.751.$$

Conduct a hypothesis test that the population mean test scores for each method are equal. Specify your null and alternative hypotheses, defining your notation clearly. Give a range for the p-value, using the following R output, and state clearly the interpretation of your result.

```
qnorm(c(0.9, 0.95, 0.975), mean = 0, sd = 1)
## [1] 1.28 1.64 1.96
```

(3 marks)

(iv) Give one criticism of the design of the experiment to compare the two methods. How might you have done things differently? ***(2 marks)***

9 A water company has inspected sections of water pipes for leaks. In each of 50 separate 1km long sections the number of leaks is recorded and tabulated below. A total number of 103 leaks were counted. Test the hypothesis at the 5% level of significance that the number of leaks in a 1km section of pipe follows a Poisson distribution.

number of leaks	0	1	2	3	4	5	6
count	5	13	14	12	5	0	1

You may use the following R output to help you (not all of which is relevant.)

```
dpois(0:3, 103/50)
## [1] 0.127 0.263 0.270 0.186
qchisq(0.95, df = c(3, 4, 5, 6))
## [1]  7.81  9.49 11.07 12.59
```

(5 marks)

10 An opinion poll is to be conducted about voting intentions at the next UK general election. Let θ be the proportion of all voters currently intending to vote Labour. Suppose there will be n participants in the survey, and let $X_i = 1$ if the ith participant says he/she is intending to vote Labour, and 0 otherwise. Assume that $X_1, \ldots, X_n$ are independent and identically distributed with each having the $Bernoulli(\theta)$ distribution.

For parts (i) and (ii) below, **do not** quote standard properties of the Bernoulli distribution: prove any necessary results about Bernoulli random variables from first principles.

(i) Prove that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of θ. ***(2 marks)***

(ii) Derive an expression for $Var(\bar{X})$. ***(2 marks)***

(iii) Is $\bar{X}$ a consistent estimator for θ? Explain your answer briefly, based on your result in part (ii). A formal proof is not required. ***(1 mark)***

End of Question Paper