The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2018–2019**

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 $\{0, 1\}$.

(i) Define a set function M on S by saying that $M(\varnothing) = 0$, $M(\{0\}) = 6$, $M(\{1\}) = 4$ and $M(S) = 3$. Show that M is not a measure. ***(2 marks)***

(ii) Another set function P on S has $P(\{0\}) = \frac{2}{3}$. If P is a probability measure, what are the values of $P(\varnothing)$, $P(\{1\})$ and $P(S)$? ***(2 marks)***

2 At a particular location in its genome, an animal may be of one of three types, denoted "AA", "Aa" or "aa". The probabilities of each type are initially believed to be $\frac{1}{9}$ for "AA", $\frac{4}{9}$ for "aa" and $\frac{4}{9}$ for "Aa".

There is a condition which the animal is believed to have with probability 0.6 if it is of type "AA", 0.5 if it is of type "Aa" and 0.01 if it is of type "aa". Given that testing shows that the animal has the condition, find the probability that the animal is of type "aa". You should define your notation carefully and explain your method. ***(4 marks)***

3 Let X be a discrete random variable taking values in $R_X = \{1, 2, 3\}$ with probability mass function given by $p_X(x) = x/6$ for $x \in R_X$.

(i) Find $E(X)$. ***(1 mark)***

(ii) Find an expression for the moment generating function of X. Hence verify you answer for $E(X)$. ***(3 marks)***

4 In a production process, each item is considered fit to sell with probability 0.8, independently of other items. A batch of three items is produced. Let X be the number of items in the batch which are considered fit to sell.

(i) What distribution would you expect X to have? You should state your answer as a standard distribution with appropriate parameters. ***(2 marks)***

(ii) The items which are fit to sell are put on sale, and, regardless of how many are on sale, each one is sold with probability 0.6, independently of the other items on sale. Let Y be the number of the batch which are sold.

(a) Conditional on $X = x$, what distribution would you expect Y to have? You should state your answer as a standard distribution with appropriate parameters. ***(2 marks)***

(b) Tabulate the joint probability function of X and Y, $p_{X,Y}(x, y)$, for $x, y \in \{0, 1, 2, 3\}$. ***(3 marks)***

(c) What is the probability that exactly one item was put on sale but remains unsold? ***(1 mark)***

5 Let X be a continuous random variable with probability density function given by

$$f_X(x) = \begin{cases} kx^2(1-x) & -1 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

(i) Find the value of k. ***(2 marks)***

(ii) What is $P(X = 1/2)$? ***(1 mark)***

(iii) Find the cumulative distribution function of X, $F_X(x)$, for $-1 \leq x \leq 1$. Hence find $P(X \leq 0)$. ***(3 marks)***

(iv) (a) Find $E(X)$ and $\mathrm{Var}(X)$. ***(4 marks)***

(b) Without further integration, find the mean and variance of the random variable $1 - X$. ***(2 marks)***

6 Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with mean μ and variance σ^2. Define $S(n) = \sum_{k=1}^{n} X_k$ and $\bar{X}(n) = \frac{S(n)}{n}$. For each of the following statements, state whether or not you can deduce the statement from results in the course.

(i) X_n has an approximately Normal distribution if n is sufficiently large; ***(1 mark)***

(ii) $\bar{X}(n)$ has an approximately Normal distribution if n is sufficiently large; ***(1 mark)***

(iii) $\bar{X}(n)$ has an exactly Normal distribution if n is sufficiently large; ***(1 mark)***

(iv) $P(\mu - 0.01 \leq \bar{X}(n) \leq \mu + 0.01) \to 1$ as $n \to \infty$. ***(1 mark)***

7 An analyst is going to study the demand for taxis at a particular taxi company. She plans to record her data in July this year: she will record the number of taxis booked in each day in July, obtaining 31 observations in total.

(i) Assuming the 31 observations are independent and identically distributed, and defining your notation carefully, suggest a suitable model for the data that will be obtained. ***(2 marks)***

(ii) The analyst suspects the demand for taxis will be higher at weekends. Give one criticism of the assumption made in part (i). ***(1 mark)***

(iii) Suggest one modification of your model, defining your notation carefully, that addresses your criticism in part (ii). ***(2 marks)***

8 An opinion poll has been conducted to see which of two candidates voters intend to vote for in a forthcoming election. 50 voters are sampled at random, and 27 voters state they intend to vote for candidate A.

(i) Calculate a 95% confidence interval for the proportion of all voters intending to vote for candidate A, using some of the following R output.

```
qnorm(c(0.95, 0.975))
## [1] 1.645 1.960
```

(2 marks)

(ii) With reference to your confidence interval, comment on whether the sample size is suitable for the opinion poll. Note that the voters in the poll have been selected at random, so there is no concern that the sample could be biased. ***(1 mark)***

9 Students can either study on an MSc course full-time over one year, or part-time over two years. The final outcome for a student on the MSc course is one of "distinction", "merit", "pass" or "fail". Numbers achieving each outcome for 43 full-time students and 55 part-time students are as follows.

Students	distinction	merit	pass	fail
Full-time	8	17	15	3
Part-time	10	20	15	10

A hypothesis test is to be conducted, with the null hypothesis that the distribution of outcomes for full-time students is the same as the distribution of outcomes for part-time students. In an R session, the data are stored in a matrix called `outcomes`. Some edited output from an R session is given below:

```
> outcomes
     [,1] [,2] [,3] [,4]
[1,]    8   17   15    3
[2,]   10   20   15   10

> chisq.test(outcomes)

Pearson's Chi-squared test

data:  outcomes
X-squared = 2.8074, df = ?, p-value = 0.4223
```

Note that the observed value of the test statistic has been reported as 2.8074.

(i) To compute the observed test statistic, what value should be used for the expected number of distinctions for part-time students? ***(2 marks)***

(ii) Assuming all the expected counts are greater than 5, what distribution would you compare the test statistic with, to test the null hypothesis? ***(1 mark)***

(iii) Draw a sketch to indicate the density function of the test statistic under the null hypothesis, the observed value of the test statistic, and the p-value as an area under a curve. ***(2 marks)***

(iv) The course director asks you if full-time students tend to get better results than part-time students. Based on the hypothesis test, state what would you tell her, and briefly explain how you have used the R output to draw your conclusion. ***(2 marks)***

(v) Suppose the hypothesis test were to be conducted under the Neyman-Pearson framework, with size 0.05. Given the observed data, and **without** doing any calculations, state what the conclusion of the test would be. ***(1 mark)***

 Question 9 continued on next page

9 (continued)

(vi) For your conclusion in part (v), explain

(a) whether you could have made a Type I error; ***(2 marks)***

(b) whether you could have made a Type II error. ***(2 marks)***

Note: in your answer, you should include definitions of Type I and Type II errors. If you mix up the definitions, full marks will still be awarded, as long as your answers are consistent with the definitions you have provided.

10 Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables, each with the $N(0, \sigma^2)$ distribution.

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

(ii) Given that $Var(X_i^2) = 2\sigma^4$, is $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ a consistent estimator of σ^2? Justify your answer.

(2 marks)

End of Question Paper

MAS113 Introduction to Probability and Statistics
Formula Sheet and R Commands

1. **Sample variance, covariance and correlation**.

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2\right).$$

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \qquad r_{XY} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}}.$$

2. **Density functions**. If $X \sim N(\mu, \sigma^2)$ then

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), \quad -\infty < x < \infty$$

If $Y \sim \chi^2_\nu$ then

$$f_Y(y) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{\nu/2-1} \exp\left(-\frac{y}{2}\right), & y \geq 0, \\ 0 & y < 0. \end{cases}$$

If $Y \sim t_\nu$ then

$$f_\nu(y) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} (1 + y^2/\nu)^{-(\nu+1)/2}, \quad -\infty < y < \infty$$

3. **Distribution theory**. If $Z \sim N(0,1)$ and $Y \sim \chi^2_\nu$, then

$$\frac{Z}{\sqrt{Y/\nu}} \sim t_\nu,$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}, \quad \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

4. **Confidence intervals**

$$\bar{x} \pm t_{n-1,0.025}\sqrt{\frac{s^2}{n}}, \qquad \left[\frac{(n-1)s^2}{\chi^2_{n-1;0.025}}, \frac{(n-1)s^2}{\chi^2_{n-1;0.975}}\right]$$

$$p \pm z_{0.025}\sqrt{\frac{p(1-p)}{n}}, \qquad \bar{x} - \bar{y} \pm t_{\nu,0.025}\sqrt{s_X^2/n + s_Y^2/m}$$

5. **Hypothesis tests**

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}} \sim t_\nu, \qquad \nu = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{(s_X^2/n)^2}{n-1} + \frac{(s_Y^2/m)^2}{m-1}}$$

$$Z = \frac{\frac{X}{n} - \frac{Y}{m} - (\theta_X - \theta_Y)}{\sqrt{P^*(1-P^*)\left(\frac{1}{n} + \frac{1}{m}\right)}} \overset{\text{approx}}{\sim} N(0,1), \qquad P^* = \frac{X+Y}{n+m}.$$

$$R_{XY}\sqrt{\frac{n-2}{1-(R_{XY})^2}} \sim t_{n-2}$$

$$X^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(r-1)(c-1)}$$

6. `pbinom(x, n, p)` will calculate the value of

$$P(X \leq \texttt{x}),$$

for $X \sim Binomial(\texttt{n},\ \texttt{p})$.

7. `ppois(x, lambda)` will calculate the value of

$$P(X \leq \texttt{x}),$$

for $X \sim Poisson(\texttt{lambda})$.

8. `qnorm(p, mean = m, sd = s)` will give the value of x such that

$$P(X \leq x) = \texttt{p},$$

for $X \sim N(\texttt{m}, \texttt{s}^2)$. The default values of `m` and `s` are 0 and 1 respectively.

9. `pnorm(x, mean = m, sd = s)` will calculate the value of

$$P(X \leq \texttt{x}),$$

for $X \sim N(\texttt{m}, \texttt{s}^2)$. The default values of `m` and `s` are 0 and 1 respectively.

10. `qt(p, df = n)` will give the value of x such that

$$P(T \leq x) = \texttt{p},$$

for $T \sim t_{\texttt{n}}$.

11. `pt(x, df = n)` will calculate the value of

$$P(T \leq \texttt{x}),$$

for $T \sim t_{\texttt{n}}$.

12. `qchisq(p, df = n)` will give the value of x such that

$$P(X \leq x) = \texttt{p},$$

for $X \sim \chi^2_{\texttt{n}}$.

13. `pchisq(x, df = n)` will calculate the value of

$$P(X \leq \texttt{x}),$$

for $X \sim \chi^2_{\texttt{n}}$.