The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2017–2018**

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

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

Please leave this exam paper on your desk
Do not remove it from the hall

Registration number from U-Card (9 digits)
to be completed by student

Blank

1 Let S be the set $\{-2, -1, 0, 1, 2\}$.

(i) Define a set function M by saying that for a subset A of S, $M(A)$ is 1 if both of -2 and 2 are elements of A, and 0 otherwise. Show that M is not a measure. (Hint: consider the sets $\{-2\}$ and $\{2\}$.) ***(2 marks)***

(ii) Consider the probability measure P defined by saying that for a subset A of S, $P(A)$ is the number of elements in A divided by 5. Define a random variable X with range $\{0, 1, 4\}$ by letting $X(s) = s^2$ for $s \in S$. Using the probabilities given by P, give the probabilities that X takes each of its possible values. ***(2 marks)***

2 A pair of birds is known to be either a hybrid pair, with probability 0.2, or a same-species pair, with probability 0.8. If they are a hybrid pair, then the probability of them raising a chick is believed to be 0.3, while if they are a same-species pair it is believed to be 0.7. If the pair are observed not to have raised a chick, find the probability that in fact they are a hybrid pair. You should define your notation carefully and explain your method. ***(3 marks)***

3 Let X be the number of cases of a rare disease in the UK which occur in June 2018. Expert opinion about the disease concludes that the expected number of cases in the month is 2. Assuming that the distribution of X is modelled as a Poisson distribution, $Poisson(\lambda)$:

(i) What value of λ would you use? ***(1 mark)***

(ii) Using your assumed value of λ, what would the variance of X be? ***(1 mark)***

(iii) Using your assumed value of λ, what would the probability be that there are no cases in June 2018? ***(1 mark)***

(iv) To find the probability that there are at least 6 cases in June 2018, an R command of the form `1-ppois(x,lambda)` could be used. What values would you use for `x` and `lambda`? ***(2 marks)***

4 Two random variables X and Y, each with range $\{1, 2, 3\}$, have joint probability mass function $p_{X,Y}$ as given in the following table.

$p_{X,Y}$	$x = 1$	$x = 2$	$x = 3$
$y = 1$	0.2	0.1	0
$y = 2$	0.1	0.2	0.1
$y = 3$	0	0.1	0.2

(i) Find the marginal probability mass function of X. ***(2 marks)***

(ii) Find $E(X)$ and $\mathrm{Var}(X)$. ***(3 marks)***

(iii) Find $P(X + Y = 3)$. ***(1 mark)***

(iv) Find $P(Y = 2|X = 2)$. ***(1 mark)***

(v) Are X and Y independent? Give a reason for your answer. ***(2 marks)***

5 (i) Explain why the following functions cannot be probability density functions for a continuous random variable.

(a)
$$f_1(x) = \begin{cases} x & 0 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

(b)
$$f_2(x) = \begin{cases} x - \dfrac{1}{2} & 0 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

(2 marks)

(ii) Let X be a continuous random variable with probability density function given by
$$f_X(x) = \begin{cases} \dfrac{3}{2}x^2 & -1 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the cumulative distribution function of X, $F_X(x)$, for $-1 \leq x \leq 1$. Hence find $P\left(X \leq -\dfrac{1}{2}\right)$. ***(3 marks)***

(b) Find $E(X)$ and $\mathrm{Var}(X)$. ***(3 marks)***

6 The random variables $Y_1, Y_2, Y_3, \ldots$ are known to all have the same distribution, to be independent of each other, and to have expectation 0 and variance 9.

(i) What are the variances of

(a) $Y_1 + Y_2$;

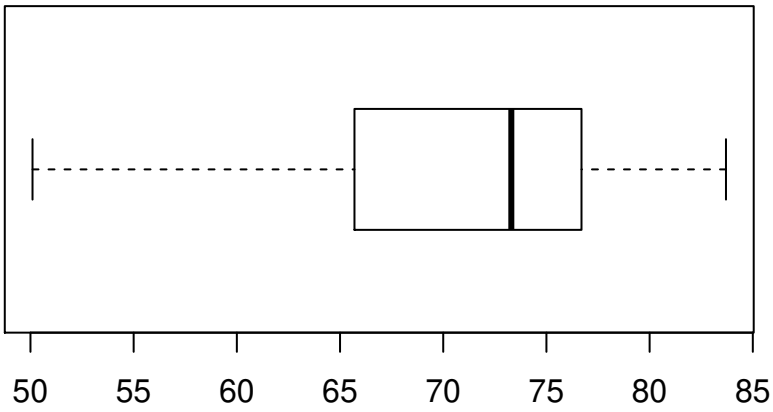
(b) $2Y_1$? ***(2 marks)***

(ii) Use Chebyshev's inequality to give an upper bound for $P(|Y_1| \geq 6)$. ***(2 marks)***

(iii) Let $S(n) = \sum_{i=1}^{n} Y_i$. What can you say about the distribution of $\frac{S(n)}{3\sqrt{n}}$ if n is large? You should state which result from the course you are using. ***(3 marks)***

7 In an R session, the vector `life2015` contains life expectancies in years for 181 countries, from a data set compiled by the World Health Organisation.

(i) A box plot of the 181 life expectancies is shown below.



Draw a sketch of the histogram of the same data. You only need to get the basic shape of the histogram correct; the precise detail is not important.

(2 marks)

(ii) Using the following R output, calculate the inter-quartile range.

```
summary(life2015)
```

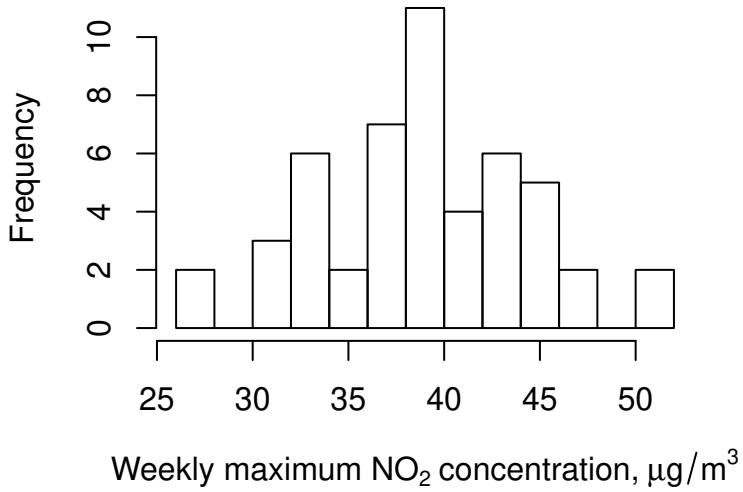
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50.10   65.70   73.30   71.21   76.70   83.70
```

(1 mark)

(iii) A region is specified for each of the 181 countries: one out of Africa, Americas Eastern Mediterranean, Europe, South-East Asia, Western Pacific. Suppose we want to compare how life expectancies vary between the regions, using a single plot. What would be the most suitable plot to do this? Draw a sketch to indicate the plot you would use. The sketch only needs to clearly show the type of plot; the values indicated in your sketch are not important.

(2 marks)

8 A monitoring station records the maximum level of the air pollutant NO_2 each week, measured in $\mu g/m^3$. In an R session, 50 measurements (maximums in each of 50 weeks) are stored in the vector `NO2`. A histogram of the measurements is plotted below.



(i) By carefully defining suitable notation for the 50 observations, state what distribution you would use to model these data. Briefly explain how to interpret the parameters in your model. ***(2 marks)***

(ii) Using the following R output, estimate the parameters of your chosen distribution in part (i)

```
sum(NO2)
## [1] 1953.89
sum(NO2 ^ 2)
## [1] 77803.75
```

(2 marks)

(iii) Based on your chosen distribution and parameter estimates, give an approximate 95% probability interval for a future weekly maximum NO_2 measurement, i.e., an interval in which you think the future measurement would lie in with probability 0.95. ***(2 marks)***

9 Garaulet et al. (2013) reported an experiment to investigate whether the timing of the main meal (lunch) had an effect on weight loss for adults on a 20 week weight-loss programme. There were 402 participants in their experiment, with 202 classified as "early eaters" and 200 classified as "late eaters". Data from the experiment are as follows, where the quantity measured is weight loss as a percentage of the participant's initial weight.

Group	sample size	sample mean	sample standard deviation
Early eaters	202	11.3%	5.8%
Late eaters	200	9.0%	7.1%

For this question, as the sample sizes are fairly large, you can assume that the t-distribution can be approximated by the normal distribution. Some R output related to the normal distribution is as follows.

```
qnorm(c(0.9, 0.95, 0.975, 0.99, 0.995, 0.9995))
## [1] 1.282 1.645 1.960 2.326 2.576 3.291
```

(i) (a) Test the hypothesis that the population mean weight loss is the same for early eaters and late eaters. For a test of size 0.05, state whether the null hypothesis would be rejected or not. ***(2 marks)***

(b) Draw a sketch of the normal distribution and rejection region, marking on it the observed value of the test statistic. ***(2 marks)***

(c) Give a suitable bound for the p-value, and state how this should be interpreted. State the command you would use in R to calculate the p-value exactly (assuming a normal distribution for your test statistic). ***(2 marks)***

(ii) Calculate a 95% confidence interval for the difference between the two population mean weight losses. Be careful to use percentages in your answer. ***(2 marks)***

(iii) In two or three sentences, comment on whether the experiment has demonstrated that weight loss is affected by the timing of the main meal, and whether any differences are likely to be large or not. You should base your answer on your results to parts (i) and (ii), but write in plain English only: do not use any technical statistical terms. ***(2 marks)***

10 Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables, each with the $Exponential(\lambda)$ distribution. Suppose we wish to estimate the parameter ϕ, defined as $\phi = 1/\lambda$. Note that, in terms of this parameter ϕ, we have

$$\mathbb{E}(X_i) = \phi, \quad \mathrm{Var}(X_i) = \phi^2.$$

(i) Is $\bar{X}$ an unbiased estimator of ϕ? Justify your answer with a short proof. ***(1 mark)***

(ii) Derive the standard error of $\bar{X}$, in terms of the sample size n and ϕ. ***(1 mark)***

(iii) From your results in parts (i) and (ii), state, with brief justification, whether $\bar{X}$ is a consistent estimator of ϕ or not. ***(1 mark)***

End of Question Paper

MAS113 Introduction to Probability and Statistics
Formula Sheet and R Commands

The purpose of this is to provide a memory aid for the exam paper. Not all formulae and R commands will be needed in the exam questions. Terms within the formulae will not be defined, but the notation will be the same as that used in your lecture notes.

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

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

3. If $X \sim N(\mu, \sigma^2)$ then
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), \quad -\infty < x < \infty$$

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

5. If $Y \sim t_\nu$ then
$$f_\nu(y) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} (1 + y^2/\nu)^{-(\nu+1)/2}, \quad -\infty < y < \infty$$

6. If $Z \sim N(0,1)$ and $Y \sim \chi^2_\nu$, then
$$\frac{Z}{\sqrt{Y/\nu}} \sim t_\nu,$$

7.
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

8.
$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

9.
$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n_X + S_Y^2/n_Y}} \sim t_\nu, \qquad \nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{(s_X^2/n_X)^2}{n_X - 1} + \frac{(s_Y^2/n_Y)^2}{n_Y - 1}} \simeq \min\{n_X - 1, n_Y - 1\}$$

10.
$$r_{XY}\sqrt{\frac{n-2}{1-(r_{XY})^2}} \sim t_{n-2}$$

11.
$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}.$$

12. `pbinom(x, n, p)` will calculate the value of
$$P(X \leq \texttt{x}),$$
for $X \sim Binomial(\texttt{n},\ \texttt{p})$.

13. `ppois(x, lambda)` will calculate the value of
$$P(X \leq \texttt{x}),$$
for $X \sim Poisson(\texttt{lambda})$.

14. `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.

15. `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.

16. `qt(p, df = n)` will give the value of x such that
$$P(T \leq x) = \texttt{p},$$
for $T \sim t_{\texttt{n}}$.

17. `pt(x, df = n)` will calculate the value of
$$P(T \leq \texttt{x}),$$
for $T \sim t_{\texttt{n}}$.

18. `qchisq(p, df = n)` will give the value of x such that
$$P(X \leq x) = \texttt{p},$$
for $X \sim \chi^2_{\texttt{n}}$.

19. `pchisq(x, df = n)` will calculate the value of
$$P(X \leq \texttt{x}),$$
for $X \sim \chi^2_{\texttt{n}}$.