Data provided: formula sheet

MAS223

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2015–2016**

Statistical Modelling and Inference **2 hours 30 minutes**

*Candidates should attempt **ALL** questions.*
The maximum marks for the various parts of the questions are indicated.
The paper will be marked out of 90.

1 Let X and Y be random variables with joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} 2x+4y & \text{if } 0<x<1 \text{ and } 0<y<\frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the marginal probability density function $f_Y(y)$ of Y. ***(3 marks)***

(b) Find the distribution function $F_Y(y)$ of Y. ***(3 marks)***

(c) Let $y \in (0, \frac{1}{2})$. Find the conditional probability density function of X, given that $Y = y$. ***(2 marks)***

(d) Show that

$$\mathbb{E}[X|Y=y] = \frac{2+6y}{3(1+4y)}.$$

and write down a formula for $\mathbb{E}[X|Y]$. ***(4 marks)***

2 Let X be an exponential random variable with rate parameter $\lambda > 0$.

(a) Find the probability density function of $Y = g(X)$, where $g(x) = e^{-\lambda X}$. ***(5 marks)***

(b) State which standard distribution is the distribution of Y. ***(1 mark)***

3 Let $\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$ be a random vector with a bivariate normal distribution, with mean vector $\boldsymbol{\mu} = \begin{pmatrix} -1 \\ 3 \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix}$.

(a) Write down the marginal distribution of Y. ***(2 marks)***

(b) Find the correlation coefficient of X and Y. ***(2 marks)***

(c) Let $U = X + Y$ and $V = X - 2Y$. Find the mean vector and covariance matrix of the random vector $\begin{pmatrix} U \\ V \end{pmatrix}$. ***(5 marks)***

4 Let X and Y be random variables with joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{2}\sin(x+y) & \text{if } 0 < x < \frac{\pi}{2} \text{ and } 0 < y < \frac{\pi}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Let $U = X + Y$ and $V = X/2$.

(a) Find the joint probability density function of U and V, stating clearly the region on which it is non-zero. ***(8 marks)***

(b) Are U and V independent? Give a reason for your answer. ***(2 marks)***

5 The random variable X is known to have Poisson distribution, with unknown parameter $\theta > 0$. Three independent samples of X take the values $\mathbf{x} = (1, 4, 2)$.

(a) Show that the likelihood function $L(\theta; \mathbf{x})$ of θ given the data $\mathbf{x}$, satisfies

$$L(\theta; \mathbf{x}) = \frac{e^{-3\theta}\theta^7}{48}.$$

(3 marks)

(b) Find the corresponding log-likelihood function $l(\theta; \mathbf{x})$. ***(3 marks)***

(c) Show that the maximum likelihood estimator of θ, given the data $\mathbf{x}$, is

$$\hat{\theta} = \frac{7}{3}$$

(5 marks)

(d) Suggest why we might think of the value obtained in (c) for $\hat{\theta}$ as a natural estimator of θ. ***(2 marks)***

6 Consider a linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ with the number of parameters in the $\boldsymbol{\beta}$ vector greater than 3. Write down the estimators for $\boldsymbol{\beta}$ and σ^2 and their sampling distributions, and hence derive the $100(1-\alpha)\%$ confidence interval for β_3. You must show clearly how you obtain the confidence interval and define all the quantities used in the confidence interval. ***(6 marks)***

7 A one way ANOVA model was used to test the effectiveness of 3 drugs against a certain disease. Nine patients were divided into three equal groups, and each group was administered a different drug. The number of days to recovery of the three groups are noted below.

Drug 1: $10, 11, 9$, **Drug 2:** $8, 9, 13$, **Drug 3:** $12, 13, 14$.

Let μ_1, μ_2 and μ_3 denote the mean recovery times for the groups.

(a) Write down the model, clearly stating any assumptions that you make. ***(2 marks)***

(b) Conduct the F-test for $H_0 : \mu_2 = \mu_3$ versus $H_a : \mu_2 \neq \mu_3$ and report the P-value as $P(F_{?,?} > ?)$. [*You will need to fill in the ? marks*] ***(8 marks)***

8 This question concerns data collected on a sample of countries in the year 1993. For each country the following three variables were computed:

$life$: Average life expectancy in years

tv : Average number of people per television set

$doctor$: Average number of people per doctor

Here is the numerical summary information

life		tv		doctor	
Min. :	51.5	Min. :	1.30	Min. :	226
1st Qu.:	64.1	1st Qu.:	3.35	1st Qu.:	457
Median :	70.0	Median :	6.30	Median :	824
Mean :	67.8	Mean :	51.98	Mean :	2934
3rd Qu.:	74.1	3rd Qu.:	23.00	3rd Qu.:	2862
Max. :	79.0	Max. :	592.00	Max. :	36660

A regression model $life = \beta_0 + \beta_1 \cdot tv + \beta_2 \cdot doctor + \epsilon$ was fit and the following output was obtained.

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	70.251957	1.087705	64.59	$< 2e-16$
tv	−0.023495	0.009647	−2.44	0.02
doctor	−0.000432	0.000202	−2.14	0.04

```
Residual standard error:  6 on 35 degrees of freedom
Multiple R-Squared:  0.44, Adjusted R-squared:  0.408
F-statistic:  13.8 on 2 and 35 DF, p-value:  3.92e-05
```

(a) What variables, if any, have skewed distributions? Justify. ***(2 marks)***

(b) How many countries were there in the data set? ***(1 mark)***

(c) What is the estimate for σ^2 ? ***(1 mark)***

(d) The regression parameters were estimated using least squares. What extra assumption is required for the validity of the t and F tests that is not needed when computing the least squares estimates? ***(1 mark)***

(e) What is the P-value of the test corresponding to the hypothesis $H_0 : \beta_1 = \beta_2 = 0$? ***(1 mark)***

(f) What is the P-value of the test corresponding to the hypothesis $H_0 : \beta_2 = 0$? ***(1 mark)***

(g) What would be the predicted life expectancy in a country with an unlimited supply of televisions and an average of 100 people per doctor ? ***(2 marks)***

8 (continued)

(h) Ethiopia had the largest residual which was an outlier. Does this also imply that Ethiopia is an influential point? ***(1 mark)***

(i) Log transformations of both the predictors were taken and the model refitted resulting in the following output:

```
Coefficients:
```

	Estimate	Std.Error	tvalue	Pr(>|t|)
(Intercept)	90.622	4.356	20.81	< 2e − 16
log(tv)	−2.916	0.591	−4.94	1.9e − 05
log(doctor)	−2.259	0.747	−3.02	0.0047

```
Residual standard error:  3.7 on 35 degrees of freedom
Multiple R-Squared:  0.787, Adjusted R-squared:  0.775
F-statistic:  64.6 on 2 and 35 DF, p-value:  1.79e-12
```

Is this model better than the previous model? Explain. ***(2 marks)***

(j) Suppose that we used number of TVs per person rather than number of persons per TV in constructing the predictor in the model used in part (i). What can we say about the estimated value of the regression coefficient for this transformed predictor? ***(2 marks)***

9 A two way analysis of variance was conducted to investigate the abilities of various treatments to induce blood clots in order to stem severe bleeding. Three different treatments have been tested, and blood samples have been taken from three patients. Each treatment has been applied to three samples from each patient. The dependent variable of interest is the time taken for a blood clot to form. The two independent variables or factors are treatment and patient. Below is part of the analysis done in `R` of the dataset `blood`, which contains the data of the experiment.

```
anova(lm(clottime ~ treat*patient,blood))

Analysis of Variance Table

Response:  clottime
```

	Df	SumSq	MeanSq	Fvalue	Pr(> F)
treat	2	a	5.8900	80.3182	1.071e − 09 ∗ ∗∗
patient	2	9.26	4.6300	b	7.325e − 09 ∗ ∗∗
treat : patient	4	0.74	d	c	0.07709 .
Residuals	18	1.32	0.0733		

```
− − −

Signif.codes : 0 ∗ ∗ ∗ 0.001 ∗ ∗ 0.01 ∗ 0.05 . 0.1  1
```

(a) Write down the full model corresponding to the above analysis defining all the quantities involved. *(3 marks)*

(b) What tests do the the first three rows of the table represent? *(3 marks)*

(c) Compute a, b, c, d. *(4 marks)*

End of Question Paper