Data provided: Supplementary Formula Sheet

MAS223

The University Of Sheffield.

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

Statistical Inference and Modelling **2 hours and 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 $T = \{(x, y)\ ;\ 0 < |y| < x < \infty\}$. Let (X, Y) be a bivariate random vector with probability density function

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{4} y^2 e^{-x} & \text{for } (x, y) \in T, \\ 0 & \text{othwerwise.} \end{cases}$$

(a) Sketch the region T. ***(2 marks)***

(b) Find the marginal probability density function $f_X(x)$ of X. ***(4 marks)***

(c) Let $x > 0$. Find the conditional probability density function of Y given that $X = x$. ***(3 marks)***

(d) Evaluate $\mathbb{E}[Y|X]$. ***(3 marks)***

(e) Consider the following claim:

> *The value of $\mathbb{E}[Y|X]$ does not depend on X, because X and Y are independent.*

Do you agree? Justify your answer briefly. ***(2 marks)***

2 Let $\mathbf{X} = (X, Y)^T$ be a random vector with a bivariate normal distribution, with mean vector and covariance matrix

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}.$$

(a) Let $U = X + Y$ and $V = 2X - Y - 1$. Find the mean vector and covariance matrix of the random vector $\mathbf{U} = (U, V)^T$. ***(5 marks)***

(b) Write down the distribution of U. ***(2 marks)***

3 (a) Let X and Y be random variables with joint probability density function

$$f_{X,Y} = \frac{1}{\pi}(x^2 + y^2)e^{-(x^2+y^2)}.$$

Let $U = X + Y$ and $V = X - Y$.

Find the joint probability density function $f_{U,V}(u, v)$ of U and V, stating clearly the region on which it is non-zero. ***(8 marks)***

(b) Let Z be uniformly distributed on the interval $(-1, 1)$. Find the distribution function $F_W(w)$ of $W = Z^2$. ***(5 marks)***

4 Let X and Y be independent random variables. Suppose that $\mathbb{E}[X] = \mathbb{E}[Y]$ and also that $\mathrm{Var}(X) = \mathrm{Var}(Y)$. Let $U = X - Y$ and $V = XY$.

(a) Show that $\mathrm{Cov}(U, V) = 0$. ***(3 marks)***

(b) Suppose, additionally, that X and Y are Bernoulli random variables, with identical distribution given by

$$\mathbb{P}[X = 0] = \mathbb{P}[X = 1] = \mathbb{P}[Y = 0] = \mathbb{P}[Y = 1] = \frac{1}{2}.$$

Are U and V independent? Justify your answer.

(3 marks)

5 Let $n \in \mathbb{N}$ and let $\mathbf{x} = (x_1, \ldots, x_n)$ be a vector of independent, identically distributed samples from the $Be(\theta, 1)$ distribution. Here, $\theta \in (0, \infty)$ is an unknown parameter.

(a) Show that $B(\theta, 1) = \frac{1}{\theta}$, where $B(\cdot, \cdot)$ denotes the Beta function. ***(1 mark)***

(b) Find the likelihood function $L(\theta; \mathbf{x})$ of θ, and the corresponding log-likelihood function $\ell(\theta; \mathbf{x})$. ***(4 marks)***

(c) Find the maximum likelihood estimator $\hat{\theta}$ for θ, given the data $\mathbf{x}$.

(5 marks)

6 Suppose we have some data as follows: $(x_1, y_1) = (1, 1.5)$, $(x_2, y_2) = (2, 1.5)$, $(x_3, y_3) = (3, 5)$, $(x_4, y_4) = (4, 7.5)$. We model this as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

where $i \in \{1, 2, 3, 4\}$ and $\epsilon_i \sim N(0, \sigma^2)$.

(a) Write this model in matrix notation. ***(2 marks)***

(b) Show that the least-squares estimators for β_0, β_1, and β_2 are $\hat{\beta}_0 = 1.625$, $\hat{\beta}_1 = -0.975$, and $\hat{\beta}_2 = 0.625$, respectively. You may use the following results:

$$\begin{pmatrix} 4 & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{pmatrix}^{-1} = \frac{1}{20} \begin{pmatrix} 155 & -135 & 25 \\ -135 & 129 & -25 \\ 25 & -25 & 5 \end{pmatrix},$$

where each sum is taken over the set $i \in \{1, 2, 3, 4\}$, and

$$\frac{1}{20} \begin{pmatrix} 155 & -135 & 25 \\ -135 & 129 & -25 \\ 25 & -25 & 5 \end{pmatrix} \begin{pmatrix} 15.5 \\ 49.5 \\ 172.5 \end{pmatrix} = \begin{pmatrix} 1.625 \\ -0.975 \\ 0.625 \end{pmatrix}.$$

(5 marks)

(c) Find an estimator of σ^2. ***(5 marks)***

(d) We wish to test $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$. Find the p-value for this hypothesis test in the form $P(F_{?,?} > ?)$. You may quote the value $RSS_r = 2.575$ for the residual sum of squares for the reduced model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. ***(3 marks)***

(e) Consider the reduced model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. When testing the hypothesis $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ for this (reduced) model, you calculate the p-value to be $p = 0.0514$. What can you conclude from this regarding the relative fits of $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and $y_i = \beta_0 + \epsilon_i$ to the data (x_i, y_i)? ***(2 marks)***

7 In 2016, the United Kingdom (UK) voted in a referendum on whether to leave the European Union (EU). In 2017, the UK had a general election where votes for members of parliament in each constituency were cast. It has been postulated that those British voters who voted Leave in 2016 were more likely to have voted for the Conservative Party in 2017.

To test this, a simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, was fitted in `R`, where i indexes the 638 constituencies in Britain, y_i is the proportion of people in constituency i who voted Conservative in 2017 (`Con2017`), and x_i is the proportion of people in constituency i who voted Leave in 2016 (`Leave2016`). The data were stored in the `voter` dataframe. Here is some `R` output:

```
> lmvoter<-lm(Con2017~Leave2016,data=voter)
> summary(lmvoter)

Call:
lm(formula = Con2017~Leave2016, data = voter)

Residuals:
 Min       1Q        Median   3Q       Max
 -0.42895  -0.11230  0.02199  0.11750  0.27088

Coefficients:
             Estimate  Std.  Error  t value  Pr(>|t|)
 (Intercept) 0.10773   0.02655      4.058    5.56e-05
 Leave2016   0.60237   0.04986      12.082   <2e-16

Residual standard error:  0.1434 on 636 degrees of freedom
Multiple R-squared:  0.1867, Adjusted R-squared:  0.1854
F-statistic:  146 on 1 and 636 DF, p-value:  < 2.2e-16

> deviance(lmvoter)
[1] 13.07105
> deviance(lm(Con2017~1,data=voter))
[1] 16.07136
```

7(continued)

(a) Write a paragraph reporting on the following aspects of the results:

- The hypothesis test used to test for a correlation between those who voted Leave in 2016 and those who voted Conservative in 2017
- What you conclude from this test and why
- The statistic that demonstrates how well a simple linear regression model fits the data
- What you can conclude from the value of this statistic
- To what extent the proportion of Leave votes in 2016 is a good predictor of the proportion of Conservative votes in 2017.

Make sure you write your answer in properly-constructed, grammatically-correct sentences. ***(7 marks)***

(b) Calculate the 95% confidence interval for β_1. You are given that $t_{(0.975,638)} = 1.963689$, $t_{(0.95,638)} = 1.647245$, $t_{(0.975,636)} = 1.963701$, $t_{(0.95,636)} = 1.647253$. ***(3 marks)***

(c) Suppose you know that 40% of people voted Leave in a particular constituency. What is the expected proportion of Conservative votes in that constituency in 2017? ***(1 mark)***

8 You are given some data on mortality in various employment sectors, together with the proportion of smokers in each sector (estimated by sampling). You want to test for the effect on mortality of (i) smoking, and (ii) whether or not the work is outdoors. The professions are split into two groups: outdoor-working, denoted by $i = 1$ and indoor-working, denoted by $i = 2$. Let $x_{i,j}$ be the smoking index of profession j within group i. Let $y_{i,j}$ be the mortality index of profession j within group i.

(a) To test the effect of smoking and outdoor-working on mortality, five different models are constructed. Each model can be described both in terms of the assumptions it makes and as a mathematical expression. Below, models M_0 and M_2 are described in terms of the assumptions made. Models M_1, M_3, and M_4 are given as equations, where you should assume that $\epsilon_{i,j} \sim N(0, \sigma^2)$ and that the $\epsilon_{i,j}$ are independent.

- Model M_0: Assumes there is no relationship between mortality and either smoking or working outdoors
- Model M_1: $y_{i,j} = \beta_0 + \beta_1 x_{i,j} + \epsilon_{i,j}$
- Model M_2: Assumes that there is a relationship between outdoor working and mortality, but not between smoking and mortality.
- Model M_3: $y_{i,j} = \beta_0 + \tau_i + \beta_1 x_{i,j} + \epsilon_{i,j}$
- Model M_4: $y_{i,j} = \beta_0 + \tau_i + \beta_{1,i} x_{i,j} + \epsilon_{i,j}$

For models M_0 and M_2, give the corresponding mathematical expressions. For each of models M_1, M_3, and M_4, describe the assumptions being made.

(8 marks)

(b) You are given the following output, testing model M_3 against model M_4

```
> lm4<-lm(mortality~outdoors*smoking,smokingdf)
> lm3<-lm(mortality~outdoors+smoking,smokingdf)
> anova(lm3,lm4)
Analysis of Variance Table

Model 1:  mortality~outdoors + smoking
Model 2:  mortality~outdoors * smoking
   Res.Df  RSS   Df  Sum of Sq  F      Pr(>F)
 1 22      7968
 2 21      6865  1   1103       3.374  0.08043.
```

What are the null and alternative hypotheses for this test? What can we conclude about the effects of smoking and outdoor work on mortality?

(4 marks)

End of Question Paper

SOME DISCRETE DISTRIBUTIONS

Name	**Parameters**	**Genesis / Usage**	**Notation**	$p(x) = \mathbb{P}[X = x]$ **(and non-zero range)**	$\mathbb{E}[X]$	$\mathrm{Var}(X)$	**Comments**
Uniform (discrete)	$k \in \mathbb{N}$	Set of k equally likely outcomes	$Unif(1, \ldots, k)$ (not standard)	$p(x) = 1/k$ $x = 1, ..., k$	$\frac{k+1}{2}$	$\frac{k^2-1}{12}$	Fair dice roll ($k = 6$)
Bernoulli trial	$\theta \in [0, 1]$	Experiment with two outcomes (typically, success = 1, fail = 0)	$Bernoulli(\theta)$	$p(x) = \theta^x(1-\theta)^{1-x}$ $x = 0, 1$	θ	$\theta(1-\theta)$	Coin toss
Binomial	$n \in \mathbb{N}$, $\theta \in [0, 1]$	Number of successes in n i.i.d. Bernoulli trials	$Bi(n, \theta)$ $B(n, p)$	$p(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ $x = 0, 1, 2, ..., n$	$n\theta$	$n\theta(1-\theta)$	Sampling with replacement $Bi(1, \theta) \equiv Bernoulli(\theta)$
Geometric	$\theta \in [0, 1)$	Number of failed i.i.d. Bernoulli trials before first success	$Geom(\theta)$	$p(x) = \theta^x(1-\theta)$ $x = 0, 1, 2, \ldots$	$\frac{\theta}{1-\theta}$	$\frac{\theta^2}{(1-\theta)^2}$	Alternative formulations might swap θ and $1-\theta$, or use $X' = X + 1$ to include the successful trial
Negative Binomial (or Pascal)	$k \in \mathbb{N}$, $\theta \in (0, 1]$	Number of i.i.d. Bernoulli trials until k^{th} success	$NegBin(k, \theta)$ (not standard)	$p(x) = \binom{x-1}{k-1}\theta^k(1-\theta)^{x-k}$ $x = k, k+1, k+2, \ldots$	$\frac{k}{\theta}$	$\frac{k(1-\theta)}{\theta^2}$	Several alternative formulations exist.
Hypergeometric	$N \in \mathbb{N}$ $k \in \{0, \ldots, N\}$ $n \in \{0, \ldots, n\}$	Number of special objects in a random sample of n objects, from a population of N objects with k special objects	$HypGeom(N, k, n)$ (not standard)	$p(x) = \binom{k}{x}\binom{N-k}{n-x}/\binom{N}{n}$ $x = 0, ..., n$	$\frac{nk}{N}$	$n\frac{N-n}{N-1}\frac{k}{N}(1-\frac{k}{N})$	
Poisson	$\lambda \in (0, \infty)$	Counting events occurring 'at random' within space or time	$Poi(\lambda)$	$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $x = 0, 1, 2, ...$	λ	λ	

SOME CONTINUOUS DISTRIBUTIONS

Name	**Parameters**	**Genesis / Usage**	**Notation**	$f(x) =$ **p.d.f.** **(and non-zero range)**	$\mathbb{E}[X]$	$\mathrm{Var}(X)$	**Comments**
Uniform (continuous)	$\alpha, \beta \in \mathbb{R}$ with $\alpha < \beta$	The uniform distribution for a continuous interval	$Unif(\alpha, \beta)$ $U(a, b)$	$f(x) = \frac{1}{\beta - \alpha}$ $x \in (\alpha, \beta)$	$\frac{\alpha+\beta}{2}$	$\frac{(\beta-\alpha)^2}{12}$	Also written as $U[\alpha, \beta]$ and similarly for open and half-open intervals.
Normal	$\mu \in \mathbb{R}$, $\sigma \in (0, \infty)$	Empirically and theoretically (via CLT, etc.) a good model in many situations.	$N(\mu, \sigma^2)$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ $x \in \mathbb{R}$	μ	σ^2	$N(0,1) \equiv$ standard normal. $X \sim N(\mu, \sigma^2) \Rightarrow$ $aX + b \sim N(a\mu + b, a^2\sigma^2)$ Hence $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$
Exponential	$\lambda \in (0, \infty)$	Inter-arrival times of random events	$Exp(\lambda)$	$f(x) = \lambda e^{-\lambda x}$ $x > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	Alternative parametrization: $\theta = \frac{1}{\lambda}$
Gamma	$\alpha, \beta \in (0, \infty)$	Lifetimes of ageing items, multi-inter-arrival times	$Ga(\alpha, \beta)$ $\Gamma(\alpha, \beta)$	$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ $x > 0$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	Alternative parametrization: $\theta = 1/\beta$, $Ga(1, \lambda) \equiv Exp(\lambda)$, $Ga(n/2, 1/2) \equiv \chi^2_n$
Log-normal	$\mu \in \mathbb{R}, \sigma \in (0, \infty)$	Quantities related to exponential growth	$logN(\mu, \sigma^2)$ (not standard)	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log x - \mu)^2}{\sqrt{2}\sigma}\right)$ $x > 0$	$e^{\mu + \frac{1}{2}\sigma^2}$	$(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$	If $X \sim logN(\mu, \sigma^2)$ then $\log X \sim N(\mu, \sigma^2)$
Chi-squared	$n \in \mathbb{N}$	Squared (normally distributed) errors, statistical tests	χ^2_n	$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}$ $x > 0$	n	$2n$	$X^2_n \equiv Ga(n/2, 1/2)$ $X_i \sim N(0,1)$ i.i.d. $\Rightarrow \sum_{i=1}^n X_i^2 \sim \chi^2_n$
Beta	$\alpha, \beta \in (0, \infty)$	Quantities constrained to be within intervals	$Be(\alpha, \beta)$	$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ $x \in [0, 1]$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$	$Be(1,1) \equiv Unif(0,1)$
Cauchy	$a, b \in \mathbb{R}$	Heavy tailed, pathological examples	$Cauchy(a, b)$	$f(x) = \frac{1}{\pi b} \frac{b^2}{(x-a)^2+b^2}$ $x \in \mathbb{R}$	undefined	undefined	$Cauchy(0,1)$ is often called 'the' Cauchy distribution
Pareto	$\alpha, \beta \in (0, \infty)$	Heavy tailed quantities	$Pareto(\alpha, \beta)$	$f(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}$ $x > \beta$	$\frac{\alpha\beta}{\alpha+1}$ if $\alpha > 1$	$\frac{\alpha^2\beta}{(\alpha-1)^2(\alpha-2)}$ if $\alpha > 2$	If $X \sim Pareto(\alpha, \beta)$ then $\log \frac{X}{\beta} \sim Exp(\alpha)$
Student t	$n \in \mathbb{N}$	Statistical tests	t_n	$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1 + \frac{x^2}{n})^{-\frac{n+1}{2}}$ $x \in \mathbb{R}$	0 if $n > 1$	$\frac{n}{n-2}$ if $n > 2$	$t_1 \equiv Cauchy(0,1)$ Can take $n \in (0, \infty)$
F	$\nu, \delta \in (0, \infty)$	Statistical tests	$F_{\nu,\delta}$	$f(x) = \frac{\nu^{\nu/2}\delta^{\delta/2}x^{\nu/2-1}}{B(\nu/2,\delta/2)(\nu x+\delta)^{(\nu+\delta)/2}}$ $x > 0$	$\frac{\delta}{\delta-2}$ if $\delta > 2$	$\frac{2\delta^2(\nu+\delta-2)}{\nu(\delta-2)^2(\delta-4)}$ if $\delta > 4$	If $X \sim \chi^2_\nu$ and $Y \sim \chi^2_\delta$ are independent then $\frac{X/\nu}{Y/\delta} \sim F_{\nu,\delta}$. If $T \sim t_\nu$ then $T^2 \sim F_{1,\nu}$. If $Z \sim Be(\alpha, \beta)$ then $\frac{\beta Z}{\alpha(1-Z)} \sim F_{2\alpha,2\beta}$.