Data provided: Supplementary Formula Sheet

MAS223

The University Of Sheffield.

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

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 (X, Y) be a bivariate random variable with joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} x \exp\left(-\frac{xy}{a(a-2)}\right) & \text{if } x > 0 \text{ and } y > 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $a \in \mathbb{R}$.

(a) Find the possible values of a, given the information above. ***(6 marks)***

(b) Suppose that $a = 1 + \sqrt{2}$. Show that

$$\mathbb{P}[Y > X > 1] = \int_1^\infty e^{-x^2}\, dx.$$

(2 marks)

2 Suppose that $(X, Y)^T$ is a bivariate normal random variable with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ given by

$$\boldsymbol{\mu} = \begin{pmatrix} -1 \\ 5 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}.$$

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

(b) What is the distribution of V? ***(2 marks)***

(c) Does the bivariate random variable $(V, V)^T$ have a bivariate normal distribution? Justify your answer. ***(2 marks)***

3 Let $\alpha > 0$. Let X be a $Be(\alpha, 1)$ random variable, which has probability density function

$$f_X(x) = \begin{cases} \alpha x^{\alpha-1} & \text{if } x \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

and let $Y = -\log(X)$.

Find the probability density function $f_Y(y)$ of Y. What is the name of this distribution? ***(8 marks)***

4 Let X and Y be independent (one dimensional) normal distributions, both with distribution $N(0, 1)$. Let

$$U = \frac{X}{Y}, \qquad V = Y.$$

(a) Find the joint probability density function $f_{U,V}(u, v)$ of (U, V). ***(8 marks)***

(b) Find the probability density function $f_U(u)$ of U. What is the name of this distribution? ***(6 marks)***

(c) Deduce, without any further calculations, that U and $1/U$ have the same distribution. ***(1 mark)***

5 Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a sequence of independent, identically distributed samples of a Gamma distribution $Ga(2, \theta)$, where the parameter $\theta \in (0, \infty)$ is unknown.

(a) Find the likelihood function $L(\theta; \mathbf{x})$ and the corresponding log-likelihood function $\ell(\theta; \mathbf{x})$, of θ given the data $\mathbf{x}$. ***(5 marks)***

(b) Find the maximum likelihood estimator $\hat{\theta}$ for θ, given the data $\mathbf{x}$. ***(5 marks)***

6 Suppose we have some data as follows: $(x_1, y_1) = (0, 0)$, $(x_2, y_2) = (1, 2)$, $(x_3, y_3) = (2, 1)$. We model this as follows:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $i \in \{1, 2, 3\}$ 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 and β_1 are $\hat{\beta}_0 = 1/2$ and $\hat{\beta}_1 = 1/2$, respectively. *(4 marks)*

(c) Find the unbiased estimator of σ^2. *(4 marks)*

(d) Draw a graph of the data-set with the best-fit line superimposed. *(2 marks)*

(e) We wish to test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Find the p-value for this hypothesis test in the form $P(F_{?,?} > ?)$. *(3 marks)*

7 Smoking is believed to be a key predictor of lung cancer. To test this, data were gathered from a variety of professions about the tendency to smoke, using a 'smoking index', and the tendency to die of lung cancer, via a 'mortality index'. We fitted a simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, in `R`, where i indexes the 25 professional groups, y_i is the mortality index of group i (`Mortality`) and x_i is the smoking index of group i (`Smoking`). The data were stored in the `lungcancer` dataframe. Here is some `R` output:

```
> lmlc<-lm(Mortality~Smoking,data=lungcancer)
> summary(lmlc)
Call:
lm(formula = Mortality~Smoking, data = lungcancer)

Residuals:
 Min      1Q       Median  3Q      Max
 -30.107  -17.892  3.145   14.132  31.732

Coefficients:
             Estimate  Std.  Error  t value  Pr(>|t|)
 (Intercept) -2.8853   23.0337      -0.125   0.901
 Smoking     1.0875    0.2209       4.922    5.66e-05

Residual standard error:  18.62 on 23 degrees of freedom
Multiple R-squared:  0.513, Adjusted R-squared:  0.4918
F-statistic:  24.23 on 1 and 23 DF, p-value:  5.658e-05
```

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

- The hypothesis test used to test for a correlation between smoking and mortality from lung cancer
- What you conclude from this hypothesis 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 smoking is a good predictor of death by lung cancer.

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

(b) The mortality index of a group is defined to be the ratio of the rate of deaths from lung cancer in that group to the rate of deaths from lung cancer among all people. Suppose the smoking index for university lecturers is 2. Assuming that the simple linear regression model is a good representation of the data, what would you expect the mortality index to be for university lecturers? Why is this answer unrealistic? How might you re-structure the model to avoid such unrealistic conclusions? ***(4 marks)***

8 You are given some data on cholesterol levels for various people. You want to test for the effect on cholesterol levels of (i) smoking, and (ii) saturated fat intake. The people are split into two groups: smokers, denoted by $i = 1$ and non-smokers, denoted by $i = 2$. Let $x_{i,j}$ be the average daily saturated fat intake of person j within group i. Let $y_{i,j}$ be the cholesterol level of person j within group i.

(a) To test the effect of smoking and saturated fat intake on cholesterol levels, five different models are constructed. Each model can be described both in terms of the assumptions it makes and as a mathematical expression. Below, model M_1 is described in terms of the assumptions made. Models M_0, M_2, 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: $y_{i,j} = \beta_0 + \epsilon_{i,j}$
- Model M_1: Assumes that there is a relationship between smoking and cholesterol level, but not between saturated fat intake and cholesterol level.
- Model M_2: $y_{i,j} = \beta_0 + \beta_1 x_{i,j} + \epsilon_{i,j}$
- 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 model M_1, give the corresponding mathematical expression. For each of models M_0, M_2, M_3, and M_4, describe the assumptions being made.

(8 marks)

(b) Draw the nesting structure for models M_0 to M_4. ***(2 marks)***

(c) Suppose that, instead of splitting the people up into smokers and non-smokers, we use a continuous 'smoking index' that gives each person a (real) number corresponding to the amount that person smokes. Assuming the effects of smoking and saturated fat intake **do not** interact, write down a model to test whether there is a relationship between cholesterol level and both smoking index and saturated fat intake. Ensure you define all your notation carefully. ***(3 marks)***

(d) Why is it not possible to perform a hypothesis test comparing the model in part (c) with model M_3? ***(1 mark)***

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 successful i.i.d. Bernoulli trials before first failure	$Geom(\theta)$ $Geo(\theta)$	$p(x) = \theta^x(1-\theta)$ $x = 0, 1, 2, \ldots$	$\frac{\theta}{1-\theta}$	$\frac{\theta^2}{(1-\theta)^2}$	Alternative formulations might swap θ and $1-\theta$, or use $X' = X + 1$ to include the successful trial
Negative Binomial	$k \in \mathbb{N}$, $\theta \in (0, 1]$	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	$Be(\alpha, \beta)$ $Beta(\alpha, \beta)$	$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ $x \in [0, 1]$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$	$Be(1,1) \equiv Unif(0,1)$
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 $\log \frac{X}{\beta} \sim Exp(\alpha)$
Weibull	$\lambda, k \in (0, \infty)$	Lifetimes, extreme values, particle sizes	$Weibull(\lambda, k)$	$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$ $x > 0$	$\lambda\Gamma(1 + 1/k)$	$\lambda^2[\Gamma(1 + 2/k)$ $+\Gamma(1 + 1/k)^2]$	If $X \sim Weibull(\lambda, k)$ then $(X/\lambda)^k \sim Exp(1)$
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 $\frac{\beta Z}{\alpha(1-Z)} \sim F_{2\alpha,2\beta}$.