Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2024	MATH2711-WE01

Title: Statistical Inference II

Time:	3 hours	
Additional Material provided:	Formula Sheet; Tables: Normal distribution, t-distribution, chi-squared distribution, signed-rank test statistic, rank-sum test statistic.	
Materials Permitted:		
Calculators Permitted:	Yes	Models Permitted: Casio FX83 series or FX85 series.

Instructions to Candidates:	Answer all questions. Section A is worth 40% and Section B is worth 60%. Within each section, all questions carry equal marks. Students must use the mathematics specific answer book.

Revision:	

SECTION A

Q1 The measurements of various attributes of houseflies were studied by Sokal and Hunter (1955). A random sample of 25 independent houseflies was taken and the length of their wings, X, was measured in millimetres. Summary statistics of those observations are as follows:

$$\sum_{i=1}^{25} x_i = 1197, \qquad \sum_{i=1}^{25} x_i^2 = 57671.$$

1.1 Find a 99% confidence interval for the population mean.

1.2 Find the approximate probability that the sample variance over-estimates the population variance by at least 75%.

1.3 State any assumptions required for your calculations, and briefly comment on how those assumptions could be verified.

Q2 The independent random variables X and Y have chi-squared distributions with a and b degrees of freedom respectively. The probability density function for the chi-squared distribution can be found on the provided Formula Sheet.

Consider the transformations $W = X + Y$ and $Z = X/(X + Y)$.

2.1 Find the joint probability density function of W and Z. Are W and Z independent?

2.2 Find the probability density function of Z alone, identify the distribution of Z with a standard distribution, and clearly state its parameters.

Q3 The Laplace distribution is a continuous probability distribution with p.d.f. given by

$$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right),$$

where $x \in \mathbb{R}$, $\mu \in \mathbb{R}$ and $b > 0$. Assuming that the location parameter μ is known and fixed:

3.1 Show that $f(x \mid b)$ belongs to the 1-parameter exponential family and clearly identify all of the exponential family components.

3.2 Suppose that we observe a sample of i.i.d. observations $\mathbf{x} = (x_1, \ldots, x_n)^T$. Show that $f(\mathbf{x} \mid b)$ also belongs to the 1-parameter exponential family, identifying again all the relevant components.

3.3 Based on your results from **3.1**, use the properties of the exponential family to express the mean and the variance of the random variable $|X - \mu|$ as a function of b.

Q4 The Borel distribution is a discrete probability distribution with probability mass function given by

$$\mathbb{P}[X = x \mid \mu] = f(x|\mu) = \frac{1}{x!} e^{-\mu x} (\mu x)^{x-1},$$

where $x \in \{1, 2, 3, \ldots\}$ and $\mu \in [0, 1]$.

4.1 Show that the maximum-likelihood estimator of μ for an i.i.d. sample of observations $x_1, \ldots, x_n$ is given by

$$\hat{\mu} = 1 - \frac{1}{\bar{x}},$$

where $\bar{x}$ is the sample mean.

4.2 Assume now that we observe the following table of count data:

x	1	2	3	4	5 or more
Occurrences	93	30	10	10	7

Assess the fit of the Borel distribution to these data using the chi-square goodness-of-fit test at a significance level of 5%.

SECTION B

Q5 Kimber (1995) studied the shape of birds' eggs, exploring the relationships between their dimensions and attributes. One of her research questions was whether the unknown height of a bird's egg, Y, could be related to the known width of the egg, x, via a Normal distribution. For a sample of n independent eggs, we express this hypothesised relationship as:

$$Y_i \sim \mathrm{N}\left(\theta x_i, \frac{1}{\tau} = \sigma^2\right),$$

for $i = 1, \ldots, n$, and where θ and τ are unknown parameters.

5.1 Find the likelihood function for the parameters θ and τ, and identify sufficient statistics for the parameters given data $(x_1, \ldots, x_n)^T$. Justify your answer.

5.2 Derive the maximum likelihood estimators $\hat{\theta}$ for θ, and $\hat{\tau}$ for τ. (You are not required to verify that you have found a maximum.)

5.3 Is $\hat{\theta}$ an unbiased estimator for θ?

5.4 Assuming that the parameter τ is known, find the exact sampling distribution of $\hat{\theta}$, and hence find an exact 95% confidence interval for θ. You should clearly state any results used in order to construct the interval.

Q6 Suppose we wish to model the total rainfall, X_i, (in mm) falling in Durham during the same month over the last n years. To do this we will use the Gamma distribution, which has commonly been used in meteorological problems to model the levels of rainfall at single locations.

Assume that the rainfall measurements, X_i, have a Gamma$(2, \beta)$ distribution, and that the X_i are conditionally independent given β.

6.1 Determine whether a Gamma prior for β is conjugate for this problem, and identify the posterior distribution and its parameters in terms of a standard distribution.

6.2 Find the Jeffreys prior for this problem, and comment on it in relation to your answer to **6.1**.

6.3 The total June rainfall measurements (in mm) recorded by the Durham University Observatory every year from 2014 to 2023 are given below:

47.2 28.4 49.8 103.4 31.2 108.4 82.6 28.8 39.8 50.4

Summary statistics for the data are: $\sum_{i=1}^{n} x_i = 570$ and $\sum_{i=1}^{n} x_i^2 = 40706.4$.

Use the data above to compute and compare the posterior distributions for β obtained when using (i) a Gamma$(1, 10)$ prior from **6.1**, and (ii) the Jeffreys prior from **6.2**.

Discuss the behaviour of the posterior distributions as the sample size, n, grows large.

6.4 Using a Gamma prior and the available data $x_1, x_2, \ldots, x_n$, find the probability density function of the next rainfall measurement, X^*, given the observed data. Clearly state any assumptions you have made.

Q7 A company is interested in assessing the quality of the circuits used in its manufacturing plants. The quality of the circuits is quantified by their lifetime which is modelled by a specific form of the Lognormal distribution with p.d.f. given by

$$f(x \mid \theta) = \sqrt{\frac{\theta}{2\pi}} \frac{1}{x} \exp\left(-\frac{\theta}{2}(\log x)^2\right),$$

where $x > 0$ and $\theta > 0$. Specifically, if we denote by X the lifetime of a circuit, then $X \sim \text{Lognormal}(\theta)$ and $\mathbb{E}[X] = e^{1/2\theta}$.

7.1 Circuit lifetimes were recorded in one of the company's manufacturing plants leading to a sample of n i.i.d. lifetime measurements $\mathbf{x} = (x_1, \ldots, x_n)^T$. The focus is on tests of the form $\mathcal{H}_0 : \mathbb{E}[X] = c_0$ against $\mathcal{H}_1 : \mathbb{E}[X] = c_1$ for fixed values $c_0 > c_1 > 1$. Express this hypothesis test mathematically in terms of the parameter θ. Derive the most powerful test and show that this is equivalent to a rejection rule of the form $\ell \leq k$, where $\ell = \sum_{i=1}^{n} (\log x_i)^2$ and k is an appropriate constant.

7.2 Derive the universally most powerful test of the hypothesis $\mathcal{H}_0 : \mathbb{E}[X] = c_0$ versus the general alternative $\mathcal{H}_1 : \mathbb{E}[X] < c_0$, working within the context of **7.1**.

7.3 The company operates two manufacturing plants in total. Lifetimes of circuits were recorded in each manufacturing plant resulting in a collection of n i.i.d. observations on X and m i.i.d. observations on Y where X and Y are independent and

$$X \sim \text{Lognormal}(\theta_X),$$
$$Y \sim \text{Lognormal}(\theta_Y).$$

The company now wishes to assess whether the performance in the two manufacturing plants is the same. Show that the generalised likelihood ratio test statistic of the hypothesis $\mathcal{H}_0 : \mathbb{E}[X] = \mathbb{E}[Y]$ versus the alternative $\mathcal{H}_1 : \mathbb{E}[X] \neq \mathbb{E}[Y]$ is given by

$$\Lambda(\mathbf{x}, \mathbf{y}) = \ell_x^{\frac{n}{2}} \ell_y^{\frac{m}{2}} \left(\frac{n+m}{n\ell_x + m\ell_y}\right)^{\frac{n+m}{2}},$$

where $\mathbf{x} = (x_1, \ldots, x_n)^T$, $\mathbf{y} = (y_1, \ldots, y_m)^T$, $\ell_x = \sum_{i=1}^{n} (\log x_i)^2$ and $\ell_y = \sum_{j=1}^{m} (\log y_j)^2$.

Q8 Two basic models for survival-time data are based on the Gamma and Weibull distributions. Consider the setting where we are given a sample $\mathbf{x} = (x_1, \ldots, x_n)^T$ of n observations and we are interested in the following Bayesian model comparison

$$\mathcal{M}_0 : \begin{cases} x_i \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \beta) \\ \beta \sim \text{Gamma}(a, b) \\ \alpha > 0 \text{ (fixed parameter)} \end{cases} \quad \text{vs.} \quad \mathcal{M}_1 : \begin{cases} x_i \overset{\text{i.i.d.}}{\sim} \text{Weibull}(k, \lambda) \\ \lambda \sim \text{InvGamma}(c, d) \\ k > 0 \text{ (fixed parameter)} \end{cases}$$

where $\beta, a, b, \lambda, c, d$ are all positive and $x_i \in (0, \infty)$ for $i = 1, \ldots, n$. Note that $\text{InvGamma}(\cdot, \cdot)$ denotes the inverse-gamma distribution and that the p.d.f. of the Weibull distribution, for fixed k, is given by

$$f(x \mid \lambda) = \frac{k}{\lambda} x^{k-1} e^{-x^k/\lambda}, \text{ for } x > 0.$$

8.1 Derive the joint sampling distribution of $\mathbf{x}$ under models $\mathcal{M}_0$ and $\mathcal{M}_1$.

8.2 Show that the Bayes factor in favour of $\mathcal{M}_0$ against $\mathcal{M}_1$ in terms of the data, the sample size n and the prior parameters a, b, c, d can be expressed as

$$\text{B}_{01} = (k\Gamma(\alpha))^{-n} \left(\prod_{i=1}^{n} x_i^{\alpha - k} \right) \frac{b^a \Gamma(c)}{d^c \Gamma(a)} \frac{\Gamma(\alpha n + a)}{\Gamma(n + c)} \frac{(\sum_{i=1}^n x_i + d)^{n+c}}{(\sum_{i=1}^n x_i + b)^{\alpha n + a}}.$$

8.3 Assume now that we want to specify the prior parameters in $\mathcal{M}_0$ and $\mathcal{M}_1$ such that $\mathbb{E}[\beta] = \mathbb{E}[\lambda] = 1/2$ and $\mathbb{V}\text{ar}[\beta] = \mathbb{V}\text{ar}[\lambda] = 1/4$. Find the values of the pairs (a, b) and (c, d) that satisfy these requirements.

8.4 Suppose that we observe a sample of 5 observations that yield the quantities $\sum_{i=1}^5 x_i = 4.57$ and $\prod_{i=1}^5 x_i = 0.05$, and that by design $\alpha = 2$ and $k = 1$. Based on the Bayes factor from **8.2** and the prior specification from **8.3** calculate the quantity $2 \log_e(B_{10})$ and infer whether there is any evidence against model $\mathcal{M}_0$. Subsequently, calculate the resulting posterior probabilities of the two models assuming equal model probabilities *a-priori*. Which model would you select based on your results?