Durham University

EXAMINATION PAPER

Examination Session: May/June	**Year:** 2023	**Exam Code:** 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 Measurements, X_i, are made on the temperature, in $^\circ C$, achieved during twelve firings of a ceramic kiln. To adequately fire pottery and harden the clay, temperatures of $950 - 1050^\circ C$ are required. Assume that the measurements may be regarded as twelve independent observations from a normal distribution with unknown mean μ and unknown variance σ^2. The data are as follows.

$$951, 965, 960, 979, 978, 1016, 947, 949, 970, 958, 1011, 982,$$

with summaries $\sum_{i=1}^{12} x_i = 11\ 666$, $\sum_{i=1}^{12} x_i^2 = 11\ 346\ 926$.

1.1 Find, approximately, the probability that the sample variance overestimates σ^2 by at least 25%.

1.2 Calculate a 99% confidence interval for μ, and a 99% confidence interval for σ^2.

1.3 Consider the statement: "there is an almost 100% chance that μ is between 950 and 1050, and so the kiln temperature will be in the desired range". Is this statement correct? Explain your answer, but no further calculation is required.

Q2 Suppose X_1, X_2, X_3 are independent and identically distributed random variables with common p.d.f.

$$f(x) = e^{-x}, \quad x > 0,$$

and zero otherwise. Consider the following transformations: $Y_1 = X_1 + X_2 + X_3$, $Y_2 = \dfrac{X_2}{X_1 + X_2 + X_3}$, and $Y_3 = \dfrac{X_3}{X_1 + X_2 + X_3}$.

2.1 Find the joint p.d.f. of $\boldsymbol{Y} = [Y_1, Y_2, Y_3]^T$.

2.2 Are Y_1 and Y_2 independent? Justify your answer.

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

$$f(x \mid a, b) = abx^{a-1}(1 - x^a)^{b-1},$$

where $x \in (0, 1)$ and $a > 0$, $b > 0$ are shape parameters. Assuming that $a = 1$:

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 $-\log(1 - X)$ as a function of b.

Q4 A small experiment was conducted to test whether the installation of cavity-wall insulation has an effect on the amount of energy consumed in houses. Ten houses were selected from a housing estate, of which five are selected at random for insulation. The total energy consumption over one winter is measured for each house. The data, in megawatt hours, are as follows.

No insulation	12.64	11.85	12.82	11.37	14.42
Insulation	12.91	9.92	9.52	10.02	10.38

4.1 Perform a non-parametric rank-sum test to investigate the null hypothesis that the insulation has no effect on energy consumption. Use the exact distribution of the test statistic and conduct the test at the 1% level of significance.

4.2 Perform an independent sample t-test for the same hypothesis, again at the 1% level of significance.

SECTION B

Q5 An experiment was performed to investigate the uptake of calcium by cells that had been in “hot” calcium suspension. A total of twenty-seven observations were made at various times after the start of the experiment. Denoting the calcium uptake by y (nmoles/mg) at time t (minutes) after the start of the experiment, theoretical considerations suggest that y is normally distributed with mean $\beta_1[1-\exp(-t/\beta_2)]$ and variance σ^2. You may assume that the uptake measurements are independent.

The log-likelihood function, $\mathcal{L}(\boldsymbol{\theta})$, for this problem was maximised numerically using the experimental data, giving the following results:

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} 4.3094 \\ 4.7967 \\ 0.5258 \end{bmatrix}, \qquad \mathcal{L}''(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} -46.8992 & 13.5902 & -0.0016 \\ 13.5902 & -5.2748 & 0.0008 \\ -0.0016 & 0.0008 & -195.3281 \end{bmatrix},$$

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) = 3.8566, \qquad (\mathcal{L}''(\hat{\boldsymbol{\theta}}))^{-1} = \begin{bmatrix} 0.0841 & 0.2168 & 0.0000 \\ 0.2168 & 0.7481 & 0.0000 \\ 0.0000 & 0.0000 & 0.0051 \end{bmatrix},$$

where $\boldsymbol{\theta} = [\beta_1, \beta_2, \sigma]$ and $\hat{\boldsymbol{\theta}}$ is the location of the maximum of $\mathcal{L}$.

5.1 Write down the the probability density function for an individual observation and deduce the log-likelihood function for the experiment.

5.2 Find an approximate 95% confidence interval for σ.

5.3 Show that for any fixed σ, the log-likelihood is maximised by the same values of β_1 and β_2. Hence show that the profile log-likelihood for σ can be written as

$$L_\sigma(\sigma) = \text{constant} - \frac{n}{2}\left[\log\sigma^2 + \frac{\hat{\sigma}^2}{\sigma^2}\right]$$

Compute the large-sample standard deviation of $\hat{\sigma}$ directly from the profile log-likelihood $L_\sigma(\sigma)$, and compare your answer with that obtained in **5.2**.

5.4 Find the maximum likelihood estimate and a 95% CI for the expected calcium uptake after 5 minutes from the start of the experiment.

Q6 A genetic study to estimate the frequency of a gene in human DNA is undertaken, where the gene can take only one of two forms (alleles): A or B. DNA sequences for this location are provided for n individuals. We denote the observed number of A alleles in these observed sequences by X, with an underlying unknown allele frequency of $\theta \in [0, 1]$.

A model for this problem is proposed based on a binomial distribution for X, where $X|\theta \sim \text{Bin}(n, \theta)$. Additionally, we assume a Beta prior distribution $\theta \sim \text{Beta}(\alpha, \beta)$ where $\alpha > 0$ and $\beta > 0$ are known constants.

6.1 Due to the expense of the DNA sequencing, only a small sample of $n = 3$ sequences could be performed. The A allelle was observed in all three samples.

Using the Beta prior with parameters $\alpha = 2$ and $\beta = 1$, find *exact* 95% prior and posterior equal-tailed credible intervals for θ. Compare the two intervals and comment on how the distribution for θ has changed after learning from the data.

6.2 Derive the Jeffreys prior for this problem, and find the corresponding posterior distribution for θ. Without further calculation, comment on how the 95% equal-tailed credible intervals constructed using this prior would compare to those found in **6.1**.

6.3 For some genetic markers, the assumption of a Beta prior may be too restrictive and a prior density with two peaks might be more appropriate. For example, we could generate such a distribution combining two Beta distributions together as follows:

$$f(\theta) = w f_{Be}(\theta|\alpha_1, \beta_1) + (1 - w) f_{Be}(\theta|\alpha_2, \beta_2)$$

where $f_{Be}(\theta|\alpha, \beta)$ denotes the p.d.f. of a Beta distribution with parameters α, β, and $w \in (0, 1)$ is a known constant weight parameter.

Using this prior, derive the posterior distribution of θ. Express the posterior distribution in terms of a similar combination of standard distributions, clearly identify their parameters, and the corresponding posterior weight parameters.

Q7 An engineer is interested in ensuring the successful operation of the production lines in a manufacturing plant. Measuring the time between repairs of a given production line as X, a model of an exponential distribution is agreed, with p.d.f. given by

$$f(x \mid \lambda) = \lambda e^{-\lambda x},$$

where $x > 0$ and $\lambda > 0$ is the rate parameter, so that $\mathbb{E}[X] = 1/\lambda$. Generally, the expected time between repairs is approximately three months.

7.1 The engineer suspects problems with this production line and records an i.i.d. sample of n inter-repair times $\mathbf{x} = (x_1, \ldots, x_n)^T$, intending to test the null hypothesis that the expected time between repairs is indeed three months (1/4 of a year) versus an alternative hypothesis under which the expected time between repairs is two months (1/6 of a year). Express this hypothesis test mathematically in terms of the parameter λ. Derive the most powerful test the engineer can use and show that this is equivalent to a rejection rule of the form $\bar{x} < k$, for an appropriate constant k.

7.2 Subsequently, the engineer wishes to test the simple null hypothesis in **7.1** versus a composite alternative hypothesis which generally states that the expected repair time is less than three months. Express, again, this test mathematically, in terms of λ, and derive the uniformly most powerful test.

7.3 Across the entire manufacturing plant, there are two production lines - the original line X and a further line Y. Times between repairs were recorded for each line giving i.i.d. observations $\mathbf{x} = (x_1, \ldots, x_n)^T$ from Line X and i.i.d. observations $\mathbf{y} = (y_1, \ldots, y_m)^T$ from Line Y, where $\mathbf{x}$ and $\mathbf{y}$ are independent and inter-repair times follow exponential distributions with respective rate parameters λ_X and λ_Y; that is, $x_i \sim \text{Exp}(\lambda_X)$ and $y_j \sim \text{Exp}(\lambda_Y)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

Now, the engineer wishes to assess whether the two production lines behave differently; thus, testing $\mathcal{H}_0 : \lambda_X = \lambda_Y$ vs. $\mathcal{H}_1 : \lambda_X \neq \lambda_Y$. Find the likelihood functions and the MLEs under the null and alternative hypotheses.

7.4 Derive the generalised likelihood ratio test for the hypothesis test stated in **7.3**, simplifying the test statistic as far as possible.

Q8 Consider the setting where we have a Weibull sampling distribution with a known (fixed) positive parameter k, and p.d.f. given by

$$f(x \mid \lambda) = \frac{k}{\lambda} x^{k-1} e^{-x^k/\lambda},$$

where $x \in [0, \infty)$ and $\lambda > 0$. We are given a sample $\mathbf{x} = (x_1, \ldots, x_n)^T$ of n observations which we assume to be conditionally i.i.d. given parameter λ.

8.1 Derive the joint sampling distribution of $\mathbf{x}$ given λ.

8.2 Consider the hypothesis test $\mathcal{H}_0 : \lambda = 1$ vs. $\mathcal{H}_1 : \lambda = 2$ and derive the Bayes factor in favour of $\mathcal{H}_0$ against $\mathcal{H}_1$.

8.3 Assume now that we want to test the null hypothesis $\mathcal{H}_0 : \lambda = 1$ vs. a more general alternative of the form $\mathcal{H}_1 : \lambda > 0$. Under the alternative we specify an inverse-Gamma distribution for parameter λ, with parameters $a = b = 1$. Derive the Bayes factor in favour of $\mathcal{H}_0$ against $\mathcal{H}_1$.

8.4 Consider the framework where we want to compare a model $\mathcal{M}_0$ with an inverse-Gamma prior on λ, as defined in **8.3**, versus a model $\mathcal{M}_1$ with an inverse-Gamma prior with parameters a and b such that the prior mode and mean are $b/(a+1) = 1$ and $b/(a-1) = 3$, respectively. Find the corresponding prior distribution under $\mathcal{M}_1$, briefly describe this model comparison framework mathematically and derive the Bayes factor in favour of $\mathcal{M}_0$ against $\mathcal{M}_1$.

8.5 Suppose that we observe $\mathbf{x} = (0.95, 1.16, 1.37)^T$ and that by design $k = 1$. Calculate the Bayes factor of **8.4** and the resulting posterior probabilities of the two models assuming equal model probabilities *a-priori*. Which model would you select?