Imperial College
London

M2S2

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)

May-June 2019

This paper is also taken for the relevant examination for the Associateship of the Royal College of Science

Statistical Modelling 1

Date: Friday 17 May 2019

Time: 14.00 - 16.00

Time Allowed: 2 Hours

This paper has 4 Questions.

Candidates should start their solutions to each question in a new main answer book.

Supplementary books may only be used after the relevant main book(s) are full.

All required additional material will be provided.

- DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO.
- Affix one of the labels provided to each answer book that you use, but DO NOT USE THE LABEL WITH YOUR NAME ON IT.
- Calculators may not be used.

1. (a) Suppose $\theta \in \mathbb{R}$ is the unknown parameter in a statistical model and suppose that T is an estimator for θ.

   (i) Define the bias and the mean squared error of T.

   (ii) Write down the equation that relates the bias, variance and mean squared error.

   (iii) Prove that the equation from Part (ii) holds true.

   (b) There is a bag containing three labeled balls, $i = 1, 2, 3$, each of which also has a number written on it, which we denote by y_i. A random sample is selected by drawing a set $\mathcal{S}$ of two balls from the bag without replacement. These two balls are then used to estimate the sum of the balls in the bag, i.e. $t = y_1 + y_2 + y_3$.

   (i) Define the term *random sample*.

   (ii) Prove that the estimator $\widehat{t}(\mathcal{S})$ defined by

$$\widehat{t}(\mathcal{S}) = \begin{cases} (3/2)y_1 + (3/2)y_2 & \text{if } \mathcal{S} = \{1, 2\} \\ (3/2)y_1 + 2y_3 & \text{if } \mathcal{S} = \{1, 3\} \\ (3/2)y_2 + y_3 & \text{if } \mathcal{S} = \{2, 3\} \end{cases}$$

   is unbiased and hence calculate its mean squared error.

   (c) We might also consider a Bayesian approach to statistical modelling, which requires us to additionally define a prior distribution. In particular, we might choose a prior that is conjugate to the likelihood for convenience.

   (i) Define the term *conjugate prior*.

   (ii) Show that the beta distribution is a conjugate prior for a likelihood that follows a Binomial(N, θ), where N is known and θ is the unknown parameter of interest.
   Recall: The probability density function (pdf) of a random variable $Z \sim beta(a, b)$ *is* $f(z) = \frac{(a+b+1)!}{(a-1)!(b-1)!} z^{a-1}(1-z)^{b-1}$, *for* $z \in [0, 1]$, $a > 0$, $b > 0$.

2. You are out hiking in the Scottish hills and you come across two bee hives close to one another. You wonder whether you could infer anything about their relative populations by considering the frequency with which the bees leave each hive. To investigate this you decide to record the time between departures for each of these two hives.

Suppose you observe n_1 inter-departure times for bee hive number 1 and n_2 inter-departure times for bee hive number 2. You decide to let $y_{11}, \ldots, y_{1n_1}$ denote the inter-departure times for hive number 1 and to let $y_{21}, \ldots, y_{2n_2}$ denote the inter-departure times for hive number 2.

You decide to work with the model

$$Y_{11}, \ldots, Y_{1n_1} \sim \text{Exp}(\lambda_1), \quad Y_{21}, \ldots, Y_{2n_2} \sim \text{Exp}(\lambda_2),$$

where all random variables are independent and the unknown parameters are $\lambda_1 > 0, \lambda_2 > 0$.

Consider the hypotheses

$$H_0 : \lambda_1 = \lambda_2 \quad \text{against} \quad H_1 : \lambda_1 \neq \lambda_2.$$

Recall: The probability density function (pdf) of a random variable $Z \sim Exp(\lambda)$ *is* $f(z) = \lambda \exp(-\lambda z)$ *for* $z > 0$.

(a) Write down and simplify the likelihood.

(b) Derive the maximum likelihood estimator under the assumption that H_0 is true.

(c) Derive the maximum likelihood estimator under the full model.

(d) Describe a likelihood ratio test for the above hypotheses. Clearly describe (and simplify if possible) the test statistic, and state the decision rule.

(e) Suppose your test rejects the null hypothesis. What can you conclude about the relative size of the two bee populations?

3. (a) Consider a sequence of estimators $(T_n)_{n\in\mathbb{N}}$. Write down the definition for each of the following concepts.

(i) $(T_n)_{n\in\mathbb{N}}$ is consistent,

(ii) $(T_n)_{n\in\mathbb{N}}$ is asymptotically unbiased.

(b) Suppose that $X_1, ..., X_n$ are independent and identically distributed Poisson(λ) random variables.

Recall: The probability mass function (pmf) of a random variable $Z \sim Poisson(\lambda)$ *is* $f(z) = \frac{\lambda^z \exp(-\lambda)}{z!}$ *for* $z = 0, 1, 2, ...,\ \lambda > 0$.

(i) Find the maximum likelihood estimator of λ.

(ii) Find the maximum likelihood estimator of $\exp(-\lambda)$.

(iii) Find an approximation for the distribution of the estimator in part (b)(i) above.

(c) A component of type A_i functions with probability θ_i, $i = 1, 2$, where θ_1, θ_2 are unknown. Mechanisms of type X consist of an A_1-component in series with an A_2-component and such a mechanism functions if and only if both components function. Mechanisms of type Y are similar except that they consist of one type A_1-component in series with two A_2-components (themselves in series). All components act independently.

Let ϕ_1 denote the probability that a type X mechanism functions, and let ϕ_2 denote the probability that a type Y mechanism functions. A random sample of n type X mechanisms and an independent random sample of n type Y mechanisms is selected. Of these, n_1 and n_2 function respectively. We also assume that the conditions $n_2 \leq n_1$ and $n_1^2 \leq nn_2$ hold.

(i) Find the maximum likelihood estimators of ϕ_1, ϕ_2.

(ii) Using the answer from part (i) above, find the maximum likelihood estimators of θ_1, θ_2.

(iii) Why is the MLE not straightforward to compute if the two conditions stated are not satisfied?

4. (a) Using vector/matrix notation, write down the general form of a linear model, then state the role of each term in the equation and its dimensions.

(b) In a linear model with second order assumptions as stated in the lectures and a full rank design matrix X, derive the covariance matrix $\text{cov}(\widehat{\boldsymbol{\beta}})$ of the least squares estimator $\widehat{\boldsymbol{\beta}}$.

(c) (i) State the Gauss Markov Theorem.

(ii) Give a specific and concrete example of a modelling problem that could be tackled with a linear model. State clearly what you are trying to estimate and describe whether or not the Gauss Markov Theorem would influence your choice of estimator.

(d) Let us assume the full rank and Normal theory assumptions. State the distribution of the following test statistic,

$$T = \frac{\mathbf{c}^T\widehat{\boldsymbol{\beta}} - \mathbf{c}^T\boldsymbol{\beta}}{\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}\frac{\text{RSS}}{n-p}}}$$

Write the test statistic as a fraction and state the properties of the denominator and numerator that allow us to show that T does indeed have the distribution you have stated.

M2S2 SOLUTIONS

NOTE: In addition to the marks for each question, a letter denotes the approximate level of difficulty of the marks. A indicates the basic, routine material (easiest 40 percent), B indicates the marks for demonstration of sound knowledge (next 25 percent), C indicates the harder material (next 15 percent) and D indicates the most challenging material (hardest 20 percent).

seen ⇓

1. (a) (i) $\text{bias}_\theta(T) = \mathrm{E}_\theta(T) - \theta$. $\text{MSE}_\theta(T) = \mathrm{E}_\theta((T-\theta)^2)$.

2A

(ii) $\text{MSE}_\theta(T) = \text{Var}_\theta(T) + (\text{bias}_\theta(T))^2$.

2A

(iii) Let $X = T - \theta$. Then $\text{Var}_\theta(X) = \mathrm{E}_\theta(X^2) - \mathrm{E}_\theta(X)^2 = \text{MSE}_\theta(T) - (\text{bias}_\theta(T))^2$. Rearranging shows the relationship.

3B

(b) (i) A random sample is a collection of independent and identically distributed (iid) samples drawn from a given probability distribution.

1A

(ii) Each of the randomly sampled pairs has probability $\frac{1}{3}$, therefore
$E[\widehat{t}] = \frac{1}{3}(\frac{3}{2}y_1 + \frac{3}{2}y_2 + \frac{3}{2}y_1 + 2y_3 + \frac{3}{2}y_2 + y_3) = y_1 + y_2 + y_3$
The given estimator for $y_1 + y_2 + y_3$ is therefore unbiased.

2A

The MSE is therefore given only by the variance of the estimator, since the bias is zero. We can calculate the variance by $\text{Var}[\widehat{t}] = E[\widehat{t}^2] - E[\widehat{t}]^2$.

$$\begin{aligned} E[\widehat{t}^2] &= \frac{1}{3}([\frac{3}{2}y_1 + \frac{3}{2}y_2]^2 + [\frac{3}{2}y_1 + 2y_3]^2 + [\frac{3}{2}y_2 + y_3]^2) \\ &= \frac{3}{2}y_1^2 + \frac{3}{2}y_2^2 + \frac{5}{3}y_3^2 + \frac{3}{2}y_1y_2 + 2y_1y_3 + y_2y_3 \end{aligned}$$

$$\begin{aligned} E[\widehat{t}]^2 &= (y_1 + y_2 + y_3)^2 \\ &= y_1^2 + y_2^2 + y_3^2 + 2y_1y_2 + 2y_1y_3 + 2y_2y_3 \end{aligned}$$

4B

$$\begin{aligned} MSE[\widehat{t}] &= \text{Var}[\widehat{t}] \\ &= E[\widehat{t}^2] - E[\widehat{t}]^2 \\ &= \frac{1}{2}y_1^2 + \frac{1}{2}y_2^2 + \frac{2}{3}y_3^2 - \frac{1}{2}y_1y_2 + y_2y_3 \end{aligned}$$

(c) (i) A family of prior probability distributions P is said to be conjugate to a family of observational distributions L, if for every prior $p \in P$ and every observational distribution $l \in L$, the resulting posterior distribution also belongs to P. seen ⇓ 2A

(ii) Let $y \in \{0, \ldots, N\}$ be the the number of successful outcomes from N trials. Then unseen ⇓

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto c_1\theta^y(1-\theta)^{N-y} \times c_2\theta^{a-1}(1-\theta)^{b-1} \\ &\propto c_3\theta^{a+y-1}(1-\theta)^{b+N-y-1} \end{aligned}$$

where we note that an exact expression for the posterior follows from using the constants $c_1 = \frac{N!}{y!(N-y)!}$, $c_2 = \frac{(a+b-1)!}{(a-1)!(b-1)!}$, and $c_3 = \frac{(a+b+N-1)!}{(a+y-1)!(b+N-y-1)!}$.

i.e. $p(\theta|y) = \text{beta}(a+y, b+N-y)$. 4C

2. (a) The likelihood may be written as

sim. seen ⇓

$$L(\lambda_1, \lambda_2) = \prod_{j=1}^{n_1} \lambda_1 \exp(-\lambda_1 Y_{1j}) \prod_{j=1}^{n_2} \lambda_2 \exp(-\lambda_1 Y_{2j})$$
$$= \lambda_1^{n_1} \exp\left(-\lambda_1 \sum_{j=1}^{n_1} Y_{ij}\right) \lambda_2^{n_2} \exp\left(-\lambda_2 \sum_{j=1}^{n_2} Y_{ij}\right).$$

3A

(b) In order to find the MLE under the null hypothesis, i.e. under H_0, we have $\lambda_1 = \lambda_2$. Denoting this joint parameter by λ, the likelihood simplifies to

$$L(\lambda) = \lambda^{n_1+n_2} \exp\left(-\lambda \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}\right).$$

This is the likelihood for iid observations from Exp(λ). The MLE then follows as

$$\log L(\lambda) = (n_1 + n_2) \log(\lambda) - \lambda \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}$$

and

$$\frac{\partial}{\partial \lambda} \log L(\lambda) = \frac{n_1 + n_2}{\lambda} - \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}.$$

Equating this expression to zero and solving for λ gives

$$\widehat{\lambda} = \frac{n_1 + n_2}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}}$$

4B

which we confirm is indeed the MLE since

$$\left(\frac{\partial}{\partial \lambda}\right)^2 \log L(\lambda) = -\frac{n_1 + n_2}{\lambda^2} < 0.$$

2A

(c) In order to find the MLE under the full model, we note that the factors in the likelihood are nonnegative and only contain different components of the parameter. We can therefore maximise them separately. Each element is the likelihood of iid observations from Exp(λ_i), and so the same argument as above holds, resulting in the MLE

$$\widehat{\lambda}_i = \frac{n_i}{\sum_j Y_{ij}}, \quad i = 1, 2.$$

4B

(d) The test statistic of the likelihood ratio test is

$$t = \frac{\sup_{\lambda_1,\lambda_2} L(\lambda_1, \lambda_2)}{\sup_{\lambda} L(\lambda)}.$$

Using the above, the denominator simplifies to

$$L(\widehat{\lambda}) = \left(\frac{n_1 + n_2}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}} \right)^{n_1+n_2} \exp(-n_1 - n_2).$$

The numerator simplifies to

$$\sup_{\lambda_1,\lambda_2} L(\lambda_1, \lambda_2) = L(\widehat{\lambda}_1, \widehat{\lambda}_2) = \left(\frac{n_1}{\sum_{i=1}^{n_1} Y_{1i}} \right)^{n_1} \left(\frac{n_2}{\sum_{i=1}^{n_1} Y_{2i}} \right)^{n_2} \exp(-n_1 - n_2).$$

Hence,

$$t = \frac{\left(\frac{n_1}{\sum_{i=1}^{n_1} Y_{1i}} \right)^{n_1} \left(\frac{n_2}{\sum_{i=1}^{n_1} Y_{2i}} \right)^{n_2}}{\left(\frac{n_1+n_2}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}} \right)^{n_1+n_2}}.$$

This can be simplified as $t = \frac{\bar{Y}^{n_1+n_2}}{\bar{Y}_1^{n_1} \bar{Y}_2^{n_2}}$ where $\bar{Y} = \frac{1}{n_1+n_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}$ and $\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$. 3D

From the lectures we know

$$2 \log t \xrightarrow{d} \chi_1^2,$$

as n_1 and n_2 tend to ∞.

Therefore in order to get an asymptotic test to the level α one would reject H_0 if $2 \log t > c$ where c is such that $\mathrm{P}(Z > c) = \alpha$ with $Z \sim \chi_1^2$. 2D

(e) If the test rejects the null hypothesis, we may argue this provides evidence that the two hive populations are indeed of different sizes. (1 mark) Note that no marks should be given if the student answers that one hive definitely is larger than the other.

However, this inference is critically based on a number of assumptions: namely, that our model is appropriate, that the inter-arrival times are indeed proportional to the size of the population, that we haven't obtained a false positive result through chance alone, that we didn't make any mistakes in the data collection etc. (1 mark) 2D

3. (a) (i) A sequence of estimators $(T_n)_{n\in\mathbb{N}}$ for $g(\theta)$ is called consistent if for all $\theta \in \Theta$, $T_n \xrightarrow{P_\theta} g(\theta)$ $(n \to \infty)$. seen ⇓ 2A

(ii) A sequence of estimators $(T_n)_{n\in\mathbb{N}}$ for $g(\theta)$ is called asymptotically unbiased if $E_\theta(T_n) \to \theta$ $(n \to \infty)$, $\forall \theta \in \Theta$. 2A

(b) (i) The log-likelihood follows as sim. seen ⇓

$$L(\lambda) = -n\lambda + s_n \log\lambda - \log\left(\prod_{i=1}^{n} x_i!\right)$$

where $s_n = \sum_{i=1}^n x_i$ (with corresponding random variable $S_n = \sum_{i=1}^n X_i$). So

$$\frac{\delta L(\lambda)}{\delta\lambda} = -n + \frac{s_n}{\lambda} \equiv 0$$

Therefore, $\widehat{\lambda} = \frac{s_n}{n} = \bar{x}$, which is a maximum likelihood since the second derivative is strictly negative. 4A

(ii) By the invariance of ML estimators to reparameterisation the maximum likelihood estimate of $\phi = \exp(-\lambda)$ is $\widehat{\phi} = \exp(-\bar{X})$. 2B

(iii) Using the central limit theorem, unseen ⇓

$$\frac{S_n - n\lambda}{\sqrt{n}} \to N(0, \lambda)$$

since $E(X) = \mathrm{Var}(X) = \lambda$ (*Note that students must remember/derive this themselves for a Poisson distribution*). So,

$$S_n \overset{\text{Asympt.}}{\sim} N(n\lambda, n\lambda)$$

And hence,

$$\bar{X} = \frac{S_n}{n} \overset{\text{Asympt.}}{\sim} N\left(\lambda, \frac{\lambda}{n}\right)$$

Note that the student could also approach this question using the general asymptotic result for the MLE and the Fisher Information, for which full marks could also be awarded. 4D

(c) (i) Let $\phi_1 = \mathrm{P}(X \text{ functions}) = \theta_1\theta_2$ and $\phi_2 = \mathrm{P}(Y \text{ functions}) = \theta_1\theta_2^2$. unseen ⇓
Let n_1 = number of type X which function $\sim$ Binomial(n, ϕ_1).
Let n_2 = number of type Y which function $\sim$ Binomial(n, ϕ_2).

Then we can write the likelihood as

$$L(\phi_1, \phi_2) = \binom{n}{n_1} \phi_1^{n_1}(1-\phi_1)^{n-n_1} \binom{n}{n_2} \phi_2^{n_2}(1-\phi_2)^{n-n_2}$$

Differentiating w.r.t. ϕ_1 and ϕ_2 (or using known results about MLEs from random samples from a binomial distribution), $\widehat{\phi}_i = \frac{n_i}{n}$, $i = 1, 2$. 2C

(ii) By functional invariance of MLEs we obtain $\widehat{\theta}_1 = \frac{n_1^2}{nn_2}$ and $\widehat{\theta}_2 = \frac{n_2}{n_1}$, since the transformation between the θ's and ϕ's is one-to-one. 3C

(iii) We note that the range for each unknown parameter θ_1, θ_2 is $[0, 1]$. If the conditions are not satisfied then the point estimate where the gradient is zero will not necessarily lie within the valid parameter space, i.e. the MLE would be on a point where the gradient is non-zero. 1C

4. (a) $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ seen ⇓

$\mathbf{Y}$ is an $n \times 1$ vector of observations.

$\mathbf{X}$ is an $n \times p$ design matrix.

$\boldsymbol{\beta}$ is an $p \times 1$ vector of parameters.

$\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random variables with zero mean describing the error. 2A

(b) *(Students have seen this in the lecture notes)* seen ⇓

Let $\mathrm{E}(\boldsymbol{Y}) = X\boldsymbol{\beta}$ be the linear model under consideration. As we are dealing with a full rank linear model, the least squares estimator is

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}.$$

Thus,

$$\begin{aligned}\operatorname{cov}(\widehat{\boldsymbol{\beta}}) =& (X^T X)^{-1} X^T \operatorname{cov}(\boldsymbol{Y})((X^T X)^{-1} X^T)^T \\ =& (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} \\ =& \sigma^2 (X^T X)^{-1}.\end{aligned}$$

5A

(c) (i) Let $c \in \mathbb{R}^p$ and let $\widehat{\boldsymbol{\beta}}$ be a least squares estimator of $\boldsymbol{\beta}$ in a linear model, where we assume full rank and second order assumptions. Then the estimator $c^T\widehat{\boldsymbol{\beta}}$ has the smallest variance among all linear unbiased estimators for $c^T\boldsymbol{\beta}$. 3A

(ii) This is an open question and any reasonable description of a linear model, where we are interested in estimating some linear combination of parameters, i.e. $c^T\boldsymbol{\beta}$, is acceptable. unseen ⇓ 3B

If we want an unbiased estimator for $c^T\boldsymbol{\beta}$, then the Gauss Markov theorem says that we should use $c^T\widehat{\boldsymbol{\beta}}$, as per part (i).

However, we may be able to find a biased estimator with lower variance, and hence lower MSE, in which case we might choose to ignore the Gauss Markov theorem. 2C

(d) This test statistic is t_{n-p}-distributed. sim. seen ⇓

We know that $c^T\widehat{\beta} \sim N(c^T\beta, c^T(X^TX)^{-1}c\sigma^2)$, and so we can define

$$A = \frac{c^T\widehat{\beta} - c^T\beta}{\sqrt{c^T(X^TX)^{-1}c\sigma^2}} \sim N(0,1)$$

Let $B = \frac{\mathrm{RSS}}{\sigma^2} \sim \chi^2_{n-p}$. We can use the property that A and B as just defined are independent, then use the fact that $\frac{A}{\sqrt{B/n}} \sim t_n$. 5D