TEMPORARY FRONT PAGE

BSc, MSc and MSci EXAMINATIONS (MATHEMATICS)

May 2018

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

Statistical Modelling I

Date: Friday, 18th May 2018

Time: 2 pm – 4 pm

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.

Statistical tables will not 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.
- Credit will be given for all questions attempted.
- Each question carries equal weight.
- Calculators may not be used.

1. Let X be the observed data and $f(x;\theta)$ be the probability density function (pdf) of a statistical model with scalar parameter θ. If T is an unbiased estimator of a function of the parameter, $\phi = \phi(\theta)$, then under regularity conditions (which in this question we assume hold) and for all θ,

$$\mathrm{Var}_\theta(T) \geq \frac{\left(\frac{d\phi}{d\theta}\right)^2}{I(\theta)}$$

where $I(\theta)$ is the expected Fisher information of the sample.

Let us consider a random sample of size n, such that the random variables $X_1, \ldots, X_n$ follow a Beta(θ,1) distribution, with some unknown parameter $\theta > 0$.

Recall: The pdf of a random variable $Z \sim Beta(\theta, 1)$ is $f(z) = \theta z^{\theta-1}$, for $0 < z < 1$, $\theta > 0$.

(a) (i) For a likelihood function L, give the definition of a maximum likelihood estimator (MLE).

(ii) Find the MLE of $\frac{1}{\theta}$.

(iii) Show that it is unbiased.

(iv) Check whether there exists any alternative unbiased estimator with lower variance.

(b) (i) By considering the expected value of a random variable with the Beta distribution, suggest an unbiased estimator for $\frac{\theta}{\theta+1}$.

(ii) Show whether the variance of your estimator attains the Cramer-Rao lower bound.

2. Answer the following questions. For multiple choice questions, read the following statements carefully and choose the single correct answer from the options provided.

(a) Describe how a pivotal quantity may be used to obtain a confidence interval.

(b) Consider a random sample of size n from a univariate Gaussian distribution with unknown mean, μ, and known variance, σ^2. The width of a typical, well-constructed, two-sided 95% confidence interval for μ

(i) would be larger than the width of a 90% confidence interval for μ.

(ii) would be smaller than the width of a 90% confidence interval for μ.

(iii) would be the same as the width of a 90% confidence interval for μ.

(iv) cannot be compared to the width of a 90% confidence interval for μ.

(v) None of the above.

(c) A radar gun is used to measure the speed of 74 randomly selected cars on the motorway, and the data is used to construct a 95% confidence interval for the mean speed of all the cars. The 95% confidence interval goes from 64.4 mph to 71.6 mph. Which of the following statements provides an appropriate interpretation of this 95% confidence interval?

(i) 95% of all cars on the motorway travel between 64.4 mph and 71.6mph.

(ii) There is a 95% probability that the mean speed of all cars on the motorway is between 64.4 mph and 71.6 mph.

(iii) There is a 95% probability that the mean speed of any group of 74 randomly chosen cars on the motorway is between 64.4 mph and 71.6 mph.

(iv) 95% of cars within any group of 74 randomly chosen cars on the motorway travel between 64.4 mph and 71.6mph.

(v) None of the above.

(d) Consider a random sample of size n from a univariate Gaussian distribution with unknown mean, μ, and known variance, σ^2. Which of the following would always result in an increase in the width of a typical, well-constructed, two-sided confidence interval?

(i) An increase in n.

(ii) A decrease in the sample standard deviation.

(iii) A decrease in the confidence level.

(iv) A decrease in the sample mean.

(v) None of the above would always cause an increase in the width of a confidence interval.

(e) Let X_1, X_2 denote a random sample from a uniform distribution $U(\theta - 0.5, \theta + 0.5)$, with unknown parameter $\theta \in \mathbb{R}$. Show that the random interval given by $[T_L, T_U]$, where $T_L = \min(X_1, X_2)$ and $T_U = \max(X_1, X_2)$, is a 50% confidence interval for θ.

(f) Consider a random sample $Y_1, ..., Y_n \sim N(\mu, \sigma^2)$, where μ and σ^2 are both unknown. Suggest an appropriate pivotal quantity for the mean parameter, μ, and derive an expression for a $(1 - \alpha)$ confidence interval. State the distribution of the pivotal quantity without proof.

3. Let $\mathbf{Y} \sim N(\mu, \mathbf{I})$ be a random vector, let A be a symmetric $n \times n$ matrix of rank $n-p$, and let B be a $p \times n$ matrix of rank p, such that $BA = 0$.

(a) Making use of any standard linear algebra results, show that there exists an $n \times (n-p)$ matrix L, such that $\text{rank}(L) = n-p$, $A = LL^T$, and L^TL is a diagonal matrix consisting of the non-zero eigenvalues of A.

(b) By considering the joint distribution of the random vector $\mathbf{Z} = [B\mathbf{Y}, L^T\mathbf{Y}]$, show that $\mathbf{Y}^TA\mathbf{Y}$ and $B\mathbf{Y}$ are independent.

(c) Consider the linear model $\boldsymbol{Y} = X\beta + \epsilon$, where $\boldsymbol{X}$ is an $n \times p$ design matrix of full rank p, $\beta \in \mathbb{R}^p$ is an unknown vector of regression coefficients and $\epsilon \sim N(0, \sigma^2 I_n)$. Recall that RSS is the residual sum of squares for the linear model.

(i) Derive the maximum likelihood estimator $\widehat{\beta}$.

(ii) Show that $\widehat{\sigma}^2 := \frac{\text{RSS}}{n-p}$ is an unbiased estimator of σ^2.

(iii) Show that $\widehat{\beta}$ and $\widehat{\sigma}^2$ are independent.

4. (a) Give the definition of the likelihood ratio test and describe the main idea behind its construction.

(b) Suppose that X and Y are continuous random variables representing the survival time for patients given two different treatments for a particular disease, such that their probability density functions take the form,

$$f_X(x;\alpha) = \alpha e^{-\alpha x} \quad (x > 0, \alpha > 0) \quad \text{and} \quad f_Y(y;\beta) = \beta e^{-\beta y} \quad (y > 0, \beta > 0)$$

Let $X_1, ..., X_n$ and $Y_1, ..., Y_n$ denote two random samples from these random variables.

(i) For the likelihood ratio test of $H_0 : \alpha = \beta$ versus $H_1 : \alpha \neq \beta$, show that the likelihood ratio statistic $\widehat{\lambda}$ can be written as

$$\widehat{\lambda} = [4u(1-u)]^n$$

where $u = \bar{x}/(\bar{x} + \bar{y})$, and where $\bar{x}$ and $\bar{y}$ represent the sample means of the random sample of X and Y respectively.

(ii) Give an asymptotic approximation for a distribution based on the test statistic $\widehat{\lambda}$ from part (i), and describe how you would decide whether to reject H_0, along with an interpretation of such a decision.

Corrections

Imperial College
London

IMPERIAL COLLEGE LONDON

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2018

This paper is also taken for the relevant examination for the Associateship.

M2S2

Statistical Modelling (Solutions)

Setter's signature	Checker's signature	Editor's signature
.....	.....	.....

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) A maximum likelihood estimator of a parameter θ is an estimator $\widehat{\theta}$ such that $L(\widehat{\theta}) = \sup_{\theta \in \Theta} L(\theta)$. 1A

(ii) The log likelihood from the n samples is sim. seen ⇓

$$l = \log L = \theta \sum_{i=1}^{n} \log(X_i) + n \log \theta - \sum_{i=1}^{n} \log(X_i)$$

Take derivative with respect to θ and set equal to zero to obtain $\sum_{i=1}^{n} \log(X_i) + \frac{n}{\theta} = 0$. Therefore MLE of θ is $-n/\sum_{i=1}^{n} \log(X_i)$, since the second derivative is negative, and the MLE of $1/\theta$ is $-\sum_{i=1}^{n} \log(X_i)/n$. 3A

(iii) To show unbiasedness we note that,

$$\begin{aligned} E\left(-\frac{\sum_{i=1}^{n} \log(X_i)}{n}\right) &= -E(\log(X_1)) \\ &= -\int_0^1 (\log(x))\theta x^{\theta-1} dx \\ &= -(\log x)x^{\theta}|_0^1 + \int_0^1 x^{\theta} \frac{1}{x} dx \\ &= \int_0^1 x^{\theta-1} dx \\ &= \frac{1}{\theta} \end{aligned}$$

where we note the calculation for $-(\log x)x^{\theta}|_0^1$ involves basic use of l'Hopital's rule to establish that $\lim(\log(x) \times x^{\theta}) \to 0$, as $x \to 0$. Alternatively, the student may make the argument that since $\log(x)$ increases more slowly than any positive power of x, so x^{θ} tends to zero more rapidly than $1/\log(x)$. One mark should be deducted if either of these arguments is not given explicitly. 3C

(iv) We compare the variance of this estimator to the Cramer-Rao lower bound, since we are considering an unbiased estimator. The variance follows as,

$$\begin{aligned} \mathrm{Var}\left(-\frac{\sum_{i=1}^{n} \log(X_i)}{n}\right) &= \frac{\mathrm{Var}(\log X_1)}{n} \\ &= \frac{1}{n}\left(E\left(\log(X_1)^2 - \frac{1}{\theta^2}\right)\right) \\ &= \frac{1}{n}\left(E\left(\log(X_1)^2\right) - \frac{1}{\theta^2}\right) \\ &= \frac{1}{n}\left(\frac{2}{\theta^2} - \frac{1}{\theta^2}\right) \\ &= \frac{1}{n\theta^2} \end{aligned}$$

where,

$$\begin{aligned} E\left((\log X_1)^2\right) &= \int_0^1 (\log x)^2 \theta x^{\theta-1} dx \\ &= (\log x)^2 x^\theta|_0^1 - \int_0^1 \frac{2}{x}(\log x) x^\theta dx \\ &= -\frac{2}{\theta}\int_0^1 (\log x)\theta x^{\theta-1} dx = \frac{2}{\theta^2} \end{aligned}$$

Note that the first term is zero via straightforward double application of l'Hopital's rule, while for the second term we can simply use the result from part (iii).

3C

The single sample expected Fisher information follows as,

$$I_1(\theta) = -E\left(\frac{\delta^2}{\delta\theta^2}\log f(x,\theta)\right) = \frac{1}{\theta^2}$$

2B

And so the Cramer-Rao lower bound follows as,

$$\frac{\left(\frac{d}{d\theta}\left(\frac{1}{\theta}\right)\right)^2}{nI_1(\theta)} = \frac{\frac{1}{\theta^4}}{n\frac{1}{\theta^2}} = \frac{1}{n\theta^2}$$

The proposed estimator therefore achieves the lower bound and there is no alternative unbiased estimator with lower variance.

2B

(b) (i) Note that the expected value follows as, unseen ⇓

$$E(X) = \int_0^1 x\theta x^{\theta-1}dx = \int_0^1 \theta x^{\theta}dx = \frac{\theta}{\theta+1}x^{\theta+1}|_0^1 = \frac{\theta}{\theta+1}$$

Therefore, $\bar{X}$ is an unbiased estimator of $\frac{\theta}{\theta+1}$. 3D

(ii) The variance of this estimator is,

$$\frac{\text{Var}(X)}{n} = \frac{1}{n}\left(E(X^2) - \left(\frac{\theta}{\theta+1}\right)^2\right) = \frac{1}{n}\left(\frac{\theta}{\theta+2} - \frac{\theta^2}{(\theta+1)^2}\right) = \frac{\theta}{n(\theta+2)(\theta+1)^2}$$

The lower bound is

$$\frac{\left(\frac{d}{d\theta}\left(\frac{\theta}{\theta+1}\right)\right)^2}{nI_1(\theta)} = \frac{\left(\frac{1}{(\theta+1)^2}\right)^2}{n(\frac{1}{\theta^2})} = \frac{\theta^2}{n(\theta+1)^4}$$

It is easy to show that

$$\frac{\theta}{(\theta+1)^2} < \frac{1}{(\theta+2)}$$

since by considering the denominators, $\frac{\theta^2}{\theta} + \frac{2\theta}{\theta} + \frac{1}{\theta} > \theta + 2$ for all positive θ, and therefore the estimator has a larger variance than the lower bound,

$$\frac{\theta^2}{n(\theta+1)^4} < \frac{\theta}{n(\theta+2)(\theta+1)^2} = \text{Var}(\bar{X})$$

3D

2. (a) Suppose $t(\mathbf{Y}, \theta)$ is a pivotal quantity for θ. Then we can find constants a_1 and a_2 such that,

seen ⇓

$$P(a_1 \leq t(\mathbf{Y}, \theta) \leq a_2) \geq 1 - \alpha$$

since we know the distribution of $t(\mathbf{Y}, \theta)$. In many case we can rearrange the terms to obtain the form,

4A

$$P(h_1(\mathbf{Y}) \leq \theta \leq h_2(\mathbf{Y})) \geq 1 - \alpha$$

Then $[h_1(\mathbf{Y}), h_2(\mathbf{Y})]$ is a random interval and the observed interval $[h_1(\mathbf{y}), h_2(\mathbf{y})]$ is a $(1 - \alpha)$ confidence interval for θ.

sim. seen ⇓

(b) (i) would be larger than the width of a 90% confidence interval on μ.

2B

(c) (v) none of the above.

2B

(d) (v) none of the above would cause an increase in the width of a confidence interval.

2B

(e) We can show that the confidence interval has 50% coverage probability by considering,

unseen ⇓

$$\begin{aligned} P(\min(X_1, X_2) < \theta < \max(X_1, X_2)) &= P(X_1 < \theta < X_2) + P(X_2 < \theta < X_1) \\ &= P(X_1 < \theta)P(X_2 > \theta) + P(X_2 < \theta)P(X_1 > \theta) \\ &= 0.5 \times 0.5 + 0.5 \times 0.5 = 0.5 \end{aligned}$$

4D

(f) We can use $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ as a pivotal quantity as long as we use the sample standard deviation for σ, since this nuisance parameter is also unknown. We can therefore substitute the sample variance for σ^2,

seen ⇓

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{\mathbf{Y}})^2$$

which results in the statistic $T = \frac{\sqrt{n}}{S}(\bar{\mathbf{Y}} - \mu) = \frac{(\bar{\mathbf{Y}}-\mu)}{\sqrt{S^2/n}}$, which follows a Student-t distribution with $n - 1$ degrees of freedom.

We then obtain a $(1 - \alpha)$ confidence interval $(\bar{\mathbf{y}} - \frac{S}{\sqrt{n}}t_{\alpha/2}, \bar{\mathbf{y}} + \frac{S}{\sqrt{n}}t_{\alpha/2})$, by considering

$$\begin{aligned} 1 - \alpha &= P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) \\ &= P(\bar{\mathbf{Y}} - \frac{S}{\sqrt{n}}t_{\alpha/2} \leq \mu \leq \bar{\mathbf{Y}} + \frac{S}{\sqrt{n}}t_{\alpha/2}) \end{aligned}$$

6A

3. (a) Using a standard result from linear algebra, $\exists$ an orthogonal matrix P such that

seen ⇓

$$P^T A P = D = \text{diag}(\text{eigenvalues of } A)$$

Can assume positive definiteness

Precisely $n - p$ elements of D are positive and the others are zero.
Hence, $A = PDP^T = PD^{\frac{1}{2}} D^{\frac{1}{2}} P^T = PD^{\frac{1}{2}}(PD^{\frac{1}{2}})^T$. Let L consist of the nonzero columns of $PD^{\frac{1}{2}}$. Then $A = LL^T$, and because P is orthogonal (i.e. $P^T P = I$) we also get $L^T L = \text{diag}(\text{nonzero eigenvalues of } A)$.

5A

(b) Since the components of Z are defined in terms of linear combinations of the elements of a multivariate normal random vector, the Z will also be jointly multivariate normal. We can calculate the covariance of the elements $B\mathbf{Y}$ and $L^T\mathbf{Y}$ as follows.
From part (a) we know that there exists some $L \in \mathbb{R}^{n\times p}$ such that $A = LL^T$. Then, using the fact that $BA = 0$,

$$\text{cov}(B\mathbf{Y}, L^T\mathbf{Y}) = B\,\text{cov}(\mathbf{Y})L = BL = BL(L^TL(L^TL)^{-1}) = BAL(L^TL)^{-1} = 0$$

Since $B\mathbf{Y}$ and $L^T\mathbf{Y}$ are jointly normal and have zero covariance, they are therefore independent. Hence $B\mathbf{Y}$ and $\mathbf{Y}LL^T\mathbf{Y} = \mathbf{Y}A\mathbf{Y}$ are also independent, using the fact that independent functions of independent random variables retain their independence.

4A

(c) (i) Consider the log likelihood $L(\boldsymbol{\beta})$.

$$L(\boldsymbol{\beta}) \propto (\boldsymbol{Y} - X\boldsymbol{\beta})^T(\boldsymbol{Y} - X\boldsymbol{\beta}) = \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^TX\boldsymbol{\beta} + \boldsymbol{\beta}^TX^TX\boldsymbol{\beta}$$

The derivative then follows as,

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2X^T\boldsymbol{Y} + 2X^TX\boldsymbol{\beta}$$

Setting equal to zero and rearranging we then obtain, $X^TX\widehat{\boldsymbol{\beta}} = X^T\boldsymbol{Y}$, from which we get the final solution, assuming that X^TX is invertible.

3A

(ii) Consider the projection matrix P, which projects onto the space spanned by the columns of the design matrix X, and the projection matrix $Q = I - P$, which projects onto the complement of that space. We start by using the expression, $\text{RSS} = Y^TQY$.

$$\begin{aligned}
\text{E(RSS)} &= \text{E}\,\text{trace}\,\text{RSS} = \text{E}\,\text{trace}(\boldsymbol{Y}^TQ\boldsymbol{Y}) = \text{E}\,\text{trace}(Q\boldsymbol{Y}\boldsymbol{Y}^T) = \text{trace}(Q\,\text{E}(\boldsymbol{Y}\boldsymbol{Y}^T)) \\
&= \text{trace}(Q[\text{cov}\,\boldsymbol{Y} + \text{E}(\boldsymbol{Y})\,\text{E}(\boldsymbol{Y})^T]) = \text{trace}(Q\sigma^2) + \text{trace}(QX\boldsymbol{\beta}(X\boldsymbol{\beta})^T) \\
&= \sigma^2\,\text{trace}(I - P) + 0 \\
&= \sigma^2(n - \text{trace}(P)) \\
&= \sigma^2(n - \text{rank}(P)) \\
&= \sigma^2(n - r).
\end{aligned}$$

Rearranging these expressions gives us the required result.

5B

(iii) From part (i) let us consider $B = (X^T X)^{-1} X^T$ and from part (ii), $A = Q$. Then B is a $p \times n$ matrix of rank p, A is a $n \times n$ matrix of rank $n - p$, and since Q projects onto the complement of the space spanned by the columns of X, then $BA = 0$, and so using part (b) we can conclude that $\widehat{\beta}$ and $\widehat{\sigma}^2$ must be independent.

sim. seen ⇓

3B

4. (a) Suppose we observe the data $\mathbf{y}$, then the likelihood ratio test compares the null hypothesis $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1 := \Theta/\Theta_0$ using the statistic

seen ⇓

$$t(\mathbf{y}) = \frac{\sup_{\theta\in\Theta} L(\theta;\mathbf{y})}{\sup_{\theta\in\Theta_0} L(\theta;\mathbf{y})} = \frac{\text{Max. likelihood under } H_0 \text{ and } H_1}{\text{Max. likelihood under } H_0}$$

3A

The main idea is therefore to compare the maximised likelihood L under H_0, where the parameter space is constrained, to the unrestricted likelihood over a larger set of possible values for the parameter. If the maximised likelihood under the larger set of possible parameter values is much larger than that under the constrained set of values, then $t(\mathbf{y})$ will be large and indicate we should reject the null hypothesis of the constrained set of parameter values.

3A

(b) (i) Under $H_0 : \alpha = \beta (= \gamma, \text{say})$, the restricted likelihood is $L_0 = \gamma^{2n} e^{-n\gamma(\bar{x}+\bar{y})}$.

unseen ⇓

So,

$$\frac{\delta \log(L_0)}{\delta\gamma} = \frac{2n}{\gamma} - n(\bar{x}+\bar{y}) = 0 \qquad \text{gives} \qquad \widehat{\gamma}_0 = 2(\bar{x}+\bar{y})^{-1}$$

Therefore,

$$\widehat{L}_0 = \widehat{\gamma}_0^{2n} e^{-n\widehat{\gamma}_0(\bar{x}+\bar{y})} = \left[\frac{2}{(\bar{x}+\bar{y})}\right]^{2n} e^{-2n}$$

Under $H_1 : \alpha \neq \beta$, the unrestricted likelihood is $L_1 = \alpha^n e^{-n\alpha\bar{x}} \beta^n e^{-n\beta\bar{y}}$. So that,

$$\left.\begin{aligned} \frac{\delta \log(L_1)}{\delta\alpha} &= \frac{n}{\alpha} - n(\bar{x}) = 0 \qquad \text{gives} \qquad \widehat{\alpha}_1 = (\bar{x})^{-1}, \\ \text{and} \\ \frac{\delta \log(L_1)}{\delta\beta} &= \frac{n}{\beta} - n(\bar{y}) = 0 \qquad \text{gives} \qquad \widehat{\beta}_1 = (\bar{y})^{-1}. \end{aligned}\right\}$$

Therefore,

$$\widehat{L}_1 = (\bar{x}^{-1})^n e^{-n\bar{x}^{-1}\bar{x}} (\bar{y}^{-1})^n e^{-n\bar{y}^{-1}\bar{y}} = (\bar{x}\bar{y})^{-n} e^{-2n}$$

Finally the likelihood ratio statistic $\widehat{\lambda}$ can be written as

$$\widehat{\lambda} = \frac{\widehat{L}_1}{\widehat{L}_0} = \left[\frac{4\bar{x}\bar{y}}{(\bar{x}+\bar{y})^2}\right]^{n} = [4u(1-u)]^{n}, \quad \text{with } u = \frac{\bar{x}}{(\bar{x}+\bar{y})}$$

(ii) Since H_0 is "nested" in H_1, i.e. Θ_0 is a subset of Θ, we can use the following asymptotic result to construct a test,

8D

sim. seen ⇓

$$2\log(t(\mathbf{y})) \xrightarrow{d} \chi^2_r \quad (\text{as } n \to \infty)$$

under H_0, where r is the number of independent restrictions on θ needed to define H_0, i.e. the number of independent parameters under the full model minus the number of independent parameters under H_0.

In this particular case, $r = 1$. We could choose a particular cut off point, such that the probability of the test statistic being greater than that point is less than some chosen probability, 0.05 say. Given the data we can evaluate the test statistic and reject the null hypothesis for values higher than the calculated cut off point.
Since $E(X) = \frac{1}{\alpha}$, $E(Y) = \frac{1}{\beta}$, rejection of the null hypothesis implies that the two treatments result in different average survival times for patients with the disease.