Imperial College
London

MATH97073

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May-June 2020

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

Statistical Theory 1

Date: 18th May 2020

Time: 09.00am - 11.30am (BST)

Time Allowed: 2 Hours 30 Minutes

Upload Time Allowed: 30 Minutes

This paper has 5 Questions.

Candidates should start their solutions to each question on a new sheet of paper.

Each sheet of paper should have your CID, Question Number and Page Number on the top.

Only use 1 side of the paper.

Allow margins for marking.

Any required additional material(s) will be provided.

Credit will be given for all questions attempted.

Each question carries equal weight.

SUBMIT YOUR ANSWERS AS ONE PDF TO THE RELEVANT DROPBOX ON BLACKBOARD INCLUDING A COMPLETED COVERSHEET WITH YOUR CID NUMBER, QUESTION NUMBERS ANSWERED AND PAGE NUMBERS PER QUESTION.

.

1. (a) Let $f(x, y; \theta_1, \theta_2, \theta_3, \theta_4)$ be the bivariate probability distribution function (pdf) for the uniform distribution on the rectangle with lower left corner (θ_1, θ_2) and upper right corner (θ_3, θ_4), and with $\theta_1, \theta_2, \theta_3, \theta_4 \in \mathbb{R}$ satisfying $\theta_1 < \theta_3$ and $\theta_2 < \theta_4$.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from this pdf. Find a four-dimensional sufficient statistic for $(\theta_1, \theta_2, \theta_3, \theta_4)$, stating all results you use from the lectures. (4 marks)

(b) Let $X_1, \ldots, X_n$ be samples drawn from a $Gamma(\alpha, \beta)$ population, where both $\alpha > 0$ and $\beta > 0$ are unknown parameters. Find a minimal sufficient statistic for (α, β), stating all results you use from the lectures. You may use the following pdf for the Gamma distribution:

$$f_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad x \in (0, \infty)$$

(4 marks)

(c) Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ has joint density $f(\boldsymbol{X}; \theta_1, \theta_2)$, and with the support of f independent of both θ_1 and θ_2.

Suppose that T_1 is sufficient for θ_1 when θ_2 is known, and that T_2 is sufficient for θ_2 when θ_1 is known. Show that (T_1, T_2) is sufficient for (θ_1, θ_2) if T_1 does not depend on θ_2, and T_2 does not depend on θ_1.

(12 marks)

(Total: 20 marks)

2. (a) Show that maximum likelihood estimators are equivariant. (5 marks)

(b) Let $X_1, \ldots, X_n$ be a random sample drawn from a Bern(θ) population, and let:

$$g(\theta) = \frac{2\theta^2 - 5}{\theta + 2}$$

Find the asymptotic distribution of $g(\overline{X}_n)$ (where $\overline{X}_n$ is the sample mean), citing any results you use from the lectures.

(7 marks)

(c) Let $X_1, \ldots, X_n$ be a random sample from a population having an unknown parameter $\theta \in \mathbb{R}$ with pdf:

$$f(x; \theta) = \frac{1}{2\theta} e^{\frac{-|x|}{\theta}} \qquad x \in \mathbb{R} \qquad \theta > 0$$

Find the maximum likelihood estimator of θ and use its Fisher information to compute its standard error. (8 marks)

(Total: 20 marks)

3. Suppose that $X_1, \ldots, X_n$ are n independent identically distributed observations from $U(\theta - a, \theta + a)$ where a is a known positive constant, and further assume θ has as prior distribution the Exponential distribution with mean 1.

(a) Obtain the posterior distribution of θ.

(3 marks)

(b) Is the prior distribution a conjugate prior? Justify your answer.

(3 marks)

(c) Find the Bayesian point estimator of θ under the squared error loss function.

(5 marks)

(d) Find the Bayesian point estimator of θ under the absolute error loss function.

(5 marks)

(e) Is the Bayes estimator obtained in part d) admissible or not?

(4 marks)

(Total: 20 marks)

4. (a) For each of the following concepts briefly outline (in no more than two paragraphs for each concept) a practical application of the concept:

(i) Sufficiency. (4 marks)

(ii) Bayesian estimators. (4 marks)

(b) Verify that the inverse Gamma distribution with parameters α and β is a member of the k-parameter exponential family. You may use the following form of its probability density function:

$$f(x;\ \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\Big\{-\frac{\beta}{x}\Big\} \qquad x \in (0, \infty)$$

(4 marks)

(c) Let $X_1, \ldots, X_n$ be a random sample from a Uniform$(\theta_1 - \theta_2, \theta_1 + \theta_2)$ population with unknown $\theta = (\theta_1, \theta_2)$, and with $\theta_1 \in \mathbb{R}$ and $\theta_2 \in \mathbb{R}^+$. Find a minimal sufficient statistic for θ, stating all results you use to derive it.

(8 marks)

(Total: 20 marks)

5. (a) We have a standard six-sided die. Let X be the number of times that a 6 occurs over n throws of the die. Let p be the probability of the event $X \geq n/4$. Find the upper bounds on p as a function of n that you can obtain using:

(i) Markov's inequality, (3 marks)

(ii) Chebyshev's inequality, (4 marks)

(iii) Chernoff bound (you may use any version of it). (6 marks)

The above calculations should suggest that Chernoff bounds give the tightest bounds for this tail event compared to Markov's and Chebyshev's inequality.

(iv) State one potential limitation of using Chernoff bounds to estimate tail probabilities of random variables which makes it sometimes preferable to use Markov's or Chebyshev's inequality instead.

(3 marks)

(b) Briefly give an example from statistical inference of an application that uses large deviation bounds on the sum of independent Bernoulli random variables.

(4 marks)

(Total: 20 marks)

Statistical Theory I: 2019-2020 Exam Solutions

Q1

a)
Cat A, seen similar

$$f(x_1, y_1, \ldots, x_n, y_n; \theta_1, \theta_2, \theta_3, \theta_4) = (\theta_3-\theta_1)^{-n}(\theta_4-\theta_2)^{-n}(\mathbb{I}_{[\theta_1,\infty)}x_{(1)})(\mathbb{I}_{(-\infty,\theta_2]}x_{(n)})(\mathbb{I}_{[\theta_3,\infty)}y_{(1)})(\mathbb{I}_{(-\infty,\theta_4]}y_{(n)})$$

Hence using the Neyman factorisation theorem, the statistic

$$T(x_1, y_1, \ldots, x_n, y_n) = (x_{(1)}, x_{(n)}, y_{(1)}, y_{(n)})$$

is sufficient for $(\theta_1, \theta_2, \theta_3, \theta_4)$.

b)
Cat B, seen similar
Consider the ratio:

$$\frac{f(x_1, \ldots, x_n; \alpha, \beta)}{f(y_1, \ldots, y_n; \alpha, \beta)}$$

After some simplification the ratio can be written as:

$$(\prod_i^n \frac{x_i}{y_i})^{\alpha-1}\Big(\exp\Big\{\beta\Big(\sum^n y_i - \sum^n x_i\Big)\Big\}\Big)$$

This ratio is independent of α and β if and only if $\prod_{i=1}^n x_i = \sum_{i=1}^n y_i$ and $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. Hence by the characterisation theorem of minimal sufficient statistics proved in the module, the two dimensional statistic $T(x_1, \ldots, x_n) = (\prod_{i=1}^n x_i, \sum_{i=1}^n x_i)$ is a minimal sufficient statistics for (α, β).

c)

Cat C+D, unseen

From the facts given in the question along with the Neyman factorisation theorem, it follows that there exist functions g_1, g_2, h_1 and h_2 such that the density function f can be written in the following ways:

$$f(x;\theta_1,\theta_2) = g_1(T_1,\theta_1,\theta_2)h_1(x,\theta_2)$$
$$f(x;\theta_1,\theta_2) = g_2(T_2,\theta_1,\theta_2)h_2(x,\theta_1)$$

where the first line follows from the sufficiency of T_1 for θ_1 when θ_2 is known, and the second line follows vice versa.

It follows that we can express $h_1(x,\theta_2)$ as

$$h_1(x,\theta_2) = \frac{g_2(T_2,\theta_1,\theta_2)h_2(x,\theta_1)}{g_1(T_1,\theta_1,\theta_2)}$$

We can set θ_1 on the RHS of the above relationship to an arbitrary value c from the parameter space of θ_1, to get:

$$h_1(x,\theta_2) = \frac{g_2(T_2,c,\theta_2)h_2(x,c)}{g_1(T_1,c,\theta_2)}$$

Hence we can write $h_1(x,\theta_2)$ as a product of two functions $G(T_1,T_2,\theta_2)$ and $H(x)$, defined as follows:

$$G(T_1,T_2,\theta_2) = \frac{g_2(T_2,c,\theta_2)}{g_1(T_1,c,\theta_2)} \quad \text{and } H(x) = h_2(x,c)$$

Returning to our original expression:

$$f(x;\theta_1,\theta_2) = g_1(T_1,\theta_1,\theta_2)h_1(x,\theta_2)$$

And rewriting h_1 in terms of G, and H to get:

$$f(x;\theta_1,\theta_2) = g_1(T_1,\theta_1,\theta_2)G(T_1,T_2,\theta_2)H(x)$$

Now set $G'(T_1,T_2,\theta_1,\theta_2) = g_1(T_1,\theta_1,\theta_2)G(T_1,T_2,\theta_2)$, so that

$$f(x;\theta_1,\theta_2) = G'(T_1,T_2,\theta_1,\theta_2)H(x)$$

And so it follows from the Neyman factorisation Theorem that (T_1, T_2) is sufficient for (θ_1, θ_2).

Alternative solution (provided by external examiner):

The conditional distribution

$X \mid T_1(X)$ does not depend on θ_1 (but may depend on θ_2),

Hence (with a little thought), the conditional distribution

$X \mid T_1(X),\ T_2(X)$ does not depend on θ_1 (but may depend on θ_2). (1)

Similarly, the conditional distribution

$X \mid T_2(X)$ does not depend on θ_2 (but may depend on θ_1),

Hence, the conditional distribution

$X \mid T_1(X),\ T_2(X)$ does not depend on θ_2 (but may depend on θ_1). (2)

Then (1) and (2) together imply that the conditional distribution

$X \mid T_1(X),\ T_2(X)$ does not depend on θ_1 or θ_2,

as required.

1 Q2

a)

Cat A, seen

Equivariance: Let $\tau = g(\theta)$ be a function of θ, Let $\widehat{\theta}$ be the MLE of θ. Then $\widehat{\tau} = g(\widehat{\theta})$ is the MLE of τ.

Proof: Let $h = g^{-1}$ denote the inverse of g. Then $\hat{\theta} = h(\hat{\tau})$. For any τ, $\mathcal{L}(\tau) = \prod_i f(x_i; h(\tau)) = \prod_i f(x_i; \theta) = \mathcal{L}(\theta)$ where $\theta = h(\tau)$. Hence, for any τ, $\mathcal{L}(\tau) = \mathcal{L}(\theta) \leq \mathcal{L}(\hat{\theta}) = \mathcal{L}(\hat{\tau})$.

b)

Cat B, seen similar

We have from the CLT that:

$$\sqrt{n}(\overline{X}_n - \theta) \xrightarrow{D} N(0, \theta(1-\theta))$$

Since $g'(\theta) = \frac{2\theta^2+8\theta+5}{(\theta+2)^2} \neq 0$ for $\theta \in [0,1]$, then it follows from the Delta method that

$$\sqrt{n}(g(\overline{X}_n) - g(\theta)) \xrightarrow{D} N\left(0, \left(\frac{2\theta^2+8\theta+5}{(\theta+2)^2}\right)^2 \theta(1-\theta)\right)$$

Which shows that the estimator is asymptotically normal.

c)

Cat C, seen similar

We have that the derivative of the log-likelihood is equal to $\frac{-n}{\theta} - \sum^{n} \frac{|x_i|}{\theta}$.

So by taking derivative and equating to zero, we find that the MLE of θ is:

$$\widehat{\theta}_{MLE} = \frac{\sum_i^n |X_i|}{n}$$

We know from the lecture that MLE estimators are asymptotically normal with standard error given by $se \approx \sqrt{1/I_n(\theta)}$, where I_n is the Fisher information. Moreover we know from the lectures that $I_n(\theta) = nI(\theta)$, and $I(\theta)$ can be computed using:

$$I(\theta) = -\mathbb{E}_\theta \left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2} \right)$$

Applying this to $\widehat{\theta}_{MLE}$ we find that:

$$\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2} = -\frac{2|X|}{\theta^3} + \frac{1}{\theta^2}$$

$$I(\theta) = -\mathbb{E}_\theta \left(-\frac{2|X|}{\theta^3} + \frac{1}{\theta^2} \right) = \frac{2}{\theta^3}\mathbb{E}_\theta(|X|) - \frac{1}{\theta^2}$$

So we need to compute $\mathbb{E}_\theta(\ |X|\ )$:

$$\mathbb{E}_\theta(\ |X|\ ) = \int_{-\infty}^{\infty} |x| f_\theta(x) dx = \int_{-\infty}^{\infty} \frac{|x|}{2\theta} \exp\left\{ \frac{-|x|}{\theta} \right\} dx = 2 \int_{0}^{\infty} \frac{x}{2\theta} \exp\left\{ \frac{-x}{\theta} \right\} dx = \theta$$

And therefore

$$I_n(\theta) = nI(\theta) = n\Big(\frac{2}{\theta^3}\mathbb{E}_\theta(|X|) - \frac{1}{\theta^2} \Big) = n\Big(\frac{2\theta}{\theta^3} - \frac{1}{\theta^2} \Big) = \frac{n}{\theta^2}$$

Hence, $\widehat{\theta}_{MLE}$ is asymptotically normal with $se \approx \sqrt{1/I_n(\theta)} = \sqrt{\frac{\theta^2}{n}}$.

2 Q3

a)
Cat A, routine
First note that we must have $\theta - a \leq x_{(1)} \leq x_{(n)} \leq \theta + a$ which gives us $x_{(n)} - a \leq \theta \leq x_{(1)} + a$.
Now, the posterior is given by

$$\begin{aligned}
\pi(\theta|x_1, ..., x_n) &= \frac{f_\theta(x_1, ..., x_n)\pi(\theta)}{\int f_\theta(x_1, ..., x_n)\pi(\theta)d\theta} = \frac{(2a)^{-n}e^{-\theta}}{\int\limits_{x_{(n)}-a}^{x_{(1)}+a} (2a)^{-n}e^{-\theta}d\theta} \\
&= \frac{e^{-\theta}}{\int\limits_{x_{(n)}-a}^{x_{(1)}+a} e^{-\theta}d\theta} = \frac{e^{-\theta}}{e^{a-x_{(n)}} - e^{-a-x_{(1)}}}, \qquad x_{(n)} - a \leq \theta \leq x_{(1)} + a.
\end{aligned}$$

b)
Cat A, unseen
No, because the posterior distribution is not an Exponential distribution. In fact, the support of θ in the posterior depends on the data (the posterior is a truncated Exponential distribution).

c) *Cat B, unseen*
Under the squared error loss, we know that the Bayes estimator is the posterior

mean, which can be directly obtained as follows:

$$\begin{aligned}
\hat{\theta}_{\text{Bayes}} &= E(\theta|x_1, ..., x_n) = \int\limits_{x_{(n)}-a}^{x_{(1)}+a} \frac{\theta e^{-\theta}}{e^{a-x_{(n)}} - e^{-a-x_{(1)}}} d\theta \\
&= \frac{1}{e^{a-x_{(n)}} - e^{-a-x_{(1)}}} \int\limits_{x_{(n)}-a}^{x_{(1)}+a} \theta e^{-\theta} d\theta \\
&= \frac{1}{e^{a-x_{(n)}} - e^{-a-x_{(1)}}} \Big[-\theta e^{-\theta} + \int e^{-\theta} d\theta \Big]_{x_{(n)}-a}^{x_{(1)}+a} \\
&= \frac{1}{e^{a-x_{(n)}} - e^{-a-x_{(1)}}} \Big[-(\theta+1) e^{-\theta} \Big]_{x_{(n)}-a}^{x_{(1)}+a} \\
&= \frac{\big(x_{(n)} - a + 1\big) e^{a-x_{(n)}} - \big(x_{(1)} + a + 1\big) e^{-a-x_{(1)}}}{e^{a-x_{(n)}} - e^{-a-x_{(1)}}}.
\end{aligned}$$

d) *Cat A, unseen*
Under absolute error loss, we know that the Bayes estimator is the posterior median, which can be obtained by solving for $\bar{\theta}$:

$$\int\limits_{x_{(n)}-a}^{\bar{\theta}} \frac{e^{-\theta}}{e^{a-x_{(n)}} - e^{-a-x_{(1)}}} d\theta = \frac{1}{2}$$

This is a straightforward integration, which simplifies to:

$$\Big[-e^{-\theta} \Big]_{x_{(n)}-a}^{\bar{\theta}} = \frac{e^{a-x_{(n)}} - e^{-a-x_{(1)}}}{2}$$

Solving for $\bar{\theta}$ yields:

$$\bar{\theta} = -\ln \Big(\frac{e^{a-x_{(n)}} + e^{-a-x_{(1)}}}{2} \Big)$$

e) *CatD, routine*
Yes, the Bayes estimator obtained in part d is admissible because it is unique (this is a result proved in the module).

3 Q4

a)

i) Sufficiency

Several applications were discussed in the module: sufficiency can be used to summarise data by providing a quantitative measure of the effectiveness of the summary, this can help adhere to privacy regulations or to save storage space when data is stored on servers for analytical purposes. In the module it was also proved proved that sufficient estimators tend to have desirable asymptotic qualities, so sufficiency can help in generating good estimators (for example via Rao-Blackwell theorem).

ii) Bayesian estimators

Bayesian estimators are useful when working in applications where there is some prior knowledge on the values that a particular parameter can take. This allows the statistical model to make full use of all knowledge available about the population which leads to better estimates.

Bayesian estimators can also be utilised to encode different loss functions, which can be useful in applications where the loss function is not the traditional squared error loss (for example if the loss function is 0-1 which does occur in many real-life applications such as survival analysis).

b)

The density function of the inverse Gamma function can be written in the following form:

$$f(x;\theta) = h(x)\exp\{\eta(\theta)T(x) - A(\theta)\}$$

Where $h(x) = 1$, $\eta(\alpha, \beta) = \big(-\beta, -\alpha{-}1\big)$, $T(x) = \big(\frac{1}{x},\ \log(x)\big)$ and $A(\theta) = \log(\frac{\beta^\alpha}{\Gamma(\alpha)})$. This verifies the distribution as being a member of the 2-parameter exponential family.

c)

A minimal sufficient statistic for θ can be derived via the following steps:

1. Using the factorisation theorem find a sufficient statistic S for θ.

2. Find the MLE of θ and show that it is unique.

3. Show that the MLE of θ (from step 2) is also a function of S (from step 1), and is therefore a sufficient statistic.

4. Use the result proved in the module which states that if the MLE estimator of a parameter is both unique and sufficient then is is also minimally sufficient.

For the the first step, note that the likelihood function of the joint distribution of $X_1, \ldots, X_n$ is:

$$f(x_1, \ldots, x_n; \theta_1, \theta_2) = (2\theta_2)^{-n} \quad \mathbb{I}_{(-\infty, x_{(1)}]}(\theta_1 - \theta_2) \quad \mathbb{I}_{[x_{(n)}, \infty)}(\theta_1 + \theta_2)$$

Hence it follows by the factorisation theorem that the two-dimensional statistic $S(X_1, \ldots, X_n) = (X_{(1)}, X_{(n)})$ is a sufficient statistic for θ.

Moreover it follows from inspecting the likelihood function that it is uniquely maximised when $\widehat{\theta_1} = (X_{(1)} + X_{(n)})/2$, and $\widehat{\theta_2} = (X_{(n)} - X_{(1)})/2$, establishing $\widehat{\theta} = (\widehat{\theta_1}, \widehat{\theta_2})$ as the unique MLE of (θ_1, θ_2). This accomplishes the second step.

Now it follows that $\widehat{\theta}$ is also a sufficient statistic, since it is a one-to-one function of the sufficient statistic S, and we have proved in the lecture that one-to-one mappings of sufficient statistics are also sufficient. This completes the third step.

Hence we have shown that $\theta = (\theta_1, \theta_2)$ has a unique MLE $\widehat{\theta} = (\widehat{\theta_1}, \widehat{\theta_2})$ which is also a sufficient statistic, and therefore it follows from a result we proved in the module that $\widehat{\theta}$ is a minimal sufficient statistic of θ.

4 Q5

a)

Let X_n be the random variable for the number of 6's after n dice rolls. Thus, we know that $X_n \sim Binomial(n, \frac{1}{6})$.

a. Markov: $P(X_n > \frac{n}{4}) < \frac{E[X_n]}{\frac{n}{4}} = \frac{\frac{n}{6}}{\frac{n}{4}} = \boxed{\frac{2}{3}}$

b. Chebyshev: $P(|X_n - E[X_n]| > \frac{n}{12}) < \frac{Var(X_n)}{(\frac{n}{12})^2} = \frac{n\frac{1 \cdot 5}{6 \cdot 6}}{(\frac{n}{12})^2} = \boxed{\frac{20}{n}}$

c. Chernoff: $P(X_n > \frac{n}{4}) < \frac{E[e^{tX_n}]}{e^{\frac{nt}{4}}} = \frac{(\frac{1}{6}e^t + \frac{5}{6})^n}{e^{\frac{nt}{4}}} \leq \frac{e^{\frac{n}{6}(e^t - 1)}}{e^{\frac{nt}{4}}}$

So choosing $t = \ln \frac{3}{2}$ gives a Chernoff bound of approximately $e^{-0.0018n}$.

iv) *Cat C* Chernoff bounds require that random variables represent *independent* events, on the other hand this is not required when using Markov's inequality or Chebyshev's inequality.

b)
Cat B
There are two particular examples mentioned in the mastery material.

The first is to compute better confidence intervals for Possion trials (the mastery material provides an example on this related to genetic sequencing).

The second is to estimate the runtime performance of randomised algorithms (the mastery material provides an example on this related to randomised set balancing algorithms).

Other examples are acceptable (there are many that can be thought of) so long as they highlight the application's use of estimating tail bounds.