Imperial College
London

M3S1

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 Theory

Date: Monday 20 May 2019

Time: 14.00 - 16.00

Time Allowed: 2 Hours

This paper has 4 Questions.

Candidates should use ONE 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.

Imperial College
London

M4/5S1

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 Theory

Date: Monday 20 May 2019

Time: 14.00 - 16.30

Time Allowed: 2 Hours 30 Minutes

This paper has 5 Questions.

Candidates should use ONE 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) The Rao-Blackwell Theorem is an important result in statistics that draws a connection between sufficiency and optimality.

Suppose that X has joint distribution $f_\theta(x)$, $T(X)$ is a sufficient statistic for θ, and that $\widehat{S}$ is an unbiased estimator of θ, with Var$(\widehat{S}) < \infty$. Set $\widehat{S}^* = E[\widehat{S}|T]$.

(i) The first assertion made by the theorem is that $\widehat{S}^*$ is an unbiased estimator for θ. Prove this assertion.

(ii) The second assertion made by the theorem is that Var$(\widehat{S}^*) \leq$ Var$(\widehat{S})$. Prove this assertion. The theorem further asserts that the inequality is strict unless $\widehat{S}$ is a function of T (you are not required to prove this).

(b) Suppose $X_1, X_2, \ldots X_n$ are independent random variables with $X_i \sim N(a_i\theta, 1)$, where $a_1, a_2, \ldots, a_n$ are given known positive constants. Let $\widehat{\theta}$ be an estimator of θ defined as:

$$\widehat{\theta} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n a_i}$$

(i) Show that $\widehat{\theta}$ is an unbiased estimator of θ and compute its variance.

(ii) Show that $S = \sum_{i=1}^n a_i X_i$ is a sufficient statistic for θ.

(iii) Explain without further calculations why $\widehat{\theta}$ cannot be the minimum variance unbiased estimator for θ.

2. (a) Suppose $X_1, \ldots, X_n$ is a random sample of size n from a Poisson distribution with parameter $\lambda > 0$. Show that a Gamma(α, β) prior on λ is a conjugate prior.

(b) Let $X_1, \ldots, X_n$ be independent and identically distributed with probability density function given as:

$$f_\theta(x) = \begin{cases} e^{-(x-\theta)} & \text{if } x \geq \theta, \\ 0 & \text{otherwise.} \end{cases}$$

(i) Find a Maximum Likelihood Estimate for θ.

(ii) Find a Method of Moments (MoM) estimate for θ.

(iii) State one significant shortcoming of using the MoM estimator derived in the previous question to estimate θ.

(iv) The previous question implies that for this particular model the MoM estimate is not an appropriate estimator. However in other contexts MoM estimators are sometimes a useful tool in statistics. State two possible advantages of Method of Moments estimation.

3. (a) (i) Suppose $X_1, \ldots, X_n$ are random variables with joint distribution $f_\theta(x)$. What does it mean to say that a statistic T is complete for a parameter $\theta \in \Theta$?

(ii) Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$, $\lambda \in R^+$. Show that $T(X) = \sum_{i=1}^{n} X_i$ is a complete statistic for λ.

(iii) Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(\theta, \theta+1)$, $\theta \in R^+$. Show that $T(X) = X_1$ is **not** a complete statistic for θ. *Also note that $T(X)$ is not an order statistic.*

(b) (i) For a given Loss function and parameter $\theta \in \Theta$, what does it mean to say that $\widehat{\theta}$ is an inadmissible estimator of θ?

(ii) Let X be a random variable with mean $\theta \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}^+$, and let $aX + b$ be an estimator of θ with $a, b \in \mathbb{R}$. The following cases each impose further restrictions on a and b:

- *Case 1:* $a > 1$.
- *Case 2:* $a < 0$.
- *Case 3:* $a = 1$, $b \neq 0$.

For each of these three cases, show that $aX + b$ is an inadmissible estimator of θ under squared error loss.

4. (a) Define what it means for a family of distributions to have the monotone likelihood ratio, and state the Karlin-Rubin Theorem.

(b) Suppose X is one observation from a population with Beta$(\theta, 1)$ distribution.

(i) Find the most powerful level α test of $H_0 : \theta = 1$ versus $H_1 : \theta = 2$.

(ii) Find a Uniformly Most Powerful test of $H_0 : \theta \leq 1$ versus $H_1 : \theta > 1$, or prove that such a test doesn't exist.

5. (a) State and prove the chain rule for joint entropy.

(b) Let the random variable X have three possible outcomes $\{a, b, c\}$. Consider two distributions on this random variable:

Symbol	$p(x)$	$q(x)$
a	$\frac{1}{2}$	$\frac{1}{3}$
b	$\frac{1}{4}$	$\frac{1}{3}$
c	$\frac{1}{4}$	$\frac{1}{3}$

Verify that the relative entropy between p and q is not equal to the relative entropy between q and p, in other words that $D(p||q) \neq D(q||p)$ (for $\log_2 3$ you can use the value 1.584).

In the following, H and I denote the entropy and the mutual information respectively.

(c) Let X_1 and X_2 be identically distributed but not necessarily independent. Let

$$\rho = 1 - \frac{H(X_2|X_1)}{H(X_1)}$$

(i) Show that:

$$\rho = \frac{I(X_1; X_2)}{H(X_1)}$$

(ii) Show that $0 \leq \rho \leq 1$.

(iii) When is $\rho = 0$?

(iv) When is $\rho = 1$?

(d) Show that if $H(Y|X) = 0$, then Y is a function of X, i.e., for all x with $p(x) > 0$, there is only one possible value of y with $p(x, y) > 0$, where $p(x, y)$ is the joint probability mass function of X and Y, and $p(x)$ is the marginal probability mass function of X.

S1 Formula Sheet

Distribution: Poisson(λ)
Parameters: $\lambda > 0$
pmf:

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad k \in \{0, 1, 2, \ldots\}$$

E[X]: λ
Var[X]: λ

Distribution: Uniform(α, β)
Parameters: $-\infty < \alpha < \beta < \infty$
pdf:

$$f_X(x) = \begin{cases} \frac{1}{\beta - \alpha} & x \in [\alpha, \beta] \\ 0 & \text{otherwise} \end{cases}$$

E[X]: $\frac{1}{2}(\alpha + \beta)$
Var[X]: $\frac{1}{12}(\beta - \alpha)^2$

Distribution: Gamma(α, β)
Parameters: $\alpha > 0$ and $\beta > 0$
pdf:

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\beta x} \qquad x > 0$$

E[X]: $\frac{\alpha}{\beta}$
Var[X]: $\frac{\alpha}{\beta^2}$

Distribution: Normal(μ, σ^2)
Parameters: $\mu \in \mathbb{R}, \quad \sigma^2 > 0$
pdf:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

E[X]: μ
Var[X]: σ^2

Distribution: Beta(α, β)
Parameters: $\alpha > 0, \quad \beta > 0$
pdf:

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad x \in [0,1]$$

E[X]: $\frac{\alpha}{\alpha+\beta}$

The *gamma function* is given by $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx$.
For positive integers n, it satisfies $\Gamma(n) = (n-1)!$.

M345S1 2018-2019 Exam Solutions

Q1

a)
i) [4 Marks, A, seen]
$E[\widehat{S}^*] = E[\ E[\widehat{S}|T]\ ] = E[\ \widehat{S}\ ] = \theta$

Where the second equality follows from the tower law

ii) [6 Marks, A, seen]
$\text{Var}(\widehat{S}) = E[\ \text{Var}(\widehat{S}|T)\ ] + \text{Var}(\ E[\widehat{S}|T]\ )$

This follows from the law of total variance. Now note that $E[\ \text{Var}(\widehat{S}|T)\ ] \geq 0$, since variance is always non-negative. And also $E[\widehat{S}|T] = \widehat{S}^*$ by definition, so putting it together we have that:

$\text{Var}(\widehat{S}) = E[\ \text{Var}(\widehat{S}|T)\ ] + \text{Var}(\ E[\widehat{S}|T]\ ) \geq \text{Var}(\ E[\widehat{S}|T]\ ) = \text{Var}(\widehat{S}^*)$

b)
i) [3 Marks, B, unseen]

$E[\widehat{\theta}] = \frac{\sum E[X_i]}{\sum a_i} = \frac{\theta \sum a_i}{\sum a_i} = \theta$

$var(\widehat{\theta}) = \frac{\sum var(X_i)}{(\sum a_i)^2} = \frac{n}{(\sum a_i)^2}$

ii) [5 Marks, D, unseen]

$$\begin{aligned} f(\boldsymbol{x};\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_i - a_i\theta)^2\right\} \\ &= (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n} x_i^2 - 2\theta \sum_{i=1}^{n} x_i a_i + \theta^2 \sum_{i=1}^{n} a_i^2\right)\right\} \\ &= \left\{\exp\left(\theta \sum_{i=1}^{n} a_i x_i - \theta^2/2 \sum_{i=1}^{n} a_i^2\right)\right\}\left\{\exp\left(-1/2 \sum_{i=1}^{n} x_i^2\right)(2\pi)^{-n/2}\right\} \\ &= g(s;\theta)h(\boldsymbol{x}) \end{aligned}$$

Which implies that S is a sufficient for θ by using Neyman Factorization Theorem.

iii) [2 Marks, C, unseen]

Because the Rao-Blackwell Theorem tells us that there exists an unbiased estimator $T^* = E[\widehat{\theta}|S]$ with variance strictly less than or equal to that of $\widehat{\theta}$.

Q2

a) [2 Marks, A, seen]
We have that the Likelihood satisfies:

$$L(\lambda|x) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_i^n (x_i!)}$$

and the prior:

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}, \quad \lambda > 0$$

So the posterior is given as:

$$\pi(\lambda|x) \propto \lambda^{\sum x_i+\alpha-1}e^{-(n+\beta)\lambda}, \quad \lambda > 0$$

Which shows that $\pi(\lambda|x)$ is gamma$(\sum x_i + \alpha, n + \beta)$.

b)
i)[5 Marks, B, bookwork]
We have that

$$f(x;\theta) = e^{-(x-\theta)}\mathbb{I}_{[\theta,\infty]}x$$

From which it follows that

$$L(\theta;x) = \prod_{i=1}^{n} e^{-(x_i-\theta)}\mathbb{I}_{[\theta,\infty]}x_i = e^{-n(\bar{x}-\theta)}\prod \mathbb{I}_{[\theta,\infty]}x_i = e^{-n(\bar{x}-\theta)}\mathbb{I}_{[\theta,\infty]}x_{(1)}$$

So we need to maximise the function $e^{-n(\bar{x}-\theta)}\mathbb{I}_{[\theta,\infty]}x_{(1)}$ with respect to θ.

This is equivalent to maximising the function $e^{-n(\bar{x}-\theta)}$ for θ in the interval $[-\infty, x_{(1)}]$, and since this is a strictly increasing function in this interval, its maximum is attained at the right endpoint of the interval.

So we conclude that $\widehat{\theta}_{MLE} = x_{(1)}$.

ii) [3 Marks, A, unseen]
set $\bar{X} = E[X] = \int_\theta^\infty xe^{(\theta-x)}dx = \theta + 1$.

Rearranging gives that $\widehat{\theta}_{MoM} = \bar{X} - 1$.

iii)[5 Marks, D, unseen]
It is possible that $\bar{X} - 1$ is larger than $x_{(1)}$, in which case $\widehat{\theta}_{MoM}$ would return an inconsistent result, since it would imply that the probability of $x_{(1)}$ should be zero.

iv)[5 Marks, D, unseen]

Full marks awarded for any two of the following points. Other points are also acceptable so long as they are reasoned appropriately.

- Method of Moments estimators are fairly simple and yield consistent estimators under weak assumptions (though these estimators are often biased).
- MLEs involve optimising functions and it is not always possible to find closed form solutions for these optimisations. This is even common for well-known distributions such as the Gamma distribution. Hence it is typical to compute MLEs by using numerical methods such as the Newton-Raphson method. The main challenge in using such methods is identifying suitable starting points that guarantee convergence, and in practice it has been shown that MoM estimates provide such suitable starting points.
- There are results in the literature that show that for a number of common distributions, MoM estimates provide better estimates (in terms of variance and bias) than MLEs, provided the sample size is not too large. I have mentioned a few of these results in the lecture (they don't need to be cited, mentioning their existence is sufficient to make the point).

Q3

a)
i) [2 Marks, A, seen]
A statistic is said to be complete for θ if for any function g, if $E_\theta[g(T)] = 0$ for all $\theta \in \Theta$, then $P_\theta(g(T)) = 0) = 1$ for all $\theta \in \Theta$.

ii)[4 Marks, A, seen]
For any function g, we have:

$$\begin{aligned}
E_\theta[g(T)] = 0 \quad \forall\theta > 0 &\iff \sum_{t=0}^{\infty} g(t)\frac{e^{n\theta}(n\theta)^t}{t!} = 0 \quad \forall\theta > 0 \\
&\iff \sum_{t=0}^{\infty} g(t)\frac{n^t}{t!}\theta^t = 0 \quad \forall\theta > 0 \\
&\iff \frac{g(t)n^t}{t!} = 0 \quad \forall t = 0, 1, 2, \ldots \quad \forall\theta > 0 \\
&\iff g(t) = 0 \quad \forall t = 0, 1, 2, \ldots, \quad \forall\theta > 0 \\
&\iff P_\theta(g(T) = 0) = 1 \quad \forall\theta > 0
\end{aligned}$$

iii)[5 Marks, C, unseen]
To show that T is not a complete statistic we need to find a non-zero function g, such that $E_\theta[g(T(X))] = 0$.

Since $T(X) = X_1$, we have that $E_\theta[g(T(X))] = E_\theta[g(X_1)] = \int_\theta^{\theta+1} g(x)dx$.

So we need to find a non-zero function $g(x)$ that integrates to zero over the interval $[\theta, \theta + 1]$. Many such functions exist, in particular trigonometric functions with periods adjusted to be equal to one (rather than 2π). For example, the function $g(x) = \sin(2\pi x)$.

b)
i) [2 Marks, A, seen]
For a given loss function L, an estimate $\hat{\theta}$ is an inadmissible estimator of θ if there exists an estimator $\bar{\theta}$ such that:

- $R_\theta(\bar{\theta}) \leq R_\theta(\hat{\theta})$ for all $\theta \in \Theta$
- $R_\theta(\bar{\theta}) < R_\theta(\hat{\theta})$ for some $\theta \in \Theta$

If no such $\bar{\theta}$ exists, then we say that $\hat{\theta}$ is admissible.

ii)[7 Marks, C, unseen]
To ease presentation we will use $\rho(a, b, \theta)$ to denote the risk of the rules. In particular we have that :

$$\rho(a, b, \theta) = R(\theta, aX + b) = a^2\sigma^2 + \{(a-1)\theta + b\}^2$$

Case 1:

$$\rho(a, b, \theta) \geq a^2\sigma^2 > \sigma^2 = \rho(1, 0, \theta)$$

So $aX + b$ is dominated by X when $a > 1$.

Case 2:
If $a < 0$ then $(a-1)^2 > 1$ and

$$\begin{aligned}\rho(a, b, \theta) &\geq \{(a-1)\theta + b\}^2 = (a-1)^2\{\theta + \frac{b}{a-1}\}^2 \\ &> \{\theta + \frac{b}{a-1}\}^2 = \rho(0, -b/(a-1), \theta)\end{aligned}$$

So $aX + b$ is dominated by the constant $-b/(a-1)$ when $a < 0$.

Case 3:
If $a = 1, b \neq 0$

$$\rho(1, b, \theta) = \sigma^2 + b^2 > \sigma^2 = \rho(1, 0, \theta).$$

so $X + b$ is dominated by X when $b \neq 0$.

Q4

a)[6 Marks, A, seen]
A family of distributions $\{f_\theta(x) : \theta \in \Theta\}$ is said to have a monotone likelihood ratio if there exists a function $T(X)$ such that for any $\theta_2 > \theta_1$, the ratio $\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)}$ is a non-decreasing function of $T(X)$.

While the statement of the Karlin-Rubin Theorem is as follow:

Suppose $X = (X_1, \ldots, X_n) \sim f_\theta(x)$ and consider testing $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$. If $\{f_\theta(x) : \theta \in \Theta\}$ has monotone likelihood ratio in $T(X)$, then the UMP test at level α would be

$$\omega(x) = \begin{cases} 1, & \text{if } T(x) \geq k, \\ 0, & \text{if } T(x) < k \end{cases}$$

for some $0 < k < \infty$ and

$$P_{\theta_0}(T(X) \geq K) = \alpha$$

b)

i) [7 Marks, B, bookwork]
By the Neyman-Pearson Lemma, the most powerful test of $H_0 : \theta = 1$ vs $H_1 : \theta = 2$ is given by Reject H_0 if $f(x|2)/f(x|1) > k$ for some $k \geq 0$. Substituting the beta pdf gives:

$$\frac{f(x|2)}{f(x|1)} = \frac{\frac{1}{\beta(2,1)} x^{2-1}(1-x)^{1-1}}{\frac{1}{\beta(1,1)} x^{1-1}(1-x)^{1-1}} = \frac{\Gamma(3)}{\Gamma(2)\Gamma(1)} x = 2x$$

Thus the MP test is Reject H_0 if $X > k/2$. We now use the α level to determine k. We have

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta) = \beta(1) = \int_{k/2}^{1} f_X(x|1)dx = \int_{k/2}^{1} \frac{a}{\beta(1,1)} x^{1-1}(1-x)^{1-1}dx = 1 - \frac{k}{2}$$

Thus $1 - k/2 = \alpha$, so the most powerful α level test is reject H_0 if $X > 1 - \alpha$.

ii)[7 marks, B, bookwork]
for $\theta_2 > \theta_1$, $f(x|\theta_2)/f(x|\theta_1) = (\theta_2/\theta_1)x^{\theta_2-\theta_1}$ is an increasing function of x because $\theta_2 > \theta_1$. So this has an MLR. By the Karlin-Rubin Theorem the test that rejects H_0 if $X > t$ is the UMP test of its size. By the argument in part (b), use $t = 1 - \alpha$ to get size α.

Q5

a)[5 Marks, A, seen]
Chain rule for joint entropy: $H(X,Y) = H(X) + H(Y|X)$.
Proof

$$\begin{aligned}
H(X,Y) &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) \\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x)p(y|x) \\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x) \\
&= -\sum_{x\in X} p(x)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}$$

b) [3 Marks, B, bookwork]
$D(p||q) = \frac{1}{2}\log\frac{3}{2} + \frac{1}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{3}{4} = \log 3 - 1.5 = 0.0849$ bits

$D(q||p) = \frac{1}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{4}{3} + \frac{1}{3}\log\frac{4}{3} = -\log 3 + \frac{5}{3} = 0.0817$ bits

c)
i)[2 Marks, C, unseen]

$$\rho = \frac{H(X_1) - H(X_1|X_2)}{H(X_1)} = \frac{I(X_1;X_2)}{H(X_1)}$$

ii) [2 Marks, C, unseen]
This follows easily from the fact that $0 \leq H(X_1|X_2) \leq H(X_1)$.

iii) [2 Marks, A, seen]
$\rho = 0$ iff $I(X_1;X_2) = 0$ i.e X_1 and X_2 are independent.

iv) [2 Marks, B, bookwork]
$\rho = 1$ iff $H(X_1|X_2) = 0$ i.e X_1 is a function of X_2.

d) [4 Marks, D, unseen]
Assume that there exists an x, say x_0, and two different values of y, say y_1 and y_2 such that $p(x_0, y_1) > 0$ and $p(x_0, y_2) > 0$. So we have that $p(x_0) > p(x_0, y_1) + p(x_0, y_2) > 0$ and also that $p(x_0, y_1)$ and $p(x_0, y_2)$ are not equal to 0 or 1.

Now using the conditional entropy formula we have that:

$$\begin{aligned} H(Y|X) &= -\sum_x \sum_y p(y|x) \log p(y|x) \\ &\geq p(x_0)(-p(y_1|x_0) \log p(y_1|x_0) - p(y_2|x_0) \log p(y_2|x_0)) \\ &> 0 \end{aligned}$$

The above inequality shows that if Y is not a function of X, then $H(Y|X)$ is not equal to 0.

On the other hand, note that $-t \log t \geq 0$ for $0 \leq t \leq 1$, and is strictly positive for t not equal to 0 or 1. This implies that if $H(Y|X) = 0$ then for every term in the sum of the conditional entropy formula above we must have that $p(y|x)$ is equal to 0 or 1, which in turn implies that Y is a function of X. Therefore the conditional entropy $H(Y|X)$ is 0 if and only if Y is a function of X.