IMPERIAL

MATH70043

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May 2024

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

Statistical Theory 1

Date: Tuesday, May 28, 2024

Time: 10:00 – 12:30 (BST)

Time Allowed: 2.5 hours

This paper has 5 Questions.

Please Answer All Questions in 1 Answer Booklet

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

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

Any required additional material(s) will be provided.

Allow margins for marking.

Credit will be given for all questions attempted.

Each question carries equal weight.

DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO

1. Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a vector consisting of random variables $X_1, \ldots, X_n \in \mathbb{R}$ with joint probability density function in a statistical model $\{f_\theta : \theta \in \mathbb{R}\}$. You may assume the statistical model satisfies the usual regularity conditions.

   (a) Define the Fisher information $I_n(\theta)$ for X. State the relationship between the Fisher information and the derivative of the score function. (2 marks)

   We say that the Fisher information *tensorizes* if $I_n(\theta) = nI_1(\theta)$, where $I_1(\theta)$ is the Fisher information of X_1.

   (b) Suppose that $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables. Show that the Fisher information tensorizes. (5 marks)

   You may assume any results about the score function proved in the lectures, provided these are clearly stated.

   (c) Consider the model $X_1 = \theta + \varepsilon_1$ and

$$X_i = \theta(1 - \sqrt{\alpha}) + \sqrt{\alpha} X_{i-1} + \sqrt{1 - \alpha}\varepsilon_i, \qquad i = 2, \ldots, n,$$

   where $\varepsilon_1, \ldots, \varepsilon_n \sim^{iid} N(0, 1)$ and $\alpha \in [0, 1)$ is a *known* constant. Compute the Fisher information $I_n(\theta)$ in this model.

   For what values of α does the Fisher information tensorize? (9 marks)

   (d) State a lower bound for the variance of an unbiased estimator of θ in this model. Briefly comment on how the value of α affects the difficulty of estimating θ. (4 marks)

(Total: 20 marks)

2. Consider i.i.d. observations $X_1, \ldots, X_n \sim^{iid} N_d(\theta, I_d)$ from the d-dimensional multivariate normal model $\{N_d(\theta, I_d) : \theta \in \mathbb{R}^d\}$, where I_d denotes the $d \times d$ identity matrix. We are interested in estimation of θ.

[Recall: for $\mu \in \mathbb{R}^d$, the $N_d(\mu, I_d)$ distribution has density function:

$$\frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{1}{2}(x-\theta)^T(x-\theta)\right\}, \qquad x \in \mathbb{R}^d.$$

You may use that for $Z \sim N_d(0, I_d)$, $\|Z\|_2^2 \sim \chi_d$ follows a chi-squared distribution with d degrees of freedom, where $\|x\|_2 = (\sum_{k=1}^d x_k^2)^{1/2}$ denotes the usual Euclidean norm on $\mathbb{R}^d$.]

(a) Show that the maximum likelihood estimator of $\theta \in \mathbb{R}^d$ is $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$. What is the distribution of $\hat{\theta}_n$? (4 marks)

(b) For $\alpha \in (0,1)$, construct a confidence set $C_n^\alpha \subseteq \mathbb{R}^d$ such that $P_\theta(\theta \in C_n^\alpha) = 1 - \alpha$. (3 marks)

(c) Consider testing the hypotheses

$$H_0 : \theta = 0, \qquad \text{against} \qquad H_1 : \theta \neq 0.$$

Compute the likelihood ratio statistic $\Lambda(x) = \Lambda(x; H_0, H_1)$ and give the distribution of $2\log\Lambda(X)$ under the null hypothesis H_0. Compare the result with the asymptotic distribution given by Wilks' theorem. (6 marks)

(d) Compute the maximum likelihood estimator for θ over the following restricted parameter spaces:

(i) $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 = 1\}$,

(ii) $\Theta = \{\theta \in \mathbb{R}^d : v^T\theta = \sum_{j=1}^d v_j\theta_j = 0\}$ for some fixed unit vector $v \in \mathbb{R}^d$.

(7 marks)

(Total: 20 marks)

3. (a) In the context of hypothesis testing, define the following terms: *power function*, *type I error*, *type II error*, *critical region* and *uniformly most powerful test*. (5 marks)

Let $X_1, \ldots, X_n \sim^{iid} N(\theta, 1)$ be i.i.d. normal random variables. Consider testing the hypotheses

$$H_0 : \theta = 0 \qquad \text{against} \qquad H_1 : \theta = \theta',$$

for some $\theta' > 0$. Let $\Phi(t)$ denote the denote the cumulative distribution function of the standard normal $N(0, 1)$ distribution and $z_{1-\alpha}$ be its $(1-\alpha)$-quantile, i.e. $\Phi(z_{1-\alpha}) = P(N(0, 1) \leq z_{1-\alpha}) = 1 - \alpha$.

(b) (i) Find the most powerful test of size $\alpha \in (0, 1)$ for testing H_0 against H_1. (7 marks)

(ii) Compute the power of the test in (b)(i) in terms of Φ and $z_{1-\alpha}$ (or z_α). (4 marks)

(c) Consider now the hypotheses

$$H_0' : \theta \leq 0 \qquad \text{and} \qquad H_1' : \theta > 0.$$

Is your test in (b) uniformly most powerful for testing H_0' against H_1'? Justify your answer. (4 marks)

(Total: 20 marks)

4. (a) In the context of decision theory, explain the meaning of the following terms: *loss function, decision rule*, the *risk function* of a decision rule and the *Bayes risk* of a decision rule with respect to a prior π.

Explain how a Bayes rule with respect to a prior π can be constructed. (6 marks)

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables coming from a density

$$f_\theta(y) = \frac{y}{\theta^2} e^{-y/\theta}, \qquad y > 0,$$

with $\theta > 0$.

[See below for some properties of distributions you may use in this question].

(b) Compute the maximum likelihood estimator $\hat{\theta}_n$ of θ. (2 marks)

(c) Compute the Jeffreys prior for the parameter θ in this model. Derive the posterior distribution for θ given $Y_1, \ldots, Y_n$ based on the Jeffreys prior. (4 marks)

For $c > 0$, consider the loss function

$$L(a, \theta) = e^{c(\frac{a}{\theta} - 1)} - c\left(\frac{a}{\theta} - 1\right) - 1.$$

(d) For the loss function L, show that the Bayes rule with respect to the Jeffreys prior derived in (c) takes the form

$$\delta(Y) = \frac{1}{c}(1 - e^{-\frac{c}{2n+1}}) \sum_{i=1}^{n} Y_i.$$

[You may find it simplifies your notation to write $T = \sum_{i=1}^n Y_i$.] (8 marks)

[Recall that if Z has inverse Gamma $IG(\alpha, \beta)$ distribution with density

$$\frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} e^{-\beta/z}, \qquad z > 0,$$

then $EZ = \frac{\beta}{\alpha - 1}$ and $E(1/Z) = \frac{\alpha}{\beta}$. If Y has Gamma(k, θ) distribution with density

$$\frac{y^{k-1}}{\theta^k \Gamma(k)} e^{-y/\theta}, \qquad y > 0,$$

then $EY = k\theta$.]

(Total: 20 marks)

5. (a) Given independent and identically distributed (i.i.d.) observations $X_1, \ldots, X_n$ with finite mean $EX_i = \mu$ and variance $\text{Var}(X_i) = \sigma^2$, explain the notion of a *bootstrap* sample $X_1^*, \ldots, X_n^*$.

Suppose we are interested in some statistic $T_n = T_n(X_1, \ldots, X_n)$ of the data. Describe how you can use $B \geq 2$ bootstrap samples to construct an estimate $\hat{\tau}_B^2$ of the variance of T_n.

(6 marks)

Consider now inference for the parameter $\theta = 1/(EX_1)$, where $\mu = EX_1 \neq 0$. A proposed 95% bootstrap confidence interval for θ is

$$C_n = \left[\frac{1}{\bar{X}_n} - 1.96\hat{\tau}_B^*, \frac{1}{\bar{X}_n} - 1.96\hat{\tau}_B^* \right], \tag{1}$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\hat{\tau}_B^2$ is the estimate in (a) based on $T_n(X_1, \ldots, X_n) = 1/\bar{X}_n$.

(b) Show that $\sqrt{n}(\frac{1}{\bar{X}_n} - \frac{1}{\mu}) \to^d N(0, v^2)$ as $n \to \infty$ for some $v^2 = v^2(\mu, \sigma)$ which you should evaluate. (4 marks)

(c) Consider a bootstrap sample $X_1^*, \ldots, X_n^*$ and let $\bar{X}_n^* = \frac{1}{n}\sum_{i=1}^n X_i^*$. Show that conditional on $X_1, \ldots, X_n$,

$$\frac{\sqrt{n}}{v_n}\left(\frac{1}{\bar{X}_n^*} - \frac{1}{\bar{X}_n}\right) | X_1, \ldots, X_n \to^d N(0,1)$$

as $n \to \infty$ for some $v_n^2 = v_n^2(X_1, \ldots, X_n)$ which you should evaluate.

(4 marks)

You may assume that the triangular central limit theorem holds for $X_1^*, \ldots, X_n^*$ conditionally on $X_1, \ldots, X_n$. This result states that if $Y_{1,n}, \ldots, Y_{n,n}$ are i.i.d. random variables with mean $EY_{i,n} = \mu_n$ and variance Var$(Y_{i,n}) = \sigma_n^2$, then

$$\frac{\sqrt{n}}{\sigma_n}\left(\frac{1}{n}\sum_{i=1}^n Y_{i,n} - \mu_n\right) \to^d N(0,1)$$

as $n \to \infty$, i.e. the central limit theorem holds when $Y_{1,n}, \ldots, Y_{n,n}$ are i.i.d. for every $n \geq 1$, but their distribution can change with n.

(d) Using your answers to (b) and (c) or otherwise, justify the use of C_n in (1) as an (asymptotic) 95% confidence interval for B and n large enough.

(6 marks)

(Total: 20 marks)

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2024

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

MATH60043/MATH70043

Statistical Theory (Solutions)

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

1. (a) The Fisher information is

seen ⇓

$$I_n(\theta) = E_\theta\left[\left(\frac{d}{d\theta}\log f_\theta(X)\right)^2\right].$$

Let $\ell_n(\theta) = \log f_\theta(x)$ denote the corresponding log-likelihood. Under the regularity conditions, $I_n(\theta) = -E_\theta[\ell_n''(\theta)]$ (or if $S_n(\theta) = \ell_n'(\theta)$ is the score function, $I_n(\theta) = -E_\theta S_n'(\theta)$).

2, A

(b) Using the i.i.d. structure, $\ell_n(\theta) = \log\prod_{i=1}^n f_{X_1,\theta}(x_i) = \sum_{i=1}^n \ell_1(\theta; x_i)$. Thus

seen ⇓

$$\begin{aligned} I_n(\theta) &= E_\theta\left[\sum_{i=1}^n\sum_{j=1}^n \ell_1'(\theta;X_i)\ell_1'(\theta;X_j)\right] \\ &= \sum_{i=1}^n E_\theta\ell_1'(\theta;X_i)^2 + \sum_{i=1}^n\sum_{j\neq i} E_\theta\ell_1'(\theta;X_i)E_\theta\ell_1'(\theta;X_j) \\ &= nI_1(\theta) + n(n-1)(E_\theta\ell_1'(\theta;X_i))^2. \end{aligned}$$

But it is proved in the lectures that under the regularity conditions, $E_\theta\ell_1'(\theta;X_1) = E_\theta S_1(\theta) = 0$ and hence the second term vanishes.

5, A

(c) We have $X_1 \sim N(\theta,1)$ and $X_i|X_1,\ldots,X_{i-1} =^d X_i|X_{i-1} \sim N(\theta(1-\sqrt{\alpha}) + \sqrt{\alpha}X_{i-1}, 1-\alpha)$. Thus the joint density equals

sim. seen ⇓

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{(x_1-\theta)^2}{2}}\prod_{i=2}^n \frac{1}{\sqrt{2\pi(1-\alpha)}}e^{-\frac{(x_i-\theta(1-\sqrt{\alpha})-\sqrt{\alpha}x_{i-1})^2}{2(1-\alpha)}},$$

giving log-likelihood and its derivatives:

$$\begin{aligned} \ell_n(\theta) &= -\frac{1}{2}(x_1-\theta)^2 - \frac{1}{2(1-\alpha)}\sum_{i=2}^n(x_i-\theta(1-\sqrt{\alpha})-\sqrt{\alpha}x_{i-1})^2 + C, \\ \ell_n'(\theta) &= (x_1-\theta) + \frac{1-\sqrt{\alpha}}{1-\alpha}\sum_{i=2}^n(x_i-\theta(1-\sqrt{\alpha})-\sqrt{\alpha}x_{i-1}), \\ \ell_n''(\theta) &= -1-(n-1)\frac{(1-\sqrt{\alpha})^2}{1-\alpha} = -1-(n-1)\frac{1-\sqrt{\alpha}}{1+\sqrt{\alpha}}, \end{aligned}$$

where the constant C does not depend on θ. We thus obtain

$$I_n(\theta) = -E_\theta[\ell_n''(\theta)] = 1 + (n-1)\frac{1-\sqrt{\alpha}}{1+\sqrt{\alpha}}.$$

Since $I_1(\theta) = 1$, the Fisher information tensorizes if and only if $I_n(\theta) = n$, i.e. $\frac{1-\sqrt{\alpha}}{1+\sqrt{\alpha}} = 1$. This happens if and only if $\alpha = 0$.

9, B

(d) Using the Cramer-Rao lower bound, we have

sim. seen ⇓

$$\mathrm{Var}_\theta(\hat{\theta}_n) \geq \frac{1}{I_n(\theta)} = \frac{1}{1+(n-1)\frac{1-\sqrt{\alpha}}{1+\sqrt{\alpha}}}.$$

The Fisher information is a strictly decreasing function of α on $[0,1)$, thus the larger the value of α, the more difficult it is to estimate θ. Even though the noise is smaller, the increasing correlation between X_{i-1} and X_i means there is less information in each additional observation. The most informative setting is when $\alpha = 0$, in which case we have i.i.d. observations.

4, C

2. (a) The log-likelihood equals

sim. seen ⇓

$$\ell_n(\theta) = \log\left(\prod_{i=1}^n \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|x_i-\theta\|_2^2}\right) = C - \frac{1}{2}\sum_{i=1}^n (x_i-\theta)^T(x_i-\theta),$$

where C is independent of θ. Taking the gradient,

$$\nabla \ell_n(\theta) = \sum_{i=1}^n (x_i - \theta) = 0,$$

which is solved by $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n x_i$ [we can also check that the Hessian is $-I_d$, which is negative-definite, and hence this is a global maximum]. We see $\hat{\theta}_n \sim N_d(\theta, \frac{1}{n} I_d)$.

4, A

(b) We have $\sqrt{n}(\hat{\theta}_n - \theta) \sim N_d(0, I_d)$ and hence $n\|\hat{\theta}_n - \theta\|_2^2 \sim \chi^2_d$ using the hint in the question. We thus have $P_\theta(n\|\hat{\theta}_n - \theta\|_2^2 \le q_d(1-\alpha)) = P(\chi^2_d \le q_d(1-\alpha)) = 1-\alpha$, where $q_d(1-\alpha)$ is the $1-\alpha$ quantile of the χ^2_d distribution. We thus have the confidence set

sim. seen ⇓

$$C_n^\alpha = \{\theta \in \mathbb{R}^d : n\|\theta - \hat{\theta}_n\|_2^2 \le q_d(1-\alpha)\}.$$

[Any valid answer acceptable].

3, A

(c) The MLEs over H_0 and H_1 are 0 and $\hat{\theta}_n = \bar{x}_n$, respectively, so

sim. seen ⇓

$$\begin{aligned}\Lambda(x) = \frac{f_{\hat{\theta}_n}(x)}{f_0(x)} &= \frac{\prod_i \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2}\|x_i - \bar{x}_n\|_2^2)}{\prod_i \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2}\|x_i - 0\|_2^2)} \\ &= \exp\left\{\frac{1}{2}\sum_{i=1}^n \|x_i\|_2^2 - \frac{1}{2}\sum_{i=1}^n \|x_i - \bar{x}_n\|_2^2\right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}2\log\Lambda(x) &= \sum_{i=1}^n \|x_i\|_2^2 - \sum_{i=1}^n \|x_i - \bar{x}_n\|_2^2 \\ &= \sum_{i=1}^n [2\langle x_i, \bar{x}_n\rangle_2 - \|\bar{x}_n\|_2^2] \\ &= 2n\|\bar{x}_n\|_2^2 - n\|\bar{x}_n\|_2^2 = n\|\bar{x}_n\|_2^2.\end{aligned}$$

But under H_0, $\sqrt{n}\bar{X}_n \sim N_d(0, I_d)$ and hence $2\log\Lambda(X) = \|\sqrt{n}\bar{X}_n\|_2^2 \sim \chi^2_d$. This matches exactly the asymptotic distribution given by Wilks' theorem.

6, A

(d) (i) Restricting to $\|\theta\|_2 = 1$, the log-likelihood derived in (a) equals

unseen ⇓

$$\ell_n(\theta) = C - \frac{1}{2}\sum_{i=1}^n \left[\|x_i\|_2^2 - 2\langle x_i, \theta\rangle_2 + \|\theta\|_2^2\right] = C' + n\langle \bar{x}_n, \theta\rangle_2,$$

where C' does not depend on θ. The inner product is maximized by the unit vector θ lying in the same direction as $\bar{x}_n$, i.e. $\hat{\theta}_n = \bar{x}_n / \|\bar{x}_n\|_2$.

(ii) We note $v^T\theta = \sum_{j=1}^d v_j\theta_j = 0$ implies $\theta_d = w_1\theta_1 + \cdots + w_{d-1}\theta_{d-1}$ for $w_j = -v_j/v_d$. Then

$$\begin{aligned}\ell_n(\theta) &= C - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^d (x_{ij} - \theta_j)^2 \\ &= C - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^{d-1}(x_{ij} - \theta_j)^2 - \frac{1}{2}\sum_{i=1}^n (x_{id} - w_1\theta_1 - \cdots - w_{d-1}\theta_{d-1})^2.\end{aligned}$$

Differentiating with respect to θ_r, $r = 1, \ldots, d-1$,

$$\frac{\partial \ell_n}{\partial \theta_r} = \sum_{i=1}^{n}(x_{ir} - \theta_r) + w_r \sum_{i=1}^{n}(x_{id} - \alpha) = 0,$$

where $\alpha = \theta_d = w_1\theta_1 + \cdots + w_{d-1}\theta_{d-1}$. Solving for θ_r gives

$$\theta_r = \frac{1}{n}\sum_{i=1}^{n} x_{ir} + \frac{w_r}{n}\sum_{i=1}^{n}(x_{id} - \alpha) = \bar{x}_r + w_r\bar{x}_d - w_r\alpha.$$

But plugging back in the definition of α:

$$\alpha = w_1\theta_1 + \cdots + w_{d-1}\theta_{d-1} = \sum_{j=1}^{d-1} w_j\bar{x}_j + (\bar{x}_d - \alpha)\sum_{j=1}^{d-1} w_j^2$$

Using that $w_j = -v_j/v_d$ and multiplying up by v_d^2,

$$\alpha\sum_{j=1}^{d} v_j^2 = \alpha = v_d^2\sum_{j=1}^{d-1} w_j\bar{x}_j + \bar{x}_d\sum_{j=1}^{d-1} v_j^2.$$

It remains only to substitute this value of α back into θ_r to obtain

$$\begin{aligned}\theta_r &= \bar{x}_r - \frac{v_r}{v_d}\bar{x}_d - v_r\sum_{j=1}^{d-1} v_j\bar{x}_j + \frac{v_r}{v_d}\bar{x}_d\sum_{j=1}^{d-1} v_j^2 \\ &= \bar{x}_r - v_r\sum_{j=1}^{d-1} v_j\bar{x}_j - \frac{v_r}{v_d}\bar{x}_d\left(1 - \sum_{j=1}^{d-1} v_j^2\right) = \bar{x}_r - v_r\sum_{j=1}^{d} v_j\bar{x}_j.\end{aligned}$$

In summary, $\hat{\theta}_n = \bar{x}_n - (v^T\bar{x}_n)v$, the projection of $\bar{x}_n$ onto the space Θ. [There are other shorter ways to derive this, e.g. using Lagrange multipliers or perhaps arguing more directly].

7, D

3. (a) The *power function* $\pi_\phi : \Theta \to [0,1]$ of a test ϕ is the probability of rejecting the null hypothesis H_0 under P_θ, i.e. $\pi_\phi(\theta) = P_\theta(\text{reject } H_0)$. seen ⇓

A *type I error* is the error of rejecting the null hypothesis H_0 when it is actually true.

A *type II error* is the error of rejecting the alternative hypothesis H_1 when it is actually true.

The *critical region* of a test ϕ is the region R where we reject the null hypothesis if the data (or statistic) falls in R, i.e. $X \in R$.

A test is *uniformly most powerful* of size α for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ if (i) it is a level α test ($\sup_{\theta\in\Theta_0} \pi_\phi(\theta) \leq \alpha$) and (ii) any other level α test ϕ^* has smaller power, i.e. $\pi_{\phi^*}(\theta) \leq \pi_\phi(\theta)$ for all $\theta \in \Theta_1$.

5, A

(b) (i) Since we have simple hypotheses, the Neyman-Pearson lemma implies the UMP test is given by the likelihood ratio test. We thus reject H_0 if the following is large: sim. seen ⇓

$$\frac{f_{\theta'}(x)}{f_0(x)} = \frac{\prod_i \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\theta')^2}}{\prod_i \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2}} = \exp\left\{2\theta' \sum_i x_i - n(\theta')^2\right\}.$$

Since $\theta' > 0$, this is equivalent to rejecting H_0 if $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is large. Since $\bar{X}_n \sim N(0, 1/n)$ under H_0, we pick t such that

$$P_0(\bar{X}_n > t) = P(N(0,1/n) > t) = P(N(0,1) > t\sqrt{n}) = 1 - \Phi(t\sqrt{n}) = \alpha,$$

or equivalently $t\sqrt{n} = z_{1-\alpha}$, the $1-\alpha$ quantile of the standard normal distribution. We thus reject H_0 if and only if $\bar{X}_n > z_{1-\alpha}/\sqrt{n}$.

7, B

(ii) The power of the test is sim. seen ⇓

$$\begin{aligned} P_{\theta'}(\text{reject } H_0) = P_{\theta'}(\bar{X}_n > t) &= P(N(\theta', 1/n) > z_{1-\alpha}/\sqrt{n}) \\ &= P(\theta' + n^{-1/2}N(0,1) > z_{1-\alpha}/\sqrt{n}) \\ &= P(N(0,1) > z_{1-\alpha} - \sqrt{n}\theta') \\ &= 1 - \Phi(z_{1-\alpha} - \sqrt{n}\theta'). \end{aligned}$$

4, C

(c) The UMP test we obtain in (b) is the same for all $\theta' > 0$, thus it will be UMP for testing the one-sided hypothesis $H_1' : \theta > 0$. sim. seen ⇓

This can also be seen more formally by noting that this family has monotone likelihood ratio in $T(X) = \bar{X}_n$ (or $\sum_i X_i$) since for $\theta_1 < \theta_2$:

$$\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} = \exp\left\{(\theta_2 - \theta_1)n\bar{X}_n + \tfrac{1}{2}(\theta_1^2 - \theta_2^2)\right\},$$

which is an increasing function of $T(x)$. Hence the Karlin-Rubin theorem tells us that UMP test is of the form reject H_0' if and only if $T(x) \geq k$ for some k (which we worked out above).

4, C

4. (a) A *loss function* is a non-negative function $L : \mathcal{A} \times \Theta \to [0, \infty)$ that determines the cost of action $a \in \mathcal{A}$ for a given parameter $\theta \in \Theta$. seen ⇓

A *decision rule* $\delta : \mathcal{X} \to \mathcal{A}$ makes a decision/action $\delta(X)$ upon observing X.

The *risk function* of δ is the expected loss under P_θ as a function of θ: $R(\delta, \theta) = E_\theta[L(\delta(X), \theta)]$.

A *Bayes rule* with respect to a prior π is any decision rule that minimize the Bayes risk $R_\pi(\delta) = E_{\theta\sim\pi}[R(\delta, \theta)]$, where the expectation is taken over the prior π.

A Bayes rule can be obtained by directly minimizing the Bayes risk. However, a more common approach is to minimize the posterior risk $R_\pi(\delta(x)) = E_\pi[L(\delta(x)), \theta)|x]$, which is the expected loss under the posterior. This is because any minimizer of the posterior risk also minimizes the Bayes risk. 6, A

(b) The log-likelihood sim. seen ⇓

$$\ell_n(\theta) = \log\left(\frac{1}{\theta^{2n}} \prod y_i e^{-\frac{1}{\theta}\sum_i y_i}\right) = -2n \log\theta + C - \frac{1}{\theta}\sum_{i=1}^n y_i,$$

where C is independent of θ. Differentiating,

$$\ell_n'(\theta) = -\frac{2n}{\theta} + \frac{\sum y_i}{\theta^2}.$$

Setting $\ell_n'(\theta) = 0$ yields $\theta = \frac{1}{2n}\sum y_i$ [note $\ell_n'(\theta) \geq 0$ if and only if $\frac{1}{2n}\sum y_i \geq \theta$, hence this is the MLE]. 2, B

(c) We have Fisher information sim. seen ⇓

$$I_n(\theta) = -E_\theta \ell_n''(\theta) = -E_\theta\left[\frac{2n}{\theta^2} - \frac{2\sum Y_i}{\theta^3}\right].$$

But since $Y_i \sim Gamma(2, \theta)$, we have $EY_i = 2\theta$. Thus $I_n(\theta) = \frac{4n\theta}{\theta^3} - \frac{2n}{\theta^2} = \frac{2n}{\theta^2}$. The Jeffreys prior is then $\pi(\theta) \propto 1/\theta$.

Turning to the posterior,

$$\pi(\theta|y_1, \ldots, y_n) \propto \prod f_\theta(y_i)\pi(\theta) \propto \theta^{-2n-1} e^{-\frac{1}{\theta}\sum y_i},$$

which we recognize as the form of an $IG(2n, \sum_i y_i)$ distribution. 4, B

(d) Let $E^\pi[\cdot|Y]$ denote the posterior expectation. The Bayes rule minimizes the posterior risk, which is the approach we take. Let $\delta = \delta(y)$ be a decision rule: unseen ⇓

$$E^\Pi[L(\delta, \theta)|Y] = e^{-c} E^\Pi[e^{c\delta/\theta}|Y] - c\delta E^\Pi[1/\theta|Y] + c - 1.$$

Differentiating with respect to δ,

$$ce^{-c} E^\Pi[\tfrac{1}{\theta} e^{c\delta/\theta}|Y] - cE^\Pi[1/\theta|Y] = 0, \tag{1}$$

which we must solve in δ. Turning to the expectations,

$$E^\Pi[1/\theta|Y] = \frac{2n}{\sum Y_i} = \frac{2n}{T}$$

using that the posterior is $IG(2n, T)$ and the hint. For the first expectation,

$$\begin{aligned} E^\Pi[\tfrac{1}{\theta} e^{c\delta/\theta}|Y] &= \int_0^\infty \frac{1}{\theta} e^{c\delta/\theta} \frac{T^{2n}}{\Gamma(2n)} \theta^{-2n-1} e^{-T/\theta} d\theta \\ &= \frac{T^{2n}}{\Gamma(2n)} \int_0^\infty \theta^{-2n-2} e^{-(T-c\delta)/\theta} d\theta \\ &= \frac{T^{2n}}{\Gamma(2n)} \frac{\Gamma(2n+1)}{(T-c\delta)^{2n+1}}. \end{aligned}$$

Substituting these two values into (1) gives

$$ce^{-c}2n\frac{T^{2n}}{(T-c\delta)^{2n+1}} - \frac{2nc}{T} = 0.$$

Rearranging,

$$e^{-c} = \left(\frac{T-c\delta}{T}\right)^{2n+1} = \left(1-\frac{c\delta}{T}\right)^{2n+1},$$

which yields

$$\delta = \frac{T}{c}(1-e^{-\frac{c}{2n+1}})$$

as required. $\boxed{8, \text{D}}$

5. (a) A bootstrap sample is a sample of n i.i.d. observations drawn from the empirical distribution function $F_n(t) = \frac{1}{n}\sum_{i=1}^n 1\{X_i \le t\}$. Alternatively, we select each X_i^* with probability $P(X_i^* = X_k | X_1, \ldots, X_n) = 1/n$ for $k = 1, \ldots, n$, i.e. we set X_i^* equal to one of the observations $X_1, \ldots, X_n$, each having equal probability $1/n$.

seen ⇓

3, M

We draw B i.i.d. bootstrap samples $X_{b,1}^*, \ldots, X_{b,n}^* \sim^{iid} F_n$ for $b = 1, \ldots, B$ large, and for each $b = 1, \ldots, B$ we compute the statistics $T_{n,b}^* = T_n(X_{b,1}^*, \ldots, X_{b,n}^*)$ based on the b^{th} bootstrap sample. The bootstrap estimate of the variance of T_n is then the sample variance of $(T_{n,b}^* : b = 1, \ldots, B)$, i.e.

$$\hat{\tau}_B^2 = \frac{1}{B}\sum_{b=1}^B \left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^B T_{n,r}^* \right)^2 .$$

(or with the centering replaced directly by T_n.)

3, M

(b) We have by the central limit theorem $\sqrt{n}(\bar{X}_n - \mu) \to^d N(0, \sigma^2)$. Using the delta method with $g(\mu) = 1/\mu$ so that $g'(\mu) = -1/\mu^2$, we get

sim. seen ⇓

$$\sqrt{n}(1/\bar{X}_n - 1/\mu) \to^d N(0, g'(\mu)^2\sigma^2) = N(0, \frac{\sigma^2}{\mu^4}).$$

4, M

(c) We note that conditionally on $X_1, \ldots, X_n$, the bootstrap sample $X_1^*, \ldots, X_n^*$ are i.i.d. observations from the empirical distribution. Thus we can apply the triangular central limit theorem to obtain

$$\frac{\sqrt{n}}{\sigma_n}\left(\bar{X}_n^* - \mu_n\right) | X_1, \ldots, X_n \to^d N(0,1)$$

with $\mu_n = \bar{X}_n$ and

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2 .$$

We can then apply the delta method as in (b) to deduce

$$\frac{\sqrt{n}}{v_n}(1/\bar{X}_n - 1/\mu_n) | X_1, \ldots, X_n \to^d N(0,1),$$

where $v_n^2 = \frac{\sigma_n^2}{\mu_n^4}$.

4, M

(d) We wish to evaluate the coverage $P_\mu(1/\mu \in C_n)$ as $n \to \infty$. We know from the limit distribution in (b) that after rearranging,

unseen ⇓

$$P\left(\frac{1}{\mu} \in \left[\frac{1}{\bar{X}_n} - 1.96\frac{\sigma}{\mu^2\sqrt{n}}, \frac{1}{\bar{X}_n} + 1.96\frac{\sigma}{\mu^2\sqrt{n}}\right]\right) \to 0.95$$

as $n \to \infty$. It thus remains to show that we can replace the unknown σ/μ^2 by the estimate $\hat{\tau}_B$ without changing this probability. But $\hat{\tau}_B^2$ is a consistent estimate of the variance of $1/\bar{X}_n^*$ (conditional on $X_1, \ldots, X_n$), so by the law of large numbers, $\hat{\tau}_B^2 \to^P \text{Var}(1/\bar{X}_n^* | X_1, \ldots, X_n)$ as $B \to \infty$. But by (c), this second variance is approximately σ_n^2/μ_n^4 for n large enough. Thus for n, B large (the latter which we can control), $|\hat{\tau}_B^2 - \sigma_n^2/\mu_n^4| = o(1)$ with probability tending to one (under the bootstrap distribution).

Using the weak law of large numbers and the continuous mapping theorem,

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2 \to^p EX_1^2 - (EX_1)^2 = \mathsf{Var}(X_1) = \sigma^2$$

as $n \to \infty$, while $\mu_n = \bar{X}_n \to^p EX_1 = \mu$. So by Slutsky's theorem, $\sigma_n^2/\mu_n^4 \to^P \sigma^2/\mu^4$. Thus with probability tending to one, we have $|\hat{\tau}_B^2 - \sigma^2/\mu^4| = o(1)$. We can thus write $\hat{\tau}_B = \frac{\sigma}{\mu^2} + o(1)$ with probability tending to one as $n, B \to \infty$, which justifies using the bootstrap estimate of variance in the confidence interval. $\boxed{6,\ \mathrm{M}}$

Review of mark distribution:
Total A marks: 31 of 32 marks
Total B marks: 22 of 20 marks
Total C marks: 12 of 12 marks
Total D marks: 15 of 16 marks
Total marks: 100 of 80 marks
Total Mastery marks: 20 of 20 marks

MATH60043 Statistical Theory

Question	Marker's comment
1	Several candidates struggled already with the definition of Fisher information. For those who were clear on the definition, (b) was generally well-answered. Part (c) caused a lot of issues, with many candidates deciding the model was i.i.d. even when it was clearly not.
2	(a) was generally well answered. (b) Many candidates wrote down a 1-dimensional answer that didn't make sense in d-dimensions. (c) Was broadly well answered. (d) This was reasonably well-answered. Some candidates came up with quite varied and interesting justifications for the MLE, so there were multiple ways to solve this question (e.g. direct computations, Lagrangians, arguing by projection).
3	This question was broadly well answered. Some candidates struggled with the notion of power, but if they understood the definition, they often made significant progress.
4	Parts (a)-(c) were generally well answered. (d) Many candidates could at least make a start and understood what was required, making significant progress on the computation.

MATH70043 Statistical Theory

Question	Marker's comment
1	Several candidates struggled already with the definition of Fisher information. For those who were clear on the definition, (b) was generally well-answered. Part (c) caused a lot of issues, with many candidates deciding the model was i.i.d. even when it was clearly not.
2	(a) was generally well answered. (b) Many candidates wrote down a 1-dimensional answer that didn't make sense in d-dimensions. (c) Was broadly well answered. (d) This was reasonably well-answered. Some candidates came up with quite varied and interesting justifications for the MLE, so there were multiple ways to solve this question (e.g. direct computations, Lagrangians, arguing by projection).
3	This question was broadly well answered. Some candidates struggled with the notion of power, but if they understood the definition, they often made significant progress.
4	Parts (a)-(c) were generally well answered. (d) Many candidates could at least make a start and understood what was required, making significant progress on the computation.
5	Parts (a)-(b) where well answered. Many candidates managed (c). Many candidates did not attempt (d).