Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2025	MATH2687-WE01

Title: Data Science and Statistical Computing II

Time:	2 hours	
Additional Material provided:	Tables: Normal distribution, t-distribution.	
Materials Permitted:		
Calculators Permitted:	Yes	Models Permitted: Casio FX83 series or FX85 series.

Instructions to Candidates:	Answer all questions. Section A is worth 40% and Section B is worth 60%. Within each section, all questions carry equal marks. Write your answer in the white-covered answer booklet with barcodes. Begin your answer to each question on a new page.

Revision:	

ED01/2025

SECTION A

Q1 (a) Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ consist of a set of n independent observations of a random variable, X, having unknown probability distribution. The resampling procedure in the non-parametric Bootstrap is justified as being equivalent to simulating from the empirical cumulative distribution function (ecdf) of the sample $\mathbf{x}$.

(i) Define the ecdf and write down the corresponding probability mass function.

(ii) Derive the expectation and variance of a sample drawn from the ecdf.

The parameter of interest, θ, is the median and you use the sample median as the estimator $S(\cdot)$. You collect data:

$$x_1 = 0.2, \quad x_2 = 0.7, \quad x_3 = 0.3, \quad x_4 = 0, \quad x_5 = 0.5$$

(b) Compute $S(\mathbf{x}^\star)$ for the bootstrap resample:

$$\mathbf{x}^\star = (x_2, x_1, x_4, x_1, x_5)$$

(c) Bootstrap simulation for the data above gives $\bar{S}^\star = 0.3$ and $\widehat{\mathrm{Var}}\big(S(\mathbf{w}, \boldsymbol{\gamma})\big) = 0.0274$. Compute the 95% Normal confidence interval (CI) and comment on whether you consider there to be any appreciable bias in the mean estimator.

Q2 The following integral cannot be evaluated analytically:

$$\int_0^2 \exp\left(-x^2\right) \sin\left(\frac{1}{x}\right) \, dx$$

(a) Express the integral as an expectation with respect to a Uniform distribution and write down the Monte Carlo estimator of this integral given simulations $\{x_1, \ldots, x_n\}$ from a Uniform$(0, 2)$ distribution.

(b) The function $f(x) = \exp(-x)$ takes higher values when x is near zero than when x is near two, for $x \in (0, 2)$. Since this is where the integrand is fluctuating fast, we might hope simulations from it would enable us to create a more accurate Monte Carlo estimate of the integral. However, $f(x)$ is not a valid probability density function (pdf) for $x \in (0, 2)$.

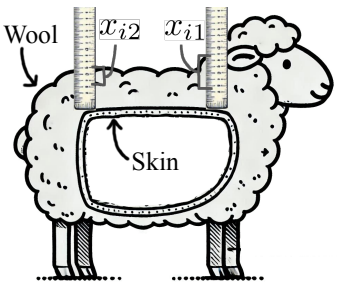
Find the normalising constant which makes $f(x)$ a valid pdf on $(0, 2)$. Hence, express the integral above as an expectation with respect to this pdf and write down the Monte Carlo estimator of the integral given simulations $\{x_1, \ldots, x_n\}$ from this pdf.

(c) You do a pilot simulation and find the terms in the Monte Carlo estimator for part (a) have variance 0.5848, whilst for part (b) they have variance 0.3787. Determine how many simulations would be needed to achieve an accuracy of ± 0.001 with 95% confidence in each case.

SECTION B

Q3 A shepherd struggles with precision while shearing his flock, leaving Shaun the Sheep frustrated by the uneven styling of his fleece. So Shaun invented the Shear-o-Matic to achieve a more consistent cut. He defines consistency as the removal of the same amount of wool across the entire body.

To assess shearing consistency, he used a ruler to measure the amount of wool removed (in cm) at the hips and shoulders of each sheep, as illustrated in the cross-section diagram. He would like to test:



H_0 : The Shear-o-Matic has similar consistency to when the shepherd cuts

vs H_1 : The Shear-o-Matic is *more* consistent in the amount of wool removed

Shaun will test this using the average across sheep of the larger minus smaller measurement ('average range'), with a smaller average range indicating more consistent shearing.

(a) (i) Write down the test statistic, T, for n sheep, using x_{i1}, x_{i2} to denote amount of wool removed for sheep i at shoulder and hip respectively.

(ii) Compute the observed test statistic, t_{obs}, for the $n = 3$ measurements:

$$\text{Shaun} : x_{11} = 1.49, \quad x_{12} = 1.87 \qquad \text{Timmy} : x_{21} = 0.57, \quad x_{22} = 0.81 \qquad \text{Shirley} : x_{31} = 1.48, \quad x_{32} = 0.94$$

The amount of wool removed by the Shear-o-Matic at both hip and shoulder are (unrealistically!) modelled as iid Exponential(λ), $f_X(x \mid \lambda) = \lambda e^{-\lambda x}, \mathbb{E}[X] = \lambda^{-1}$.

(b) (i) The shepherd averages a range of wool removed between hip and shoulder of 1.5cm. Find the parameter value λ_0 that corresponds to the null hypothesis to be tested. You may use the fact that the range between two iid Exponential(λ) random variables also has distribution Exponential(λ).

(ii) The null hypothesis is $H_0 : \lambda = \lambda_0$. What is the alternative hypothesis in this scenario? $H_1 : \lambda < \lambda_0$, $H_1 : \lambda \neq \lambda_0$, or $H_1 : \lambda > \lambda_0$

(c) 100 simulations of the test statistic under the null (ordered, 10 per line) follow:

0.21, 0.34, 0.38, 0.39, 0.43, 0.47, 0.51, 0.51, 0.51, 0.52,
0.54, 0.56, 0.57, 0.58, 0.61, 0.62, 0.62, 0.63, 0.67, 0.68,
$\cdots$ 60 other simulations $\cdots$
2.09, 2.09, 2.12, 2.16, 2.16, 2.24, 2.26, 2.27, 2.38, 2.4,
2.47, 2.48, 2.56, 2.73, 2.75, 3.17, 3.48, 3.68, 3.77, 3.78

Estimate the p-value based on this (small) Monte Carlo simulation. Is there evidence the Shear-o-Matic gives a more consistent cut?

(d) The shepherd cares about removing a lot of wool, because he needs to sell it. Before switching, he wants Shaun to demonstrate the Shear-o-Matic removes *more* wool overall. He proposes a new test statistic $T_2 = \frac{1}{n}\sum_{i=1}^{n} x_{i1} + x_{i2}$.

The sheep dog broke his computer, so he cannot do more simulations. Compute the new observed test statistic t_{obs2} and, noticing that $T_2 \geq T$, show one can reuse the (incorrect) simulations above to find a lower bound for the p-value of this new test. Can you draw any firm conclusions?

Q4 The probability density function (pdf) of a Normal$(1, 1)$ truncated to $[0, \pi]$ is:

$$f(x) = k \exp\left(-\frac{(x-1)^2}{2}\right) \quad \text{if } x \in [0, \pi]$$

and $f(x) = 0$ otherwise, where k is an unknown normalising constant. This pdf does not have a simple closed form cumulative distribution function (cdf).

(a) The following function is a valid pdf:

$$\tilde{f}(x) = \begin{cases} \frac{4\pi - 2x}{3\pi^2} & \text{if } x \in [0, \pi] \\ 0 & \text{otherwise} \end{cases}$$

If 0.06 and 0.77 are two values simulated from a Uniform(0,1) distribution, use them to produce two simulations from $\tilde{f}(x)$ by inverse transform sampling.

(b) (i) Prove that $\tilde{f}(x)$ can be used as the proposal in a rejection sampler to produce samples from $f(x)$ without knowing k, and find the value of the constant c required for the rejection sampling algorithm.

(ii) In this case, is the constant c the expected number of iterations needed for rejection sampling to produce a sample from $f(x)$? Why or why not?

Your computer has a malfunction that slows down calculating exponentials, making it very slow to calculate $f(x)$. A friend thinks up a clever 'squeezing' tweak to rejection sampling to speed up simulation from $f(x)$. The idea only relies on having an easy to calculate function $h(x)$ such that $0 \leq h(x) \leq f(x)\ \forall\, x \in [0, \pi]$. Their algorithm is as follows, with c having the value you already found in (b)(i):

1. Simulate $u_1 \sim \text{Unif}(0, 1)$ and simulate $x \sim \tilde{f}(\cdot)$ from the proposal.
2. If $u_1 \leq \dfrac{h(x)}{c\tilde{f}(x)}$ then accept the simulation x, else continue to 3.
3. Simulate a new $u_2 \sim \text{Unif}(0, 1)$.
4. If $u_2 \leq \dfrac{f(x) - h(x)}{c\tilde{f}(x) - h(x)}$ then accept the simulation x, else reject x.

(c) Given a proposal x, prove the probability x is accepted in step 2 or 4 is $f(x)/(c\tilde{f}(x))$ and deduce that any x accepted by this algorithm has pdf $f(\cdot)$.
Hint: "deduce" by following the same logic as for proving samples from standard rejection sampling have pdf $f(\cdot)$, using the probability calculated.

(d) Prove the probability that step 4 must be checked is:

$$1 - \frac{1}{c}\int_0^\pi h(x)\, dx$$

(e) The following cubic function satisfies the requirements for $h(\cdot)$:

$$h(x) = 0.11x^3 - 0.7x^2 + x + 0.45$$

What is the probability that you do *not* have to calculate an exponential?
Assume that calculating $f(\cdot)$ takes 5 seconds, calculating $\tilde{f}(\cdot)$ and $h(\cdot)$ each take 0.1 seconds, and all other operations/simulation are effectively instant. What is the expected time *saving* per proposal evaluated using this squeezing method versus standard rejection sampling?