Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2022	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.

Revision:	

SECTION A

Q1 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. We are interested in estimating some real-valued parameter θ using a statistic $S(\cdot)$.

1.1 Write down the estimator $\hat{\theta}$ and give detailed steps for how to estimate its variance via the Bootstrap.

1.2 The parameter of interest, θ, is the probability that X is zero. Write down the equation for a statistic, $S(\cdot)$, to estimate this parameter.

You collect data:

$$x_1 = 0, x_2 = 4, x_3 = 1$$

Assuming you perform Bootstrap resampling, write down all possible values for $S(\cdot)$ and the probability of observing each one.

1.3 Find $\mathbb{E}[\bar{S}^\star]$ without performing any simulation and estimate the implied bias in using this statistic to estimate θ.

Q2 **2.1** State the inverse sampling algorithm to simulate a random variable X having cumulative distribution function (cdf) $F(x)$. Take care as part of your answer to precisely define the generalised inverse cdf.

2.2 Prove that samples generated by inverse sampling have the required cdf, $F(x)$.

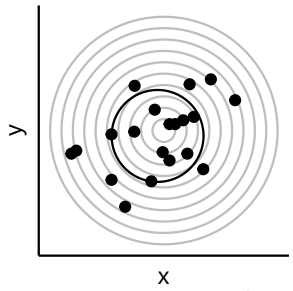
2.3 Let the random variable X have probability density function,

$$f(x) = \begin{cases} \frac{\pi}{2}\sin(\pi x) & x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

and suppose that 0.42 and 0.89 are two values simulated from a Uniform(0,1). Use these Uniform simulations to produce simulations of X using the inverse sampling algorithm.

SECTION B

Q3 In the science of ballistics, the Circular Error Probability (CEP) is used to evaluate the precision of firearms. It is defined to be the radius of a circle, centred on the mean, which is expected to include 50% of bullets fired. For example:



By making simplifying assumptions, we can directly model the random radial distance of each shot, R, from the mean as Rayleigh distributed with probability density function:

$$f(r) = \begin{cases} \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) & \text{when } r \in [0, \infty) \\ 0 & \text{otherwise} \end{cases}$$

The Department of Defence needs to test if the manufacturer's claim of a CEP of 0.5 is plausible given 20 shots at a test target. Those 20 shots resulted in an empirical CEP radius of 0.64.

They ask you to perform the test:

$$H_0 : \text{CEP} = 0.5 \text{ versus } H_1 : \text{CEP} > 0.5$$

3.1 Find the cdf of the Rayleigh distribution. Show that the median of the Rayleigh distribution is $\sigma\sqrt{2\log 2}$. Find the value of σ which corresponds to the null hypothesis to be tested.

3.2 In order to perform the above test, your colleague suggests using the test statistic

$$T = h(R_1, \ldots, R_{20}) = \text{median}(R_1, \ldots, R_{20}) - 0.5$$

Would this be ok and why (or why not)?

3.3 Describe in detail how to conduct a Monte Carlo hypothesis test in this particular setting (eg state what is simulated etc).

3.4 Following the steps you provided, your colleague has produced 100 simulations of the test statistic, ordered and shown as follows:

$$\begin{array}{ccccc} -0.159, & -0.145, & -0.136, & -0.133, & -0.127, \\ -0.120, & -0.115, & -0.115, & -0.111, & -0.110, \\ & & \ldots \text{ 80 other values } \ldots & & \\ 0.115, & 0.117, & 0.120, & 0.129, & 0.130, \\ 0.134, & 0.152, & 0.164, & 0.168, & 0.179. \end{array}$$

Estimate the p-value based on this (small) Monte Carlo simulation. What would you tell the Department of Defence?

3.5 Define the resampling risk (do not try to actually calculate its value).

Q4 A random variable, X, following the semicircle distribution has probability density function:

$$f(x) = \begin{cases} \frac{2}{\pi}\sqrt{1-x^2} & \text{when } x \in (-1,1) \\ 0 & \text{otherwise} \end{cases}$$

You may assume the knowledge that the semicircle distribution has:

$$\mathbb{E}[X] = 0, \text{Var}(X) = \frac{1}{4} \text{ and } \mathbb{E}[X^4] = \frac{1}{8}$$

We are interested in the behaviour of an estimator $\hat{\mu}_n$ of the parameter $\mu = \mathbb{E}[X^2]$ found via Monte Carlo.

4.1 (i) Assume you have n Monte Carlo simulations $\{x_1, \ldots, x_n\}$ from the semicircle distribution. Write down the equation for the Monte Carlo estimator, $\hat{\mu}_n$, as well as the variance of $\hat{\mu}_n$ as a function of n.

(ii) How large should n be so that a 95% confidence interval around $\hat{\mu}_n$ has size ± 0.1?

4.2 The estimator $\hat{\mu}_n$ requires simulations from the semicircle distribution.

(i) Show that the Uniform distribution on the interval $[-1, 1]$ can be used as a proposal distribution in a rejection sampler to produce samples from the semicircle distribution.

(ii) Write down in full detail the rejection sampling algorithm to produce simulations of X in this setting, and demonstrate the algorithm by using the following simulations (in the order provided):

$$\begin{aligned} -0.33, \quad 0.82, -0.42 &\text{ are simulations from a Uniform}(-1,1) \\ 0.75, \quad 0.61, \quad 0.01 &\text{ are simulations from a Uniform}(0,1) \end{aligned}$$

4.3 Assume the most computationally costly part is generation of Uniform simulations.

Your friend notices that you could just use the Uniform$(-1,1)$ proposals directly in an importance sampler to estimate μ without needing a second Uniform to test for acceptance/rejection.

You run the importance sampler using 1000 Uniform$(-1,1)$ simulations and calculate the weights $\{w_i\}_{i=1}^{1000}$. You find,

$$\sum_{i=1}^{1000} w_i = 1000.9 \text{ and } \sum_{i=1}^{1000} w_i^2 = 1082.4$$

By finding the effective sample size, calculate how many more Uniform simulations (**in total**) a rejection sampler would need in order to achieve approximately the same accuracy as this importance sampling estimate.