Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2024	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. Students must use the mathematics specific answer book.

Revision:	

ED01/2024

SECTION A

Q1 **1.1** State the inverse transform 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.

1.2 Let the random variable X have probability density function (pdf):

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

Simulate two values from this pdf via inverse transform sampling using the following values simulated from the Uniform$(0, 1)$ distribution:

$$0.33, \quad 0.71$$

1.3 Evaluating the following integral usually requires the Leibniz integral rule:

$$\int_0^1 \frac{x - 1}{\log x} \, dx$$

Instead, express this integral as an expectation with respect to the pdf $f(x)$ in question **1.2**, above, and write down the Monte Carlo estimator of this integral given simulations $\{x_1, \ldots, x_n\}$ from $f(x)$.

Q2 Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a sample of n independent observations sampled uniformly at random (without replacement) from a finite population of size N. We are interested in estimating some real-valued parameter θ using a statistic $S(\cdot)$.

2.1 Write down the estimator $\hat{\theta}$ and give full details of the *population bootstrap* method for estimating its variance in the finite population setting. Be sure to cover the case where n does not divide N.

2.2 Imagine the parameter we wish to estimate, θ, is the mean (so $S(\cdot)$ is the sample mean). Prove that the variance of the sample mean in the finite population setting described above is:

$$\text{Var}(\bar{X}) = \left(\frac{N - n}{N - 1} \right) \frac{\sigma^2}{n}$$

where σ^2 is the population variance.

2.3 You collect 2 observations from a population of size $N = 4$:

$$x_1 = 1, x_2 = 0$$

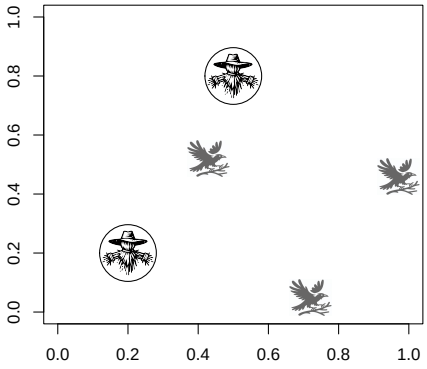
(i) We do not know the population variance. Write down the formula giving an unbiased estimator for the variance of the sample mean, based on the sample variance, and calculate the estimate for this situation.

(ii) Your colleague does not know the result you used in the last part and estimates the variance of the sample mean using *population bootstrap*. Without performing any simulation, determine all possible values for $S(\cdot)$ and the probability of observing each one.

SECTION B

Q3 Wallace has invented a new scarecrow, called the "Scaratronic", which he claims deters birds from encroaching on farmers' crops. However, an initial analysis showed no significant difference in the *number* of birds landing in fields with or without Scaratronic scarecrows.

Wallace now claims that although just as many birds may visit the field, they are more likely to avoid the areas near the Scaratronic devices. This could limit crop damage to smaller sections of the field. A local farming cooperative has asked you to conduct an analysis of Wallace's new claim, by testing the hypothesis:



$$\begin{aligned} H_0 &: \text{birds land uniformly at random} \\ \text{versus } H_1 &: \text{birds tend to land further away} \\ &\quad \text{from Scaratronic devices than if} \\ &\quad \text{landing uniformly randomly} \end{aligned}$$

The figure above shows a 1km×1km square field, the location of 2 Scaratronic devices (in the circles), and the location of 3 bird sightings.

$$\begin{aligned} \text{Scaratronic coordinates}: &\quad \{(x_{si}, y_{si})\}_{i=1}^{2} = \{(0.20, 0.20), (0.50, 0.80)\} \\ \text{Bird coordinates}: &\quad \{(x_{bj}, y_{bj})\}_{j=1}^{3} = \{(0.43, 0.52), (0.72, 0.05), (0.97, 0.46)\} \end{aligned}$$

3.1 (a) Your colleague suggests the test statistic should be the sum of the distances of all bird locations to their nearest Scaratronic device. Would this be ok and why (or why not)?

(b) Write the test statistic mathematically.

(c) Compute the observed test statistic value, t_{obs}.

3.2 Describe in full detail how to conduct a Monte Carlo hypothesis test in this particular setting (ie explain step-by-step what is simulated, calculated, etc in detail).

3.3 Following the steps you provided, your colleague has produced 100 simulations of the test statistic, ordered and shown (10 per line) as follows:

0.37, 0.41, 0.43, 0.45, 0.45, 0.45, 0.49, 0.57, 0.60, 0.61,
0.61, 0.63, 0.64, 0.65, 0.66, 0.66, 0.68, 0.70, 0.70, 0.72,
$\cdots$ 60 other simulations $\cdots$
1.21, 1.22, 1.23, 1.23, 1.24, 1.24, 1.25, 1.26, 1.29, 1.29,
1.36, 1.48, 1.51, 1.54, 1.58, 1.59, 1.61, 1.64, 1.65, 1.75

Estimate the p-value based on this (small) Monte Carlo simulation. What would you tell the local farming cooperative?

3.4 Define the resampling risk (do <u>not</u> try to actually calculate its value).

3.5 We cannot compute the resampling risk here since we do not know the exact sampling distribution of the test statistic. Instead, write the p-value as an integral and so compute a 95% confidence interval for the Monte Carlo estimator of the p-value. Is the Monte Carlo simulation done by your colleague sufficiently large to make a decision at the $\alpha = 5\%$ level of significance?

Q4 We saw in tutorials that it is easy to inverse transform sample random variables, X, having probability density function (pdf):

$$f(x \mid \alpha) = \begin{cases} \alpha x^{\alpha-1} & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

for any $\alpha > 0$.

4.1 Your friend didn't do the tutorial and so immediately tried rejection sampling instead. They considered using a Uniform(0,1) distribution as proposal. For what values of α will this work?

4.2 Consider only the set of values for which the Uniform(0,1) distribution is a valid proposal. As a function of α, what is the probability that any single iteration of the rejection sampler will be accepted?

For the rest of this question we are interested in the behaviour of an estimator $\hat{\mu}_n$ of the parameter $\mu := \mathbb{E}\left[-\sqrt{X}\right]$ when $\alpha = 2$.

4.3 Assume you have n Monte Carlo simulations $\{x_1, \ldots, x_n\}$ from the above pdf with $\alpha = 2$. Write down the equation for the Monte Carlo estimator, $\hat{\mu}_n$, and calculate the variance of $\hat{\mu}_n$ as a function of n.

4.4 Your friend suggests that it would be more efficient to use an importance sampler to compute $\hat{\mu}_n$, using Monte Carlo simulations from a proposal pdf with the same form, but with a different parameter value α, say $\alpha = a \neq 2$. Calculate (in terms of a) the variance of such an importance sampling estimator for $\hat{\mu}_n$, where the target still has $\alpha = 2$, and state the valid range of values for a.

4.5 For what range of values of a does the importance sampling estimator have lower variance than the standard Monte Carlo estimator?