MATH 70098 Ethics in DS and AI: Part I
Checkpoint 1 Questions

Zak Varty

QUESTION SHEET

This checkpoint question sheet is for you to assess your own progress through Ethics Part 1 and to identify any areas that need further clarifications. It is recommended that you make a written attempt at all questions before solutions are made available to you.

This activity is expected to take approximately **120 minutes** of effort. The available marks for each question are indicated in square brackets, with a total of **32 marks** available.

1   Toy Shop Anonymity [6]

A toy shop is trying to compare spending in two different locations and collects the spending information on its customers.

a) What is the maximum value of k for which the purchase values are k-anonymised in Table 1? Give a short justification for your answer. [2]

town	spend
A	29.99
A	17.11
A	33.51
A	00.10
A	10.00
A	07.45
A	21.99
A	32.50
A	20.00
B	45.99
B	22.11
B	04.99
B	00.25

Table 1: Scenario A

age	town	spend
12	A	29.99
11	A	17.11
10	A	33.51
11	A	00.10
12	A	10.00
10	A	07.45
10	A	21.99
12	A	32.50
11	A	20.00
11	B	45.99
11	B	22.11
11	B	04.99
11	B	00.25

Table 2: Scenario B

gender	age	town	spend
F	12	A	29.99
F	11	A	17.11
F	10	A	33.51
F	11	A	00.10
F	12	A	10.00
M	10	A	07.45
M	10	A	21.99
F	12	A	32.50
F	11	A	20.00
M	11	B	45.99
M	11	B	22.11
F	11	B	04.99
F	11	B	00.25

Table 3: Scenario C

b) The shop owner wants to account for the age of her customers in her comparison. What is the maximum value of k for which the purchase values are k-anonymised in the augmented Table 2? Give a short justification for your answer. [2]

c) She also suspects that there may be a gender gap in pocket money. When gender is included, what is the maximum value of k for which the purchase values are k-anonymised in the augmented table 3? Give a short justification for your answer. [2]

2   Estimating prevalence of study drug use [16]

A survey is designed to estimate the proportion, p, of students engaging in the use of performance enhancing drugs during exam season. To protect individual respondents a randomised response survey

design is used, where each student is asked to answer the question "Have you taken performance enhancing drugs during this exam season? (Yes/No)".

A random number generator is used to decide whether each student responds to the question directly or gives a random response. With probability q the student answers directly. If the student answers randomly then they reply "Yes" with probability q and "No" with probability $1 - q$.

a) Let Y be the event of responding 'Yes", D be the event of having taken performance enhancing drugs and R be the event of responding randomly. Draw a probability tree to describe this randomised response survey. [3]

b) What is the probability that a respondent replies "Yes" in this survey? [3]

c) Janine replied "Yes" in the survey. What is the probability that she had not taken performance enhancing drugs? [2]

d) By equating the sample and population proportions of respondents answering "Yes", derive the method of moments estimator $\hat{P}$ for p based on n responses to this survey design. In your derivation, denote the survey responses $Y_1, Y_2, \ldots, Y_n$ where $Y_i = 1$ if the student answered "Yes" and $Y_i = 0$ if the student answered "No". Let the mean of these responses be $\bar{Y}$. [2]

e) Explain why values of 0 and 1 for q would not be suitable in this survey design. [2]

f) For which values of q does the method of moments estimate $\hat{p}$ yield a valid probability? Why is it challenging to select q to satisfy this condition? [2]

g) Of the 150 students surveyed using the randomised response survey design using q = 0.4, 99 students responded "No"'. Calculate a point estimate for the proportion of students who have taken performance enhancing drugs this exam season. [2]

3   Ethical AI in Social Media [10]

A social media company collects information about its users and their browsing history. The company use this information to construct and deploy an automated system that suggests future content and advertisements to each user based on their characteristics and previous activity.

In this context, briefly explain each of the five principles of ethical AI as given in "A unified framework of five principles for AI in society" (Floridi and Cowls 2019) and give an example of what each could mean in practical terms. [10]