1. A lake contains three different types of carp.

There are an estimated 450 mirror carp, 300 leather carp and 850 common carp.

Tim wishes to investigate the health of the fish in the lake.

He decides to take a sample of 160 fish.

(a) Give a reason why stratified random sampling cannot be used.

(1)

(b) Explain how a sample of size 160 could be taken to ensure that the estimated populations of each type of carp are fairly represented.

You should state the name of the sampling method used.

(2)

As part of the health check, Tim weighed the fish.

His results are given in the table below.

Weight (w kg)	Frequency (f)	Midpoint (m kg)
$2 \leqslant w < 3.5$	8	2.75
$3.5 \leqslant w < 4$	32	3.75
$4 \leqslant w < 4.5$	64	4.25
$4.5 \leqslant w < 5$	40	4.75
$5 \leqslant w < 6$	16	5.5

(You may use $\sum fm = 692$ and $\sum fm^2 = 3053$)

(c) Calculate an estimate for the standard deviation of the weight of the carp.

(2)

Tim realised that he had transposed the figures for 2 of the weights of the fish.

He had recorded in the table 2.3 instead of 3.2 and 4.6 instead of 6.4

(d) Without calculating a new estimate for the standard deviation, state what effect

(i) using the correct figure of 3.2 instead of 2.3

(ii) using the correct figure of 6.4 instead of 4.6

would have on your estimated standard deviation.

Give a reason for each of your answers.

(2)

a) it isn't possible to have a sampling frame

(we don't know the exact number of carp in the lake &

Question 1 continued

don't have a list of all of them)

b) use quota sampling.

estimated total: 1600 → 10x sample size

divide estimated numbers of each type by 10:

catch 85 common, 45 mirror, & 30 leather carp.

ignore any fish caught once the quota for that type is full.

c) $\sigma = \sqrt{\frac{\Sigma fm^2}{n} - \left(\frac{\Sigma fm}{n}\right)^2}$

$= \sqrt{\frac{3053}{160} - \left(\frac{692}{160}\right)^2}$

$= 0.6129...$

we use the frequency in each class.

d) i. the data point stays in the same class, so this would not change the standard deviation.

ii. 116 ~~this~~ is outside the available classes, so does change the mean by a small amount. $6.4 - 4.6 = 1.8 \simeq 3\sigma$ so the estimate of σ will increase.

$\frac{692}{160} = 4.3...$ so 4.6 is close to the mean, 6.4 is far from it

↗ cardinal direction

2. Helen is studying one of the qualitative variables from the large data set for Heathrow from 2015.

not autumn or winter ←

↘ daily mean wind speed

She started with the data from 3rd May and then took every 10th reading.

There were only 3 different outcomes with the following frequencies

↘ outcomes are:
- light
- moderate
- fresh

Outcome	*A*	*B*	*C*
Frequency	16	2	1

↘ lots of "light" daily mean wind speed

(a) State the sampling technique Helen used.

(1)

(b) From your knowledge of the large data set

(i) suggest which variable was being studied,

(ii) state the name of outcome *A*.

(2)

George is also studying the same variable from the large data set for Heathrow from 2015. He started with the data from 5th May and then took every 10th reading and obtained the following

Outcome	*A*	*B*	*C*
Frequency	16	1	1

Helen and George decided they should examine all of the data for this variable for Heathrow from 2015 and obtained the following

Outcome	*A*	*B*	*C*
Frequency	155	26	3

(c) State what inference Helen and George could reliably make from their original samples about the outcomes of this variable at Heathrow, for the period covered by the large data set in 2015.

(1)

(a) Systematic sampling (1)
(she took every 10th reading, meaning she chose her sample at a regular interval)

(b)(i) daily mean wind speed (1)

(ii) light (1)

Question continued

(c) Occurence of Variable A = $\frac{155}{155+26+3} \times 100 = 84.23\ \%$

variable A occurs most which is around 84% of the time. (1)

(Total for Question ´ is 4 marks)

3. Charlie is studying the time it takes members of his company to travel to the office. He stands by the door to the office from 08 40 to 08 50 one morning and asks workers, as they arrive, how long their journey was.

(a) State the sampling method Charlie used.

(1)

(b) State and briefly describe an alternative method of non-random sampling Charlie could have used to obtain a sample of 40 workers.

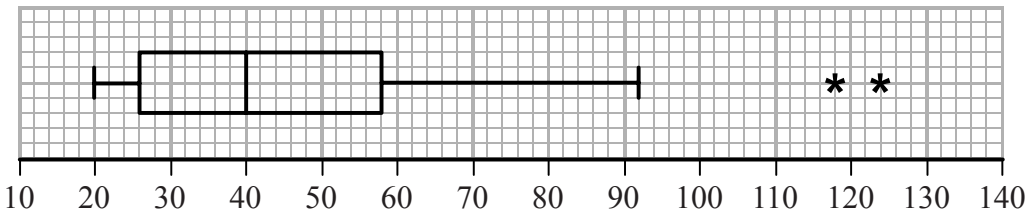
(2)

Taruni decided to ask every member of the company the time, x minutes, it takes them to travel to the office.

(c) State the data selection process Taruni used.

(1)

Taruni's results are summarised by the box plot and summary statistics below.



$$n = 95 \qquad \sum x = 4133 \qquad \sum x^2 = 202\,294$$

(d) Write down the interquartile range for these data.

(1)

(e) Calculate the mean and the standard deviation for these data.

(3)

(f) State, giving a reason, whether you would recommend using the mean and standard deviation or the median and interquartile range to describe these data.

(2)

Rana and David both work for the company and have both moved house since Taruni collected her data.

Rana's journey to work has changed from 75 minutes to 35 minutes and David's journey to work has changed from 60 minutes to 33 minutes.

Taruni drew her box plot again and only had to change two values.

(g) Explain which two values Taruni must have changed and whether each of these values has increased or decreased.

(3)

a) Opportunity sampling. - (1) (or convenience sampling)

b) Quota sampling - Charlie could ask 20 men and 20 women how long their journey was. - (1)

↳ Could've also given 'take 4 people every 10 minutes' for the 2nd mark as another example. (1)

c) A census - (1)

d) IQR = UQ - LQ

IQR = 58 - 26

= 32 - (1)

e) $\bar{x} = \frac{\Sigma x}{n}$

$\bar{x} = \frac{4133}{95}$

$\bar{x} = 43.5$ (3s.f.) - (1)

$\sigma_x = \sqrt{\left(\frac{\Sigma x^2}{n}\right) - \left(\frac{\Sigma x}{n}\right)^2}$

$\sigma_x = \sqrt{\frac{202294}{95} - \left(\frac{4133}{95}\right)^2}$ - (1)

$\sigma_x = 15.4$ (3s.f.) - (1)

f) Due to outliers in the data, the median and interquartile range should be used - as outliers affect the mean and standard deviation. - (2)

g) The median and upper quartile will change, (1) as there are now more values less than 40 (median) (1) and so the median will decrease with this, as will the upper quartile - (1)

→ For first mark, could have said that the value at 20, the lower quartile at 26, and the outliers will not change.

4. (a) State one disadvantage of using quota sampling compared with simple random sampling.

(1)

In a university 8% of students are members of the university dance club. so $p = 0.08$

A random sample of 36 students is taken from the university. and $n = 36$

The random variable X represents the number of these students who are members of the dance club.

(b) Using a suitable model for X, find

(i) $P(X = 4)$

(ii) $P(X \geqslant 7)$

(3)

Only 40% of the university dance club members can dance the tango.

(c) Find the probability that a student is a member of the university dance club and can dance the tango.

use multiplication rule for AND

(1)

A random sample of 50 students is taken from the university.

$P(X < 3)$

(d) Find the probability that fewer than 3 of these students are members of the university dance club and can dance the tango.

(2)

a) One disadvantage is that quota sampling is not random, so it cannot be used reliably for inferences. ①

OR

more likely to be biased.

OR

Not random / less random

b) i) $X \sim B(36, 0.08)$ ① where $n = 36$, $p = 0.08$

$P(X = 4) = 0.16738... = 0.167$ (3.s.f) ①

ii) $P(X \geqslant 7) = 1 - P(X \leqslant 6)$

$= 1 - 0.97776..$

$= 0.022233..$

$= 0.0222$ (3 s.f) ①

c) $P(\text{dance club AND tango}) = 0.08 \times 0.4$

$= 0.032$ ①

or 3.2 % or $\frac{4}{125}$

Question continued.

d) T = people who can dance the tango

$T \sim B(50, 0.032)$ (1) ← 0.032 from c)

$P(T < 3) = P(T \leq 2) = 0.785$ (3.s.f) (1)

(Total for Question is 7 marks)