

Populations and Samples

- 1 A headteacher wants to find out how the 1200 students in her school travel to school. She asks 60 students, chosen from the school register, how they travel.
State the population and the sample in this survey. [2]
- 2 State one advantage and one disadvantage of using a census rather than a sample to collect data from a population. [2]
- 3 State whether each of the following is a parameter or a statistic.
(a) The mean mass of all 480 eggs laid on a farm in one week.
(b) The mean mass of a random sample of 20 of these eggs.
(c) The proportion of all the voters in a town who support a proposal.
(d) The proportion of a sample of 500 of these voters who support it. [4]
- 4 For each investigation, state whether a census or a sample is more suitable, giving a reason.
(a) Finding the number of students in a class of 24 who have a pet.
(b) Finding the mean time before a type of battery goes flat, for a batch of 10 000 batteries. [2]
- 5 In a random sample of 60 students from a school of 1500 students, 15 say that they walk to school.
Estimate the number of students at the school who walk to school. [2]
- 6 Explain what is meant by a sampling frame and give an example of a suitable sampling frame for a survey of the members of a library. [2]
- 7 A factory makes 8000 light bulbs each day. To check quality, the manager switches on bulbs and records how many hours each lasts before it fails.
(a) Give a reason why the manager should use a sample rather than a census. [2]
(b) State what the sampling unit is in this investigation. [1]
- 8 A council wants the views of the 40 000 residents of a town. It uses the electoral register as its list of people to choose from.
(a) Write down the name for a list of this type. [1]
(b) Give one reason why the electoral register may not be a perfect list for this survey. [1]
(c) The mean income of the residents in the sample is $\bar{x}$. The mean income of all residents is μ . State which of these is a statistic and which is a parameter. [1]
(d) Explain why a second sample from the same register would probably give a different value of $\bar{x}$. [1]
- 9 The yolk masses (in grams) of a random sample of 8 eggs from a farm are 12.1, 11.8, 12.6, 12.0, 11.5, 12.4, 12.2 and 11.9.
Calculate the sample mean and state what it estimates. [2]
- 10 Each interview in a survey takes 12 minutes. There are 2400 people in the population.
(a) Find, in hours, the time needed to interview everyone. [2]
(b) Find the time needed to interview a sample of 150 people. [1]
- 11 A manufacturer claims that the mean mass of a bag of sugar is 1000 g. A random sample of 25 bags has a mean mass of 996 g.
(a) State whether 996 g is a statistic or a parameter. [1]
(b) Explain why this sample alone does not prove that the manufacturer's claim is false. [1]
- 12 80 deer in a forest are caught, tagged and released. Later, 50 deer are caught and 8 of them are tagged.
Estimate the number of deer in the forest and state one assumption you have made. [3]

- 13** A population consists of the four values 4, 7, 9 and 12. Find the population mean and the population variance. [3]

- 14** A small population consists of the five values 2, 3, 5, 8 and 12. Samples of size 2 are taken without replacement and the mean of each sample is found.
- (a) Find the population mean μ . [2]
 - (b) List all possible samples of size 2 and the mean of each. [2]
 - (c) A sample is chosen at random. Find the probability that its mean is within 1 of μ . [2]
 - (d) Comment on what your list in part (b) shows about sample means. [1]

- 15** Five random samples of size 10 and five random samples of size 100 were taken from the same population. The sample means are shown in the table.

Sample	1	2	3	4	5
Mean, $n = 10$	48	55	41	60	46
Mean, $n = 100$	51.2	50.4	52.1	49.6	50.9

- (a) Find the range of the sample means for each sample size. [2]
 - (b) State, with a reason, which sample size gives more reliable information about the population mean. [2]
 - (c) Use all the data for the larger samples to give an estimate of the population mean. [1]
- 16** A population consists of the values 1, 2 and 3, each equally likely to be chosen. A random sample of size 2 is taken with replacement and the sample mean $\bar{X}$ is calculated.
- (a) List the nine possible samples and find the probability distribution of $\bar{X}$. [3]
 - (b) Show that the mean of the distribution of $\bar{X}$ equals the population mean. [2]

- 17** Two numbers are chosen at random, without replacement, from the integers 1, 2, 3, 4, 5. Let M be the larger of the two numbers.
- (a) Find the probability distribution of M . [3]
 - (b) Find $E(M)$ and explain why M is a poor estimate of the mean of the population. [3]

- 18** A firm has 600 invoices. A random sample of 40 invoices has a total value of £5130.
- (a) Estimate the mean value of an invoice. [1]
 - (b) Estimate the total value of all 600 invoices. [2]
 - (c) Give one reason why the estimate in part (b) might be unreliable. [1]

- 19** A statistician takes many samples of each size from a population and records the range of the sample means. The results are shown.

Sample size	Range of sample means
5	18.4
20	6.1
80	2.4

- (a) Describe the relationship between the sample size and the range of the sample means. [1]
 - (b) Interviewing costs £4 per person plus a fixed £250. A budget of £500 is available. Find the largest sample size that can be afforded, and state its effect on reliability. [3]
- 20** A school has 2400 students. A sample of 5% of the students is taken, with equal numbers from each of the 6 year groups.
- (a) Find the size of the sample. [1]
 - (b) Find the number sampled from each year group. [1]
 - (c) Year 7 has 520 students and Year 11 has 380. Explain why equal numbers from each year may not give a representative sample. [2]

- 21** A sample X_1, X_2, X_3, X_4 is taken from a population with unknown mean μ and unknown variance σ^2 .
- (a) For each expression below, state whether or not it is a statistic, giving a reason.
- (i) $\frac{X_1 + X_2 + X_3 + X_4}{4}$ (ii) $X_1 - \mu$ (iii) the largest of X_1, X_2, X_3, X_4 (iv) $\frac{X_1^2 + X_2^2}{\sigma^2}$ [4]
- (b) Explain why the sample mean is a random variable but the population mean μ is not. [2]
-
- 22** A council wants to survey the 24 000 households in a town. A census costs £3.20 per household. A sample survey has a fixed cost of £1600 plus £3.20 for every household surveyed.
- (a) Find the cost of a census. [1]
- (b) The council wants the sample survey to cost exactly one third of the census. Find the sample size. [3]
- (c) In the sample, 4125 households support a new bus route. Estimate how many of the 24 000 households support it. [2]
- (d) Explain why the answer to part (c) is unlikely to be exactly correct. [1]
-
- 23** A group has 5 girls and 3 boys. A simple random sample of 3 people is chosen from the group.
- (a) Show that the probability that a particular person is in the sample is $\frac{3}{8}$. [3]
- (b) Find the probability that the sample contains more girls than boys. [3]
-
- 24** A group of 10 people is to be sampled. Four selections are made, and a particular person in the group is called Sam.
- (a) The four people are chosen as a simple random sample without replacement. Find the probability that Sam is in the sample. [2]
- (b) Instead, each selection is made at random from all 10 people, so that a person can be chosen more than once. Find the probability that Sam is chosen at least once. [3]
- (c) For the selections in part (b), find the probability that all four people chosen are different. [2]
- (d) Use your answer to part (c) to explain why sampling without replacement is normally used. [1]
-
- 25** A pollster asks 1000 randomly chosen voters whether they support a new law. In five separate polls of the same size, the percentages in favour are 48, 52, 51, 47 and 50.
- (a) Find the range of these percentages. [1]
- (b) Combine the five polls into a single poll and find the percentage in favour. [2]
- (c) A newspaper reports that "a majority of voters are in favour" after seeing the 52% poll. Comment on this claim. [2]
-
- 26** A population of 6 values 1, 2, 3, 4, 5, 6 is sampled by choosing 3 values without replacement, all samples being equally likely.
- (a) How many different samples are there? Find the probability that the sample mean is at least 4. [3]
- (b) Show that the mean of all the possible sample means is equal to the population mean. [2]
-
- 27** To estimate the number of fish in a lake, 150 fish are caught, tagged and returned. The next week a sample of 120 fish is caught and 18 of them carry tags.
- (a) Estimate the number of fish in the lake. [3]
- (b) State two assumptions made in your calculation. [2]
- (c) Some fish learn to take the bait and are more likely to be caught again. Explain the effect on the estimate in part (a). [2]
- (d) A third week, a further 30 fish are caught and 5 carry tags. Treating the two samples as one sample of 150 fish, find a new estimate of the number of fish. [2]

- 28** A student makes three statements about a survey of the 5000 residents of a village. For each statement say whether it is correct, and explain your answer.
- (a) "A sample of 10% of the residents will always be more representative than a sample of 5%." [3]
 - (b) "The sample mean is 52.3 minutes but the population mean is 50.0 minutes, so the sample must be biased." [3]
 - (c) "Two random samples of the same size from the same population must have the same mean." [2]
-
- 29** A population consists of the integers 1 to 6, each equally likely. Samples of size 3 are taken without replacement, all samples being equally likely. For each sample both the mean $\bar{X}$ and the median M are found.
- (a) Show that there are 20 possible samples and find the probability distribution of M . [4]
 - (b) Find $E(M)$ and $\text{Var}(M)$. [3]
 - (c) The variance of $\bar{X}$ is 0.583333. State, with a reason, which of $\bar{X}$ and M is the better statistic for estimating the population mean. [3]
-
- 30** Tanks are numbered 1, 2, ..., N . Five captured tanks have serial numbers 14, 52, 33, 71 and 40, and they are regarded as a random sample from the N tanks.
Two students estimate N .
- (a) Anna doubles the sample mean and subtracts 1. Find her estimate. [2]
 - (b) Ben uses the largest serial number m and the sample size n in the formula $m + \frac{m}{n} - 1$. Find his estimate. [2]
 - (c) Explain why the largest serial number on its own is not a good estimate of N , and why Ben's method is preferred to Anna's. [3]

Sampling Methods

- 1** A club has 80 members, listed in alphabetical order. Describe how to select a simple random sample of 5 members. [2]
-
- 2** A company has 600 employees on a numbered list. A systematic sample of 40 employees is required.
- (a) Find the interval between selected employees. [1]
 - (b) The first employee selected is number 7. Write down the next two numbers selected. [1]
 - (c) Find the number of the last employee selected. [1]
-
- 3** Name the sampling method described in each case.
- (a) Every 10th customer through a shop door is asked to complete a survey.
 - (b) The names of all 300 staff are put in a hat and 20 names are drawn out.
 - (c) A researcher interviews 30 men and 30 women, taking whoever is available in the street, until each quota is filled. [3]
-
- 4** A school has 500 girls and 300 boys. A stratified sample of 80 students is taken. Find the number of girls and the number of boys in the sample. [2]
-
- 5** Explain what is meant by stratified sampling, and state one situation in which it is better than simple random sampling. [2]
-
- 6** State one advantage and one disadvantage of quota sampling. [2]

- 7** A school has 1200 students. A stratified sample of 120 students is taken, using the year groups as strata. The numbers in each year group are shown.
Find how many students should be sampled from each year group.

Year	7	8	9	10	11
Students	270	250	240	230	210

[3]

- 8** An interviewer in a shopping centre is told to interview 50 men and 50 women, of whom at least 20 of each must be under 30.
- (a) Name the sampling method being used. [1]
- (b) Give one advantage of this method over simple random sampling. [1]
- (c) Give two disadvantages. [2]

- 9** A club has 60 members numbered 01 to 60. A simple random sample of 4 members is to be chosen using the following two-digit random numbers, read in order:
47 83 12 05 12 59 71 33 60 21 08
Write down the numbers of the members chosen, explaining how you dealt with unsuitable numbers. [3]

- 10** A student surveys the first 25 people she meets outside the school library about how long they spend studying each week.
- (a) Name this sampling method. [1]
- (b) Give one disadvantage of this method for estimating the time all students spend studying. [1]

- 11** A school has 180 students in Year 12 and 120 students in Year 13. A stratified sample of 30 students is required.
- (a) Find the number of students to be sampled from each year group. [2]
- (b) Describe how the 12 Year 13 students in the sample should be chosen. [2]

- 12** A quota sample of 100 people is to match a town where 48% of the adults are men, and 25% of all adults are aged under 30.
Find the number of men in the sample and the number of people under 30, and state one other factor that could be used to set quotas. [3]

- 13** A radio station asks its listeners to vote for their favourite song by texting a number.
Name the type of sampling used and explain why the result may not be representative of all the town's residents. [3]

- 14** For each situation, name the most suitable sampling method and give a reason for your choice.
- (a) A leisure centre has a numbered list of 2400 members and wants an unbiased sample of 120 with as little effort as possible.
- (b) A town has three age groups of very different sizes and the survey must represent each group in proportion.
- (c) A researcher has one afternoon to question 100 people in a street, with fixed numbers of each gender and age band. [6]

- 15** A company with 440 employees wants a stratified sample of 55 employees, using the four departments as strata.

Department	Admin	Sales	Production	Research
Employees	40	120	200	80

- (a) Calculate the number of employees to sample from each department. [3]
 (b) Explain how the employees in the Research department could be selected. [1]
 (c) Give two advantages of using a stratified sample rather than a simple random sample of 55 from all employees. [2]

- 16** A school has 400 students, of whom 20 are in the sixth form. A simple random sample of 10 students is taken.

- (a) Find the probability that no sixth-form student is in the sample. Give your answer to 3 significant figures. [3]
 (b) Explain how stratified sampling would avoid this problem. [2]

- 17** A shop is open seven days a week. The manager takes a systematic sample of the daily takings by choosing every 7th day.

Explain why this may give a biased estimate of the mean daily takings. [2]

- 18** A population has three strata of sizes 450, 350 and 200. A stratified sample of size n is required.

- (a) Find the numbers sampled from each stratum when $n = 40$. [2]
 (b) Find the numbers sampled from each stratum when $n = 25$ and explain what the difficulty is. [3]

- 19** A market researcher wants a sample of 200 shoppers from a town with 10 000 adults. The town has 5400 adults aged under 35, 3600 aged 35 to 59 and 1000 aged 60 or over.

- (a) Find the quota for each age group so that a quota sample of 200 matches the make-up of the town. [3]
 (b) Give two differences between a quota sample and a stratified sample. [2]
 (c) Explain why a quota sample could still be biased even though the quotas match the town. [2]

- 20** There are 340 employees at a company. Describe in detail how to use a table of random numbers to select a simple random sample of 12 employees. [4]

- 21** A list of 1000 customers is numbered 1 to 1000. A systematic sample of 25 customers is taken.

- (a) Describe how the sample would be selected. [2]
 (b) Customers 100 and 101 are next to each other on the list. If a simple random sample of 25 is taken instead, find the probability that both are chosen. [2]
 (c) Write down the probability that both are chosen in the systematic sample, and explain what this shows about the two methods. [2]
 (d) A machine puts an item with a fault on every 4th product. Products on a conveyor are inspected using a systematic sample of every 8th product. Explain why this could give a misleading result. [2]

- 22** A stratified sample of 90 people is taken from a town. The strata are three age groups A, B and C. Stratum B has 2000 members in the population. The sample sizes and the sample means of weekly spending (£) are shown.

Stratum	A	B	C
Sample size	18	40	32
Sample mean	24.5	31.2	28.6

- (a) Find the total population of the town. [2]
 (b) Find the number of people in A and in C. [2]
 (c) Calculate an estimate of the mean weekly spending of the whole town. [3]
 (d) Explain why your answer to part (c) is equal to the mean of all 90 sampled values. [1]

- 23** The table shows the numbers of staff in three departments of a company.
 A stratified sample of 50 staff is taken, first by department and then, within each department, by gender.

Department	Women	Men
Sales	60	40
Admin	90	60
Tech	30	120

- (a) Find the number of staff to be sampled from each department. [3]
 (b) Find how many women and men should be sampled from the Tech department. [2]
 (c) Explain why this sample cannot include exactly 12.5 staff from Sales, and suggest how a fair choice between 12 and 13 could be made. [2]

- 24** A list has 20 members numbered 1 to 20. A systematic sample of 5 members is taken with a random start from 1 to 4.

- (a) Show that every member has probability $\frac{1}{4}$ of being in the sample. [2]
 (b) Explain why members 1 and 2 cannot both be in the sample. [1]
 (c) A simple random sample of 5 members is taken instead. Find the probability that members 1 and 2 are both in the sample, and comment. [3]

- 25** A researcher wants the opinions of the 900 members of a golf club. She puts all the names in alphabetical order and picks every 10th name, starting from the 4th.

- (a) Show that this gives a sample of size 90. [1]
 (b) A member says that this is not a simple random sample. Explain why he is correct. [2]
 (c) Give one situation in which this method would be satisfactory and one in which it would not. [2]

- 26** A county has 12 schools. A researcher picks 3 schools at random and then picks 20 students at random from each chosen school. School A has 400 students and school B has 100 students.

- (a) Find the probability that a particular student at school A is in the sample. [2]
 (b) Find the same probability for a student at school B. [1]
 (c) Hence explain why the method does not give a simple random sample of all students, and suggest an improvement. [3]

- 27** A group of 10 people contains 6 women and 4 men. A sample of 5 people is to be chosen.

- (a) The sample is a simple random sample. Find the probability that it contains exactly 3 women. [3]
 (b) The sample is instead a stratified sample by gender. Write down the probability that it contains exactly 3 women and explain your answer. [2]
 (c) Show that any individual person has probability $\frac{1}{2}$ of being selected under each method. [3]
 (d) Use your answers to explain one benefit of stratified sampling here. [1]

- 28** A university has 3000 students in four faculties. A researcher wants a sample of 150 students to estimate the mean number of hours students spend studying each week. The faculty sizes are shown in the table. The table also gives the mean study hours found in the sample from each faculty.

Faculty	Science	Arts	Business	Law
Students	1200	900	600	300
Mean study hours	14.2	11.5	9.8	16.4

- (a) Find how many students should be taken from each faculty for a stratified sample. [2]
- (b) Describe how a systematic sample of the 60 Science students could be taken from the list of 1200 Science students. [2]
- (c) Another researcher interviews the first 150 students who leave the library. Give two reasons why this is unlikely to give a good estimate of the study hours. [2]
- (d) Use the stratified sample results to estimate the mean study hours for the whole university. [2]
- (e) A student suggests using the mean of the four faculty means, which is 12.975. Explain why this is not appropriate. [1]
-
- 29** A population is divided into three strata of sizes 1200, 600 and 200. A stratified sample of 60 is taken, with strata samples of size 36, 18 and 6, and the sample means for the strata are 52, 64 and 80 respectively.
- (a) Verify that the sample sizes are proportional to the stratum sizes. [2]
- (b) Estimate the population mean using the whole sample. [3]
- (c) A student suggests the estimate $\frac{52 + 64 + 80}{3}$. Explain why it is poorer, and calculate by how much it differs from your answer in (b). [3]
-
- 30** A population of $N = 60$ units is numbered 1 to 60 and a sample of $n = 10$ is taken.
- (a) In a simple random sample, show that the probability that a given unit is selected is $\frac{n}{N}$ and that the probability that two given units are both selected is $\frac{n(n-1)}{N(N-1)}$. Evaluate the second for these values. [4]
- (b) A systematic sample uses interval 6 and a random start in 1 to 6. Find the probability that units 1 and 7 are both selected, and the probability that units 1 and 2 are both selected. [3]
- (c) Use your answers to explain how systematic and simple random samples differ even though each unit has the same chance of selection in both. [2]

Sampling Bias

- 1** A cinema manager asks the first 20 people to arrive for a film whether they like the cinema's snacks.
- (a) Explain what is meant by a biased sample. [1]
- (b) Give one reason why this sample may be biased. [1]
-
- 2** State one source of bias in each of the following surveys.
- (a) A radio station asks listeners to phone in to say whether they support funding for local radio.
- (b) A study of exercise habits is carried out at 10 am on a weekday in a gym.
- (c) A question reads: "Don't you agree that the dreadful new road layout is a disaster?" [3]
-
- 3** Each of these survey questions is badly worded. Explain what is wrong and rewrite each one.
- (a) "How much did you enjoy our excellent service?"
- (b) "Do you eat healthily and exercise regularly?" [4]
-
- 4** A fast-food restaurant asks its customers how healthy they think their diet is. Give a reason why this sample could be biased. [1]

- 5** A question in a survey reads: "Don't you agree that the new bypass is a waste of money?" Explain what is wrong with this question and rewrite it in a neutral form. [2]
-
- 6** A questionnaire is sent to 200 customers and 50 reply. Find the response rate as a percentage, and give one reason why a low response rate can cause bias. [2]
-
- 7** A council asks residents: "Most people agree that the council wastes money. Do you agree that council tax should be cut?"
- (a) Explain why this question is likely to produce biased results. [1]
- (b) Write a better version of the question. [1]
- (c) The council posts the questionnaire to 500 residents and 80 reply. Find the response rate. [1]
- (d) Explain why the low response rate may bias the results. [1]
-
- 8** Two surveys are planned.
Survey A: to find how many hours people spend online, names are taken from a telephone directory.
Survey B: to find how often people shop in town centres, an online-only questionnaire is advertised on a shopping website.
- (a) Give two reasons why the telephone directory may be a biased sampling frame for Survey A. [2]
- (b) Give two reasons why Survey B is likely to be biased. [2]
-
- 9** A researcher selects people from the local telephone directory to ask about internet shopping habits.
Give two reasons why the sample may be biased. [2]
-
- 10** A surveyor stands outside a station at 10 am on a weekday and asks 50 people about their journey times to work.
Give two reasons why this sample may not be representative of all commuters. [2]
-
- 11** A website invites visitors to vote on whether a new tax is fair. Of the 240 who vote, 180 say it is fair.
- (a) Find the percentage who think the tax is fair. [1]
- (b) Explain why this does not show that 75% of the population think the tax is fair. [2]
-
- 12** A survey question asks: "How much do you spend on lunch each week? A: £0 to £5, B: £5 to £10, C: £10 to £15, D: £15 or more".
Give two criticisms of the response options. [2]
-
- 13** A set of scales always reads 0.5 kg too high. A student weighs 30 random pupils on them and finds a mean of 54.3 kg.
State what type of error this is and find the best estimate of the true mean mass. [2]
-
- 14** A survey about a new supermarket is sent by post to 1500 households. 240 questionnaires are returned and 168 of those say they are in favour.
- (a) Find the response rate, and the percentage of respondents who are in favour. [2]
- (b) Explain why 70% may not be a good estimate of the percentage of all 1500 households in favour. [1]
- (c) Find the lowest and highest possible percentages of all 1500 households that are in favour. [3]
-

- 15** A school council asks a sample of 100 students whether they support a new canteen menu. The school has 900 students. Information is shown in the table.

	Girls	Boys
Students in school	540	360
Students in sample	30	70
In favour (sample)	18	28

- (a) Find the proportion of the sample in favour. [1]
 (b) Explain why the sample is not representative of the school. [1]
 (c) Using the proportions of girls and boys in the school, calculate a better estimate of the proportion of the whole school in favour. [3]
 (d) Comment on how your answer affects the conclusion about support for the menu. [1]

- 16** In a survey of 400 people about a sensitive issue, 340 reply. Of those who reply, 120 say "yes".

- (a) Find the proportion of respondents who say yes, to 3 significant figures. [1]
 (b) Find the lowest and highest possible proportions of all 400 people who would say yes. [3]
 (c) Comment on the usefulness of the part (a) estimate. [1]

- 17** A poll sample of 100 people contains 70 women and 30 men, but the population is half men and half women. In the sample 60% of the women and 40% of the men support a proposal.

- (a) Find the percentage of the sample who support the proposal. [2]
 (b) Give an adjusted estimate of the percentage of the population who support the proposal. [2]
 (c) Explain why the sample estimate is biased. [1]

- 18** The age distribution of the population of a town is compared with the age distribution of the 100 people who completed a survey.

Age group	Population (%)	Survey respondents
18-34	30	12
35-54	35	30
55+	35	58

- (a) Find the number of respondents expected in each age group if the survey were representative. [1]
 (b) Identify the groups that are over- and under-represented and suggest one possible cause. [2]
 (c) In the survey, 25% of 18 to 34s, 50% of 35 to 54s and 70% of those aged 55+ support a proposal. Find the percentage of the sample in support and the weighted estimate for the population. [3]

- 19** A lab technician measures a standard mass known to be 50.0 g ten times on a balance. The readings are 50.6, 50.9, 50.4, 50.8, 50.7, 50.5, 50.9, 50.6, 50.7, 50.5 g.

- (a) Find the mean reading and estimate the bias of the balance. [3]
 (b) A new sample reads a mean of 72.15 g on the same balance. Find the corrected estimate and explain why taking more readings does not remove the bias. [2]

- 20** For each survey say which source of bias is the most serious: A undercoverage of the sampling frame, B non-response, C leading question, D self-selection.

- (a) "Do you not agree that the council should plant more trees?"
 (b) A radio phone-in asks listeners whether they support the local football team.
 (c) A survey of all residents is taken from the electoral register of a university town in which many students are not registered.
 (d) 2000 questionnaires are posted out and 150 are returned. [4]

- 21** A college has three classes of sizes 10, 20 and 90 students.
- (a) Find the mean class size. [2]
 - (b) A student is chosen at random from the 120 students and asked the size of their class. Let X be the answer given. Find the probability distribution of X and hence $E(X)$. [4]
 - (c) Explain why $E(X)$ is much larger than the mean class size. [1]
 - (d) Suggest how the sampling should be done to estimate the mean class size without this bias. [1]
-
- 22** A company selects 400 customers at random and asks about their weekly spending (£). Only 260 reply, with a mean of £34.50. A follow-up survey of 40 randomly chosen non-respondents gives a mean of £21.30.
- (a) Use the follow-up result to estimate the mean weekly spending of all 400 customers. [3]
 - (b) Find the size and direction of the error in using only the 260 replies. [2]
 - (c) Give two reasons why the answer to part (a) is still only an estimate. [2]
-
- 23** Two plans are proposed to find the proportion p of a town's residents who support a new cycle lane.
- Plan 1: 100 residents are chosen at random from the electoral register. 62 reply, and 31 of those support the lane.
- Plan 2: a website poll attracts 400 votes, of which 328 support the lane.
- (a) Find the proportion supporting the lane in each plan. [2]
 - (b) Which plan do you think gives the better estimate of p ? Give two reasons. [3]
 - (c) State one weakness of Plan 1. [1]
-
- 24** A postal survey is sent to 1000 randomly chosen households. 600 reply and 360 of them support a proposal. A random sample of 40 of the non-respondents is then visited, and 10 of them support the proposal.
- (a) Find the proportion of respondents who support the proposal and the proportion of the 40 who support it. [2]
 - (b) Estimate the number of the 1000 households who support the proposal. [3]
 - (c) Explain why the original figure of 60% was misleading. [1]
-
- 25** A charity wants 240 completed questionnaires from its supporters. Experience suggests that 30% of supporters reply to a postal questionnaire, and 75% reply when the questionnaire is hand-delivered with a follow-up visit. A postal questionnaire costs £0.80 and a hand-delivered one costs £3.00.
- (a) Find how many questionnaires must be sent out by each method to expect 240 replies. [2]
 - (b) Find the cost of each method. [2]
 - (c) Give one reason why the more expensive method may be better, and one reason why it may not be worth the extra cost. [2]
 - (d) Explain why sending out more postal questionnaires would not remove non-response bias. [1]
-
- 26** A telephone survey uses a list of landline numbers. 80% of the households in a town have a landline and 20% do not. A proposal is supported by 70% of households with a landline and by 40% of those without.
- (a) Find the percentage of all households that support the proposal. [2]
 - (b) Find the expected percentage supporting in the telephone survey and the size of the bias. [2]
 - (c) Suggest a way of reducing this bias. [1]

- 27 A school wants to estimate the proportion p of its students who have copied homework. Each of 200 randomly chosen students tosses a fair coin in private. If the coin shows heads the student answers "yes". If it shows tails the student answers truthfully. 130 students answer "yes".
- Explain why this method may give a more accurate estimate than asking the question directly. [2]
 - Show that the probability that a randomly chosen student answers "yes" is $\frac{1}{2} + \frac{1}{2}p$. [2]
 - Use the results to estimate p . [2]
 - A second school uses a biased coin that shows heads with probability 0.6, with the same rules, and also finds that $\frac{130}{200}$ answer "yes". Estimate p for this school. [2]
 - Explain why the estimate from the second school is likely to be less reliable than the first, even though the number of students is the same. [2]

- 28 A researcher stands outside a gym at 6 pm on a Monday and asks the first 40 people leaving whether they exercise at least four times a week. 34 people say yes. She concludes that 85% of the 60 000 adults in the town exercise at least four times a week.

Age group	18-34	35-54	55+
Adults in town	21000	24000	15000

- Identify four separate weaknesses in her method and conclusion. [4]
 - A better design uses a stratified sample of 400 adults, with the three age groups in the table as strata. Find the sample size for each group. [2]
 - Describe how the adults in one group could be chosen. [2]
 - In the stratified sample, 56, 48 and 22 people in the three groups exercise at least four times a week. Estimate the percentage of adults in the town who do so. [2]
- 29 A population consists of the three values 1, 2 and 3, each equally likely. Samples of size 2 are taken with replacement, so there are nine equally likely samples. The variance of a sample is defined as $\frac{1}{n}\sum(x - \bar{x})^2$, where $n = 2$ is the sample size.
- Show that the variance of the population is $\frac{2}{3}$. [2]
 - List the sample variance for each of the nine samples. [3]
 - Find the mean of the nine sample variances and explain why it shows that this sample variance is a biased estimator of the population variance. [3]
 - Show that multiplying each sample variance by $\frac{n}{n-1}$ removes the bias. [1]

- 30 A survey is run on 100 respondents to estimate the percentage of a population who use a certain app. The table shows, for each age group, its percentage of the population, the percentage of that group who use the app, and the number of respondents.

Age group	% of population	% using app	Respondents
Under 30	40	14	20
30 to 59	40	30	50
60 and over	20	56	30

- Estimate the percentage of the population using the app from the survey as it stands. [3]
- Find the percentage using the app in the population. [2]
- Calculate the bias in the survey estimate and identify its cause. [2]
- Describe how the survey could be redesigned to avoid this bias. [2]

Measures of Location and Spread

- 1 Find the mean, median and mode of 3, 8, 5, 9, 4, 8, 6, 8, 12. [3]

- 2 The ordered data 12, 15, 17, 18, 21, 24, 29, 35 are given. Find the range and the interquartile range, taking the quartiles to be the medians of the lower and upper halves of the data. [3]

- 3 The ordered stem and leaf diagram shows the scores of 15 players in a game. Find the median, the mode and the range of the scores.

```

1 | 2 4 7
2 | 1 3 3 6 9
3 | 0 2 5 5 5 8
4 | 1

```

Key: 2 | 1 means 21

[3]

- 4 The table shows the shoe sizes of 30 students. Find the mode, the median and the range of the sizes.

Size	3	4	5	6	7
Frequency	2	7	11	6	4

[3]

- 5 For 10 values, $\Sigma x = 120$ and $\Sigma x^2 = 1580$. Find the variance and the standard deviation. [3]

- 6 The number of faults found in each of 35 items is shown in the table. Find the mean number of faults.

Faults	0	1	2	3	4
Frequency	4	9	12	8	2

[2]

- 7 The table shows the number of goals scored by a team in each of 40 matches. Find the mean, median and mode number of goals.

Goals	0	1	2	3	4	5
Matches	4	9	12	8	5	2

[4]

- 8 The table summarises the times, in minutes, taken by 60 students to complete a puzzle.

Time t (min)	0-10	10-20	20-30	30-40	40-60
Frequency	6	14	22	11	7

- (a) Estimate the mean time. [2]
(b) State the class that contains the median. [1]
(c) Use your calculator to estimate the standard deviation. [2]

- 9 The ordered data 23, 25, 26, 28, 29, 31, 33, 34, 36, 40, 45 are given. Taking the quartiles to be the medians of the lower and upper halves (excluding the median), find the median, the quartiles and the interquartile range. [4]

- 10 Find, from the definition $\Sigma(x - \bar{x})^2/n$, the variance and standard deviation of 4, 7, 9, 12, 13, 17. [4]

- 11** The table shows the times t seconds taken by 60 people to complete a task. Estimate the mean time and state the modal class.

Time t (s)	Frequency
$0 \leq t < 10$	6
$10 \leq t < 20$	14
$20 \leq t < 30$	25
$30 \leq t < 50$	15

[3]

- 12** The ages of the seven people in a lift are 34, 36, 35, 37, 33, 35 and 120.

(a) Find the mean and the median age.

[2]

(b) Explain why the median is a better average here.

[1]

- 13** The heights of 60 seedlings, measured to the nearest centimetre, are summarised in the table.

Height (cm)	4 to 6	7 to 9	10	11 to 14	15 to 20
Frequency	8	15	12	20	5

(a) Write down the lower and upper class boundaries of the class 4 to 6 and of the class 10.

[2]

(b) Estimate the mean height.

[3]

- 14** For 12 values of x , $\sum x = 480$ and $\sum x^2 = 19\,800$.

(a) Find the mean and the standard deviation of the values.

[3]

(b) A 13th value, 61, is added. Find the new mean and standard deviation.

[3]

- 15** In a test, class A (20 students) had a mean mark of 62 and a standard deviation of 8. Class B (30 students) had a mean of 55 and a standard deviation of 10. Find the mean and the standard deviation of the marks of all 50 students.

[6]

- 16** The table summarises the distances d km travelled to work by 50 employees.

Distance d (km)	Frequency
$0 \leq d < 5$	8
$5 \leq d < 10$	12
$10 \leq d < 20$	20
$20 \leq d < 30$	8
$30 \leq d < 40$	2

(a) Estimate the mean distance.

[2]

(b) Estimate the standard deviation.

[3]

- 17** The ages of the 80 members of a club are summarised in the table. Use linear interpolation to estimate the median and the interquartile range.

Age a (years)	Frequency
$20 \leq a < 30$	10
$30 \leq a < 40$	22
$40 \leq a < 50$	28
$50 \leq a < 70$	16
$70 \leq a < 90$	4

[5]

- 18** The times, in seconds, taken by 11 runners to complete a short course are 12.8, 13.1, 13.4, 13.6, 13.9, 14.1, 14.4, 14.9, 15.8, 17.6 and 19.5.

A coefficient of skewness is defined as $\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$.

- (a) Calculate the mean and the standard deviation of the times. [3]
 (b) Write down the median. [1]
 (c) Evaluate the coefficient of skewness and state the type of skew. [3]

- 19** The back-to-back stem and leaf diagram shows the marks of 12 students in group A and 12 students in group B in the same test.

Group A		Group B
8 5 2	2	
8 6 3 1	3	1 5 8
6 4 1	4	2 4 5 7 9
7 2	5	1 3 5
	6	3

Key: 6 | 3 | 1 means 36 for group A and 31 for group B

- (a) Find the median and the interquartile range for each group, taking quartiles to be the medians of the lower and upper halves. [4]
 (b) Compare the marks of the two groups. [2]

- 20** The table shows the number of pets owned by each of 40 pupils. The mean number of pets is 1.95. Find a and b.

Pets	0	1	2	3	4
Frequency	a	12	15	b	3

[5]

- 21** The lengths of 80 leaves are summarised in the table. Use linear interpolation, taking the quartiles to be the $\frac{n}{4}$ th, $\frac{n}{2}$ th and $\frac{3n}{4}$ th values.

Length x (mm)	10-20	20-25	25-30	30-40	40-60
Frequency	8	20	26	18	8

- (a) Estimate the lower quartile, the median and the upper quartile. [6]
 (b) Describe the skew of the distribution, giving a reason. [1]

- 22** Five numbers are 4, 6, a, b and 10, where $a < b$. Their mean is 7 and their variance is 4.

- (a) Show that $a + b = 15$. [2]
 (b) Show that $a^2 + b^2 = 113$. [2]
 (c) Find the values of a and b. [3]

- 23** The masses, in kg, of 12 parcels are 52, 55, 58, 49, 61, 53, 57, 54, 56, 59, 51 and 89. A value is called an outlier if it is more than 2 standard deviations from the mean.

- (a) Calculate the mean and the standard deviation. [4]
 (b) Identify any outliers by this rule. [2]
 (c) Remove the outlier and recalculate the mean. Comment on the effect. [3]

- 24** The table shows the scores x in a game played 34 times. Some of the frequencies are given in terms of a constant k.

Score x	1	2	3	4	5
Frequency	2k	k + 4	3k - 1	k + 1	2

- (a) Find the value of k. [2]
 (b) Find the mean and the standard deviation of the scores. [4]
 (c) Find the median score. [1]

- 25** Two machines A and B fill bags of flour. The masses x grams of 40 bags from each machine give these summaries.
Machine A: $\Sigma x = 2016$, $\Sigma x^2 = 101\,650$. Machine B: $\Sigma x = 2004$, $\Sigma x^2 = 100\,631$.
The target mass is 50 g.
- Calculate the mean and standard deviation for each machine. [4]
 - Compare the performance of the machines, referring to the target mass. [2]
-
- 26** A class of 30 students took a test. The marks of the whole class had mean 58 and standard deviation 12. The 18 boys had a mean mark of 61 and a standard deviation of 10.
- Find the mean mark of the 12 girls. [3]
 - Find the standard deviation of the girls' marks. [4]
-
- 27** For 25 readings, a student calculated a mean of 48 and a standard deviation of 6. She then found that one reading, recorded as 70, should have been 40.
- Show that $\Sigma x = 1200$ and $\Sigma x^2 = 58\,500$ for the original readings. [3]
 - Find the correct mean. [2]
 - Find the correct standard deviation. [3]
 - Explain why the standard deviation has decreased. [1]
-
- 28** For n values $x_1, x_2, \dots, x_n$ with mean $\bar{x}$:
- Show that $\Sigma(x - \bar{x})^2 = \Sigma x^2 - n\bar{x}^2$. [3]
 - Show that for any constant a , $\Sigma(x - a)^2 = \Sigma(x - \bar{x})^2 + n(\bar{x} - a)^2$. [3]
 - Deduce the value of a for which $\Sigma(x - a)^2$ is least. [1]
 - For the values 2, 4, 6, 8, 10 use part (b) to show that $\Sigma(x - 4)^2 = 60$. [2]
-
- 29** The ordered stem and leaf diagram shows the marks of 15 students in a test. The digits a and b are unknown.
The median mark is 43 and the interquartile range is 19. Take quartiles to be the medians of the lower and upper halves of the data (excluding the median when there is an odd number of values).
- | | | |
|---|--|---------|
| 2 | | 3 8 |
| 3 | | 1 4 6 9 |
| 4 | | 0 a 4 7 |
| 5 | | 1 b 5 |
| 6 | | 2 6 |
- Key: 2 | 3 means 23
- Find the lower quartile. [2]
 - Find the value of a . [2]
 - Find the value of b . [3]
 - Show that none of the marks is an outlier, using the rule that outliers lie more than $1.5 \times$ IQR beyond a quartile. [2]
-
- 30** The integers 1, 2, ..., n are used as a data set.
- Show that the variance of the data is $\frac{n^2 - 1}{12}$. You may use $\Sigma r = \frac{n(n+1)}{2}$ and $\Sigma r^2 = \frac{n(n+1)(2n+1)}{6}$. [5]
 - Hence find the standard deviation of 1, 2, ..., 20, to 3 significant figures. [2]
 - Find the mean and standard deviation of the data 5, 10, 15, ..., 100. [3]

Histograms and Cumulative Frequency

- 1 The table shows the heights of 88 students.
Calculate the frequency density of each class.

Height h (cm)	140-150	150-160	160-180	180-200
Frequency	12	30	36	10

[3]

- 2 The masses of 50 parcels are summarised in the table.

Mass m (g)	0-20	20-40	40-60	60-80	80-100
Frequency	5	12	20	9	4

- (a) Write down the cumulative frequencies for $m < 20$, $m < 40$, $m < 60$, $m < 80$ and $m < 100$. [2]
 (b) State the coordinates of the first point you would plot on a cumulative frequency graph. [1]

- 3 The times taken, to the nearest minute, by 40 people to read a leaflet are grouped as 1 to 3, 4 to 6, 7 to 10 and 11 to 20.

Write down the class boundaries and the class width of each class.

[3]

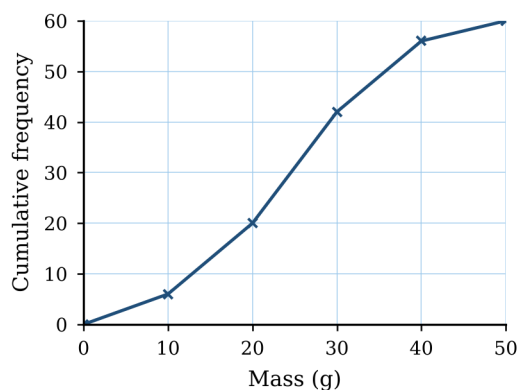
- 4 A bar in a histogram is 4 units wide and 2.5 units high. Find the frequency it represents.
 Another bar represents 35 observations and has width 10. Find its height.

[2]

- 5 In a histogram, a rectangle of width 2 cm and height 6 cm represents a frequency of 30.
 Another rectangle in the same histogram has width 3.5 cm and height 4 cm.
 Find the frequency represented by the second rectangle.

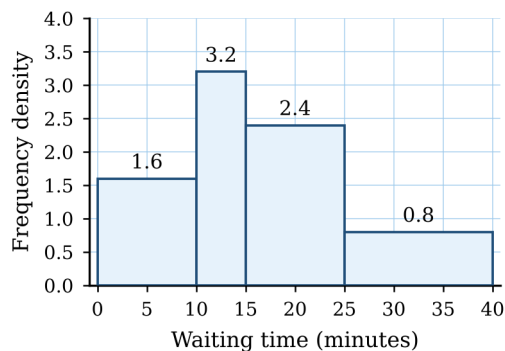
[2]

- 6 The cumulative frequency graph shows the masses of 60 eggs. Use it to estimate the median mass and the number of eggs with mass less than 35 g.



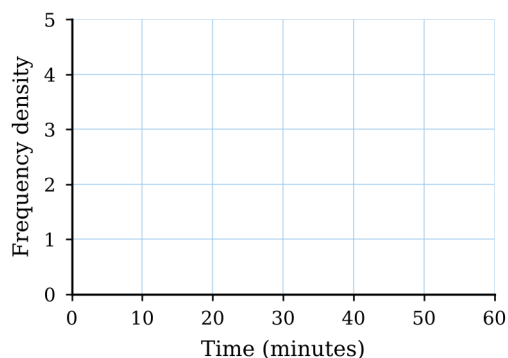
[2]

- 7 The histogram shows the waiting times of passengers at a station.



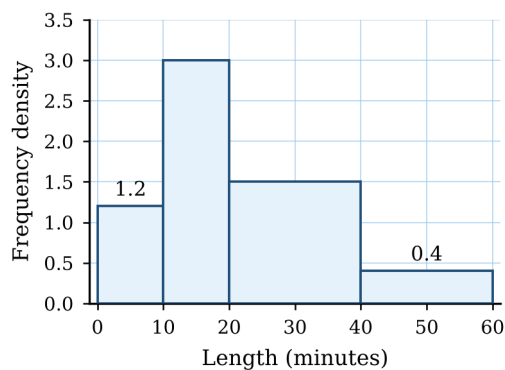
- (a) Find the number of passengers who waited between 15 and 25 minutes. [2]
 (b) Find the total number of passengers. [2]

- 8 The times, in minutes, that 88 customers spent in a shop are: $0 \leq t < 5$: 8 customers; $5 \leq t < 10$: 20; $10 \leq t < 20$: 30; $20 \leq t < 40$: 24; $40 \leq t < 60$: 6.
 Draw a histogram to represent these data on the grid below.



[5]

- 9 The histogram shows the lengths of 80 phone calls.



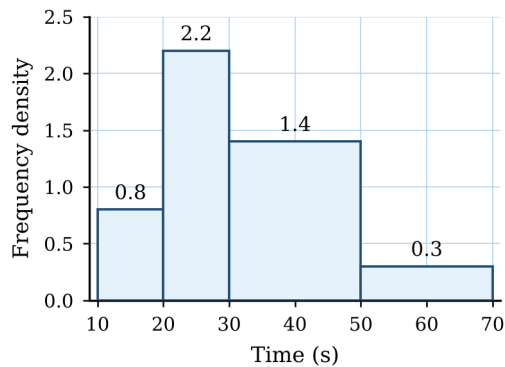
- (a) Find the frequency of each class and check that the total is 80. [3]
 (b) State the modal class. [1]

- 10 Complete the table, where ? marks the missing values.

Class	Frequency	Frequency density
$0 \leq x < 5$	15	?
$5 \leq x < 10$	?	4.2
$10 \leq x < 20$	28	?
$20 \leq x < 40$	?	0.6

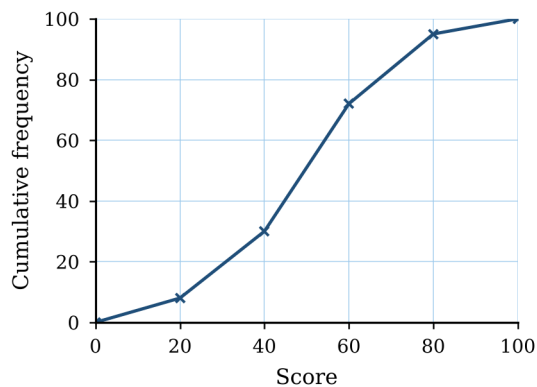
[4]

- 11** The histogram shows the times, in seconds, taken by competitors to solve a puzzle. Estimate the number of competitors who took between 24 and 40 seconds.



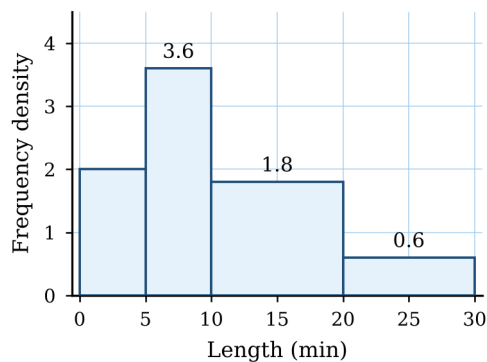
[3]

- 12** The cumulative frequency graph shows the scores of 100 candidates in a test.



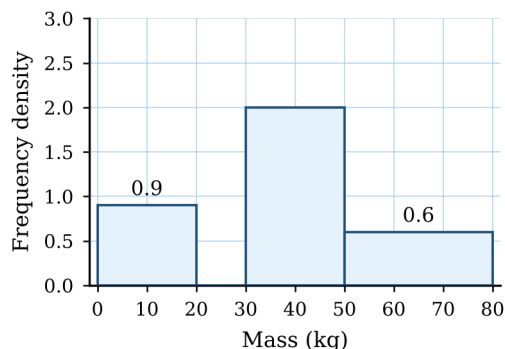
- (a) Estimate the median score. [1]
(b) Estimate the number of candidates who scored more than 70. [2]

- 13** The histogram shows the lengths in minutes of 52 films clips. Estimate the mean length.



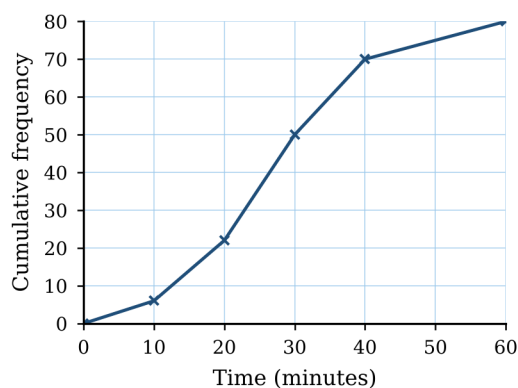
[4]

- 14** The histogram is incomplete. It shows the masses of 100 animals, but the bar for the class $20 \leq m < 30$ has not yet been drawn.



- (a) Find the number of animals in the class $20 \leq m < 30$. [2]
 (b) Find the height of the missing bar. [1]
 (c) Estimate the number of animals with mass between 40 and 65. [3]

- 15** The cumulative frequency polygon shows the times taken by 80 runners to complete a course. Use linear interpolation between the plotted points, which are (0, 0), (10, 6), (20, 22), (30, 50), (40, 70) and (60, 80).



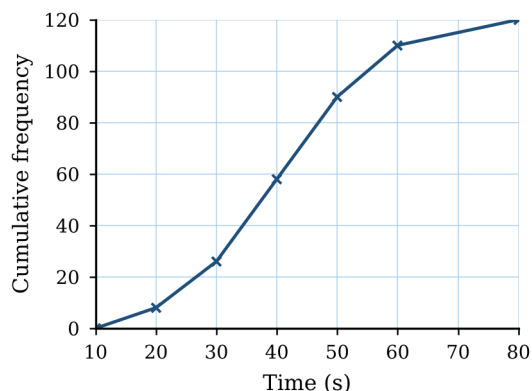
- (a) Estimate the median time. [2]
 (b) Estimate the interquartile range. [3]
 (c) Estimate the number of runners who took longer than 45 minutes. [1]

- 16** The table shows the number of hours, to the nearest hour, that 105 people spent exercising last week.

Hours (nearest hour)	1 to 3	4 to 7	8 to 9	10 to 14	15 to 23
Frequency	15	28	24	20	18

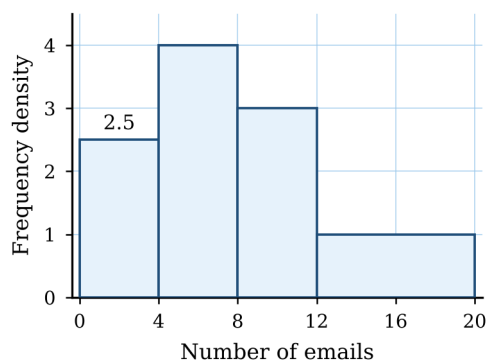
- (a) Write down the class boundaries and calculate the frequency density of each class. [3]
 (b) Estimate the number of people who exercised for between 5.5 and 11.5 hours. [3]
 (c) Estimate the median number of hours. [3]

- 17** The cumulative frequency graph shows the times taken by 120 students to answer a quiz.



- (a) Estimate the median and the interquartile range. [4]
 (b) Estimate the 90th percentile. [2]

- 18** The histogram shows the number of emails received by 46 workers in a day.



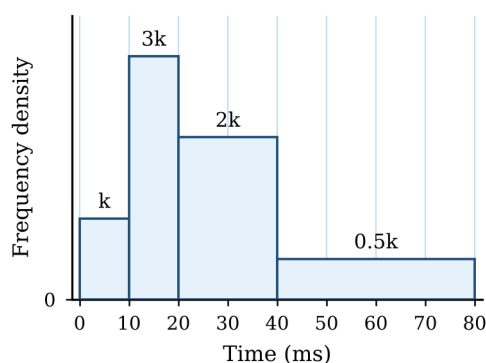
- (a) Show that the total of the frequencies is 46. [2]
 (b) Estimate the median. [3]

- 19** The cumulative frequencies of the ages of 80 visitors to a museum are shown in the table.

Age a (years)	$a < 10$	$a < 20$	$a < 30$	$a < 50$	$a < 80$
Cumulative frequency	5	18	46	70	80

- (a) Find the frequency and the frequency density of each class. [3]
 (b) Estimate the number of visitors aged between 25 and 40. [2]

- 20** The histogram shows the distribution of reaction times (ms) of 120 people. The heights of the bars are k , $3k$, $2k$ and $0.5k$.



- (a) The class widths are 10, 10, 20 and 40. Show that $100k = 120$ and find k . [3]
 (b) Find the number of people with a reaction time of less than 30 ms. [2]

- 21** The table summarises the heights of 120 saplings.

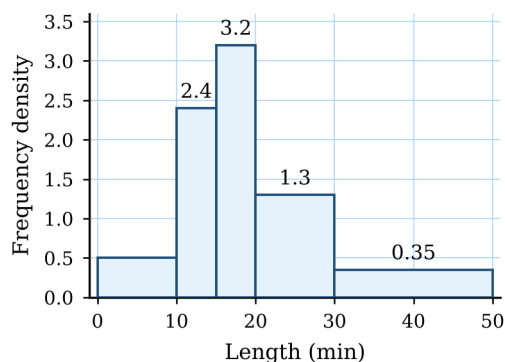
Height h (cm)	10 to 20	20 to 25	25 to 30	30 to 35	35 to 40	40 to 60
Frequency	6	15	34	36	20	9

- (a) Estimate the mean and the standard deviation of the heights. [4]
 (b) Use linear interpolation to estimate the percentage of the saplings with height within one standard deviation of the mean. [3]
 (c) For a Normal distribution about 68% of the values lie within one standard deviation of the mean. Compare this with your answer to part (b) and comment on whether the data could be modelled by a Normal distribution. [2]

- 22** In a histogram, the bar for the class $20 \leq t < 24$ is 2 cm wide and 4.5 cm high, and represents 18 observations. The class $24 \leq t < 36$ has frequency 30, the class $36 \leq t < 60$ has frequency 18, and the total frequency is 100. The remaining observations are in the class $0 \leq t < 20$.

- (a) Find the width of the bar for the class $24 \leq t < 36$. [2]
 (b) Find the height of this bar. [2]
 (c) Find the width and height of the bar for $36 \leq t < 60$. [2]
 (d) Find the width and height of the bar for $0 \leq t < 20$. [2]

- 23** The histogram summarises the lengths of 53 songs, in minutes.



- (a) Find the frequency of each class. [2]
 (b) Estimate the mean and the standard deviation. [4]
 (c) Estimate the median and comment on the skewness. [3]

- 24** The table shows the lengths, to the nearest minute, of 82 phone calls.

Length (nearest minute)	1 to 2	3 to 4	5	6 to 8	9 to 14
Frequency	10	16	20	18	18

- (a) Estimate the mean and the standard deviation of the lengths. [4]
 (b) Use linear interpolation to estimate the median length. [2]
 (c) In a histogram of these data the class 1 to 2 is drawn as a rectangle of base 2.5 cm and height 4 cm. Find the base and height of the rectangle for the class 9 to 14. [3]

- 25** The marks of 200 students in an examination are summarised in the table.

Mark x	Frequency
$0 \leq x < 20$	14
$20 \leq x < 40$	46
$40 \leq x < 50$	52
$50 \leq x < 60$	48
$60 \leq x < 80$	30
$80 \leq x < 100$	10

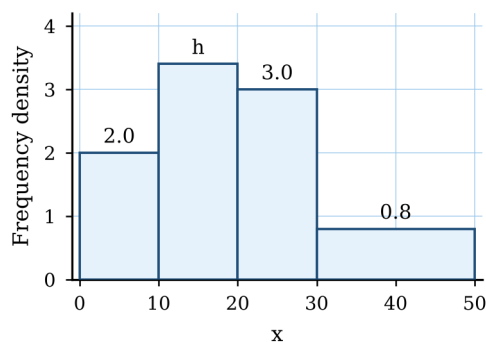
- (a) Find the cumulative frequencies. [1]
 (b) The pass mark is set so that 65% of the students pass. Estimate the pass mark. [3]
 (c) The top 10% of students are awarded grade A. Estimate the lowest mark for grade A. [3]

- 26** The table shows the masses of a sample of $(50 + x)$ fruits. The median mass, estimated by linear interpolation, is 23.75 g.

Mass m (g)	Frequency
$0 \leq m < 10$	10
$10 \leq m < 20$	x
$20 \leq m < 30$	24
$30 \leq m < 50$	16

[6]

- 27** The histogram shows the distribution of a variable x . The frequency densities of the bars are marked, and the bar for $10 \leq x < 20$ has frequency density h . An estimate of the mean of x is 20.



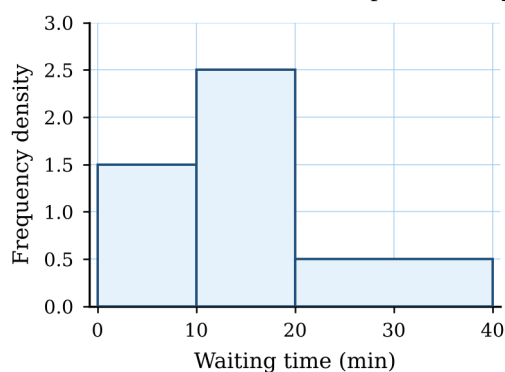
- (a) Show that $h = 3.4$. [4]
 (b) Estimate the median of x . [2]
 (c) Estimate the probability that a randomly chosen value of x lies between 15 and 35. [3]
 (d) Explain why the mean is higher than the median. [1]

- 28** The table shows the cumulative frequencies of the scores in an examination for 200 students at school A and 160 students at school B. Use linear interpolation throughout.

Score s , $s <$	20	40	50	60	80	100
School A	8	40	84	140	188	200
School B	16	56	88	112	148	160

- (a) Estimate the median score at each school. [3]
 (b) Estimate the interquartile range at each school. [3]
 (c) Estimate the percentage of students scoring more than 70 at each school. [3]
 (d) Compare the performance of the two schools. [2]

- 29** The histogram shows the distribution of the times, in minutes, that 50 customers spent waiting. Assume that within each class the times are spread evenly.



- (a) Find the frequency of each class. [2]
 (b) A customer is chosen at random from those who waited more than 10 minutes. Find the probability that the customer waited more than 25 minutes. [4]
 (c) Two customers are chosen at random from all 50, without replacement. Find the probability that both waited less than 10 minutes. [3]

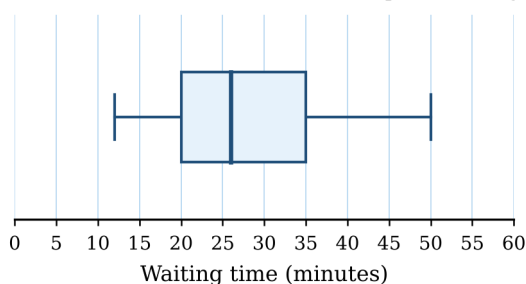
- 30** The table summarises the ages of the 60 people on a coach trip.

Age a (years)	Frequency
$0 \leq a < 10$	8
$10 \leq a < 20$	18
$20 \leq a < 30$	21
$30 \leq a < 50$	13

- (a) Estimate the mean age using class midpoints. [3]
 (b) Calculate the smallest and the largest possible values of the true mean. [3]
 (c) The coach company records that the ages of the 60 people actually add up to 1300. Find the true mean and the error in the estimate in part (a). [2]
 (d) Explain why the estimate in part (a) must lie between the two values in part (b), and why using midpoints is reasonable. [2]

Box Plots and Outliers

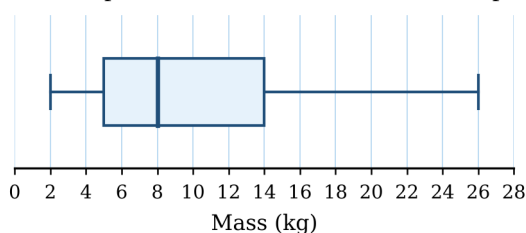
- 1** The box plot summarises the waiting times, in minutes, of patients at a clinic. Write down the median, the interquartile range and the range.



[3]

- 2** The ordered data 4, 7, 9, 12, 14, 18, 21, 30 are given. Taking the quartiles to be the medians of the lower and upper halves, write down the five-number summary (minimum, lower quartile, median, upper quartile, maximum). [3]

- 3** The box plot summarises the masses of 80 parcels.



- (a) How many parcels have a mass of at most 8 kg? [1]
 (b) How many parcels have a mass between 5 kg and 14 kg? [1]
 (c) What percentage of the parcels have a mass greater than 14 kg? [1]

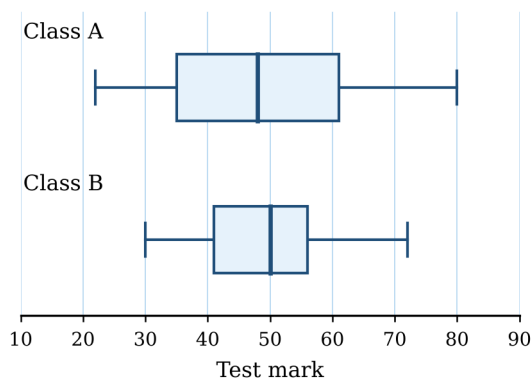
- 4** A data set has lower quartile 20 and upper quartile 32. Find the limits beyond which a value is classed as an outlier, using the rule "more than $1.5 \times \text{IQR}$ from the nearer quartile". [3]

- 5** Find the median and the quartiles of 6, 9, 11, 14, 15, 18, 21, 22, 30, taking the quartiles to be the medians of the lower and upper halves (excluding the median). [3]

- 6** A data set has $Q_1 = 30$ and $Q_3 = 42$. Which of the values 8, 11, 62, 70 are outliers under the $1.5 \times \text{IQR}$ rule? [3]

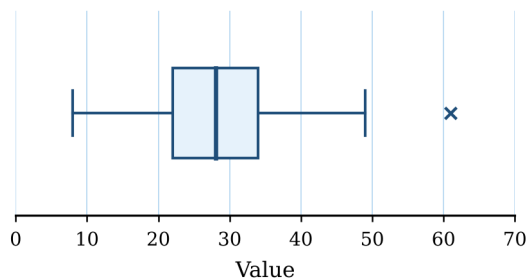
- 7** A data set has lower quartile 34 and upper quartile 46. Its smallest value is 12, its second smallest is 20, its largest is 71 and its second largest is 58. A value is an outlier if it is more than $1.5 \times \text{IQR}$ below Q_1 or above Q_3 .
- (a) Determine which values, if any, are outliers. [3]
- (b) Write down where the whiskers of a box plot of the data would end. [1]

- 8** The box plots show the marks of two classes in the same test.



- (a) Find the interquartile range of each class. [2]
- (b) Compare the marks of the two classes. [2]
- 9** A data set has minimum 12, lower quartile 20, median 26, upper quartile 34 and maximum 47. There are no outliers. Draw a box plot for the data. [3]
- 10** The times, in seconds, taken by 9 students to complete an obstacle course are 14, 17, 18, 21, 22, 25, 26, 29 and 58. Take quartiles to be the medians of the lower and upper halves of the data (excluding the median).
- (a) Find the median and the quartiles. [3]
- (b) Show that 58 is an outlier. [2]
- (c) Draw a box plot for the data. [2]
- 11** A data set has $Q_1 = 18$, median = 22 and $Q_3 = 35$. Describe the skewness of the data, supporting your answer with a calculation. [3]
- 12** The mean of a data set is 64 and the standard deviation is 6. A value is an outlier if it is more than 2 standard deviations from the mean. Decide whether 49 and 75 are outliers. [3]
- 13** The twelve values 3, 5, 6, 8, 9, 11, 12, 14, 15, 17, 18, 45 are given. Find the quartiles and use the $1.5 \times \text{IQR}$ rule to identify any outliers. [5]
- 14** A data set has minimum 18, lower quartile 40, median 47, upper quartile 55 and maximum 84. Its largest value that is not 84 is 66. A value is an outlier if it lies more than $1.5 \times \text{IQR}$ beyond a quartile.
- (a) Show that 84 is an outlier but 18 is not. [3]
- (b) Draw a box plot of the data, marking any outlier with a cross. [3]

- 15** The box plot shows the data for a sample. The lower quartile is 22, the median is 28 and the upper quartile is 34. The whiskers end at 8 and 49, and one outlier is marked at 61. Outliers are values more than $1.5 \times \text{IQR}$ beyond a quartile.

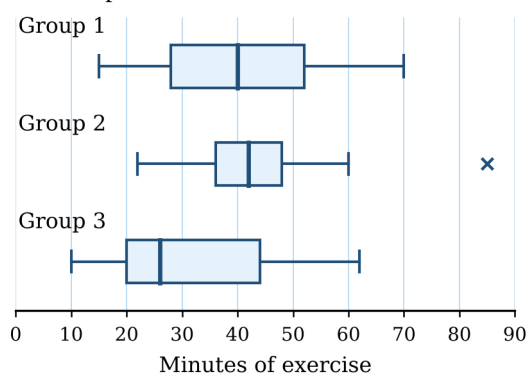


- (a) Find the interquartile range. [1]
 (b) Show that 61 is an outlier. [2]
 (c) Explain what the position of the upper whisker tells you about the data above 49. [2]
 (d) Show that the data have no low outliers. [1]
- 16** The numbers of emails received in one day by 15 members of staff are 33, 12, 55, 24, 19, 47, 26, 81, 15, 22, 38, 17, 30, 21, 28.
- (a) Draw an ordered stem and leaf diagram for the data. [3]
 (b) Find the median and the quartiles. [3]
 (c) Determine whether there are any outliers, using the rule that outliers lie more than $1.5 \times \text{IQR}$ beyond a quartile. [2]

- 17** A set of 14 times (in seconds) is 14, 16, 17, 19, 20, 21, 22, 23, 25, 26, 27, 29, 31 and 98. The value 98 is removed as an outlier.
- (a) Find the median of the 14 values and of the 13 remaining values. [2]
 (b) Find the mean before and after removing 98, and comment on which average changes more. [4]

- 18** The nine values 12, 14, 15, 17, 19, 20, 22, 23 and x are in increasing order, where $x \geq 23$. Taking quartiles as the medians of the lower and upper halves (excluding the median):
- (a) Find the quartiles and the interquartile range. [4]
 (b) Find the range of values of x for which x is an outlier. [3]

- 19** The box plots show the number of minutes of exercise taken per day by three groups of people.



- (a) Which group has the greatest interquartile range? Give its value. [2]
 (b) Describe the skewness of Group 3. [2]
 (c) Show that 85 in Group 2 is an outlier. [2]
- 20** A box plot shows the weekly sales (in £) of a shop over 120 weeks. The lower quartile is £400, the median is £520 and the upper quartile is £700.
- (a) How many weeks had sales above £700? [1]
 (b) How many weeks had sales between £400 and £700? [1]
 (c) A manager says "Sales were between £400 and £520 in half of the weeks." Explain the mistake and correct it. [2]

- 21** Twelve people in each of two groups were timed doing a puzzle. The times, in order, for group A are shown. For group B the summary is: minimum 22, lower quartile 35, median 45, upper quartile 55, maximum 64 (minutes).

Take quartiles to be the medians of the lower and upper halves of the data.

	Times (minutes)
Group A	31 34 36 38 40 41 43 45 47 49 52 78

- (a) Find the median and quartiles for group A. [3]
 (b) Show that 78 is an outlier for group A, and that group B has no outliers. [2]
 (c) Draw box plots for both groups on the same scale. [2]
 (d) Compare the two groups, saying whether you would use the mean or the median to describe group A and why. [2]

- 22** The temperatures, in °C, of a set of samples have median 40 and lower quartile 34. The upper fence for outliers ($Q_3 + 1.5 \times \text{IQR}$) is 64.

- (a) Find the upper quartile. [3]
 (b) A sample has temperature 15 °C. Determine whether it is an outlier. [2]
 (c) The temperatures are converted to °F using $F = 1.8C + 32$. Find the median and the interquartile range in °F. [2]

- 23** A sensor records 25 readings. Summary statistics are $n = 25$, $\Sigma x = 757$ and $\Sigma x^2 = 27631$. The reading 95 is thought to be a recording error.

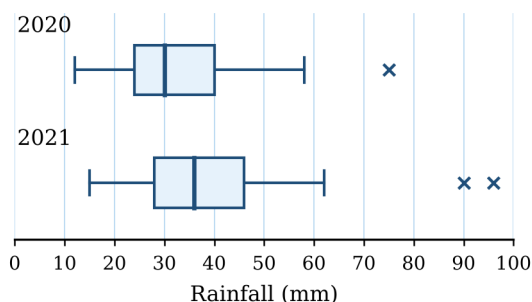
- (a) Calculate the mean and standard deviation of the 25 readings. [3]
 (b) Show that 95 is more than 2 standard deviations from the mean. [1]
 (c) Remove 95 and find the new mean and standard deviation. [3]
 (d) Comment on the effect of the outlier on the mean and standard deviation. [1]

- 24** The table summarises the times (in minutes) taken by 100 people to complete a form. Use linear interpolation to estimate the quartiles, and decide where the upper whisker of a box plot must end at most.

Time t (min)	Frequency
$0 \leq t < 10$	10
$10 \leq t < 20$	24
$20 \leq t < 30$	30
$30 \leq t < 40$	20
$40 \leq t < 60$	12
$60 \leq t < 100$	4

[7]

- 25** The box plots show the monthly rainfall (mm) at a weather station over the years 2020 and 2021.



- (a) For 2021, verify that the plotted outliers are correct, given that the whiskers end at 15 and 62. [3]
 (b) Compare 2020 and 2021 in context. [3]

- 26** The ordered stem and leaf diagram shows the scores of 23 employees in a skills test. The company decides to retrain any employee whose score is less than $Q_1 - \text{IQR}$.

Take quartiles to be the medians of the lower and upper halves of the data (excluding the median).

0	8
1	8
2	5 7 9
3	1 4 4 6 8
4	0 2 3 5 6 7 9
5	0 1 3 3 4 5

Key: 4 | 2 means 42

- Find the median and the interquartile range. [3]
- Show that only one employee will be retrained. [2]
- Using the rule that an outlier lies more than $1.5 \times \text{IQR}$ below Q_1 , state whether the employee who is retrained has an outlier score. [2]
- Describe the skewness of the data, giving a reason. [2]

- 27** For a normal distribution, a box plot uses the rule that a value is an outlier if it is more than $1.5 \times \text{IQR}$ below Q_1 or above Q_3 . Let $Z \sim N(0, 1)$. You may use $P(Z < 0.6745) = 0.75$.

- Write down the quartiles of Z and show that the IQR is 1.349 (to 4 s.f.). [1]
- Show that the outlier fences for Z are at ± 2.698 (to 4 s.f.). [2]
- Find the probability that a value of Z is classed as an outlier. [2]
- The variable X has distribution $N(50, 8^2)$. Find the outlier fences for X . [2]
- In a random sample of 500 values from a normal distribution, estimate the number of outliers. [2]

- 28** The numbers of hours that 20 students spent on a project are:
21, 24, 25, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 38, 39, 41, 42, 44, 47, 96.
Take quartiles to be the medians of the lower and upper halves of the data.

- Find the median and the quartiles. [3]
- Show that 96 is an outlier and that it is the only outlier. [2]
- Calculate the mean and standard deviation of the data (i) including and (ii) excluding the value 96. [3]
- Comment on the effect of the outlier, and say which summary measures you would report if the outlier was a genuine value. [2]

- 29** A set of values has lower quartile $2k$ and upper quartile $6k$, where k is a positive constant. A value is an outlier if it is more than $1.5 \times \text{IQR}$ above the upper quartile or below the lower quartile.

- Show that the upper limit for outliers is $12k$ and find the lower limit. [3]
- The value 50 is an outlier and the value 42 is not. Find the range of possible values of k . [4]
- If k is an integer, find k and hence the quartiles. [2]

- 30** Fifteen house prices (in £1000s) are 20, 22, 23, 24, 24, 25, 26, 27, 28, 29, 30, 31, 33, 70, 95. Two rules for finding outliers are compared.

Rule A: more than $1.5 \times \text{IQR}$ beyond the quartiles (quartiles as medians of the halves).

Rule B: more than 2 standard deviations from the mean.

- Apply Rule A. [4]
- Apply Rule B, using the mean and standard deviation of all 15 values. [5]
- Explain why the two rules disagree about 70, and which rule is more reliable here. [2]

Coding of Data

- 1** Five values of x are 23, 27, 31, 22 and 27. They are coded using $y = x - 20$.
 (a) Write down the five coded values. [1]
 (b) Find the mean of the coded values. [1]
 (c) Hence find the mean of x . [1]
-
- 2** Data are coded using $y = \frac{x - 50}{10}$. The coded data have mean 2.4 and standard deviation 1.3. Find the mean and standard deviation of x . [3]
-
- 3** The variable x has variance 16. Find the standard deviation of $y = 3x + 5$, and the variance of $z = \frac{x}{4}$. [3]
-
- 4** The variable x has mean 6 and standard deviation 1.5. Find the mean and the standard deviation of $y = 2x + 5$. [2]
-
- 5** The heights of a group of people have mean 165 cm and standard deviation 8 cm. Find the mean and standard deviation in metres. [2]
-
- 6** For 12 values of x , the coded variable $y = \frac{x}{5}$ has $\Sigma y = 48$. Find the mean of x . [2]
-
- 7** For 20 values of x , $\Sigma(x - 100) = -35$ and $\Sigma(x - 100)^2 = 500$. Find the mean and the standard deviation of x . [4]
-
- 8** Temperatures in a town have a mean of 17.5 °C and a standard deviation of 3.2 °C. Temperatures in °F are found using $F = 1.8C + 32$.
 (a) Find the mean and standard deviation in °F. [2]
 (b) A different town has a mean of 59 °F and a standard deviation of 9 °F. Find the mean and standard deviation in °C. [2]
-
- 9** Data are coded using $y = \frac{x - a}{4}$. The coded data have mean 2.5 and standard deviation 1.8, and the mean of the original data x is 53. Find the value of a , and the standard deviation and the variance of x . [4]
-
- 10** Lengths x mm are coded using $y = \frac{x - 150}{5}$. The table shows the coded values for 35 objects. Find the mean and standard deviation of the lengths.
- | y | -2 | -1 | 0 | 1 | 3 |
|-----------|----|----|----|---|---|
| Frequency | 4 | 9 | 12 | 8 | 2 |
- [5]
-
- 11** Prices in £ have mean 25 and standard deviation 4. VAT at 20% is added to each price, so that the new price is $1.2x$. Find the mean and the variance of the new prices. [3]
-
- 12** Test scores x have mean 45 and standard deviation 6. The scores are standardised using $z = \frac{x - 45}{6}$. Find the standardised score of a student who scored 57, and the value of x that gives $z = -1.5$. [3]
-
- 13** The values x have median 14 and interquartile range 6. Find the median and the interquartile range of $y = 3x - 2$. [3]
-

- 14** The table shows the masses (class midpoints x , in grams) of 50 objects. The data are coded using $y = \frac{x - 30}{10}$.

Class midpoint x	10	20	30	40	50
Frequency f	4	11	19	12	4

- (a) Find Σfy and Σfy^2 . [2]
 (b) Find the mean of x . [2]
 (c) Find the standard deviation of x . [2]

- 15** Times x (seconds) for 25 trials are coded using $y = 4x - 30$. It is found that $\Sigma y = 250$ and $\Sigma y^2 = 3400$. Find the mean and standard deviation of x . [6]

- 16** For 10 values of x , $\Sigma(x - 30) = 25$ and $\Sigma(x - 30)^2 = 140$.
 (a) Find the mean and the variance of x . [4]
 (b) Find Σx and Σx^2 . [3]

- 17** Temperatures in $^{\circ}\text{F}$ have mean 68.9 and standard deviation 7.2. They are converted to $^{\circ}\text{C}$ using $C = \frac{5}{9}(F - 32)$.
 (a) Find the mean and the standard deviation in $^{\circ}\text{C}$, to 3 significant figures. [4]
 (b) A temperature of 83.3 $^{\circ}\text{F}$ is recorded. Express it in $^{\circ}\text{C}$ and state how many standard deviations it is above the mean. [2]

- 18** For 20 values of x , $\Sigma x = 410$ and $\Sigma x^2 = 8800$. The values are coded using $y = x - 20$.
 (a) Find Σy and Σy^2 . [3]
 (b) Hence find the variance of y , and write down the variance of x . [2]

- 19** For 40 values of x , the coded variable $y = \frac{x - 60}{5}$ has $\Sigma y = -16$ and $\Sigma y^2 = 210$. Find the mean and standard deviation of x . [5]

- 20** The lengths x metres of 150 metal rods are summarised in the table. The class midpoints are coded using $y = 50(x - 1.05)$.

Length x (m)	1.00 to 1.02	1.02 to 1.04	1.04 to 1.06	1.06 to 1.08	1.08 to 1.10
Frequency	9	31	52	38	20

- (a) Find Σfy and Σfy^2 , and hence the mean and the standard deviation of y . [3]
 (b) Hence find the mean and the standard deviation of x . [2]
 (c) Use linear interpolation to estimate the median length, and use the coefficient $3(\text{mean} - \text{median})/\sigma$ to describe the skewness. [3]

- 21** Scores x have mean 12 and standard deviation 4. They are transformed using $y = ax + b$, where $a > 0$, to give scores y with mean 100 and standard deviation 10.
 (a) Find the values of a and b . [4]
 (b) Find the transformed score for $x = 18$. [1]
 (c) A transformed score of 130 is obtained. Find the original x , and show that it is the same number of standard deviations above the mean in both scales. [2]

- 22** A set of scales reads 0.3 kg too high. Using it, the masses of 60 parcels were recorded with a mean of 4.52 kg, a standard deviation of 0.40 kg and a median of 4.45 kg.
 (a) Find the correct mean mass. [1]
 (b) State the correct standard deviation and explain why. [2]
 (c) Convert the correct mean and standard deviation to pounds, using 1 kg = 2.2 lb. [2]
 (d) Find the correct median in pounds. [2]

- 23** Two groups of values of x are summarised using $x - 50$. Group 1: $n = 20$, $\Sigma(x - 50) = 40$, $\Sigma(x - 50)^2 = 500$. Group 2: $n = 30$, $\Sigma(x - 50) = -30$, $\Sigma(x - 50)^2 = 800$.
- (a) Find the mean of x for each group. [2]
- (b) Find the mean and standard deviation of all 50 values combined. [5]
-
- 24** The variable x has mean 4 and standard deviation 2. The variable y is given by $y = 10 - 3x$.
- (a) Find the mean and variance of y . [3]
- (b) A student says "the standard deviation of y is -6 because the coefficient of x is -3 ". Explain the error and give the correct value. [2]
-
- 25** Five values y are obtained from five values x using $y = 2x - 7$. Four of the y values are 5, 9, 13 and 17, and the mean of x is 8.
- (a) Find the fifth value of y , and hence the five values of x . [4]
- (b) Find the standard deviation of x . [3]
-
- 26** A sample of n values of x has mean 6 and standard deviation 2.
- (a) Write down, in terms of n , the values of Σx and Σx^2 . [3]
- (b) Show that $\Sigma(x - 2)^2 = 20n$. [2]
- (c) Find the two possible values of the constant k for which $\Sigma(x - k)^2 = 20n$. [3]
-
- 27** The data $x_1, x_2, \dots, x_n$ have mean $\bar{x}$ and variance σ^2 . They are coded using $y = ax + b$, where a and b are constants.
- (a) Show that the mean of y is $a\bar{x} + b$. [2]
- (b) Show that the variance of y is $a^2\sigma^2$. [3]
- (c) For 40 values of x , $\Sigma x = 1260$ and $\Sigma x^2 = 41\,130$. Find the mean and standard deviation of x . [2]
- (d) Find a and b so that $y = ax + b$ (with $a > 0$) has mean 0 and standard deviation 1. [2]
-
- 28** For n values of x , $\Sigma(x - c) = 60$ and $\Sigma(x - c)^2 = 375$, where c is a constant. The mean of x is 24 and the standard deviation of x is 3.
- (a) Show that $3n^2 - 125n + 1200 = 0$. [5]
- (b) Hence find n , giving a reason for your choice. [3]
- (c) Find c . [2]
-
- 29** A variable x has mean μ and standard deviation σ . The standardised variable is $z = \frac{x - \mu}{\sigma}$.
- (a) Show that z has mean 0 and standard deviation 1. [4]
- (b) Maths marks have mean 62 and standard deviation 12; Physics marks have mean 54 and standard deviation 6. Lena scored 74 in Maths and 63 in Physics. Compare her performances using standardised scores. [4]
- (c) What Maths mark would give Lena the same standardised score as her Physics mark? [2]
-
- 30** The marks m of an examination have mean 48 and standard deviation 8. The marks are adjusted using $m' = am + b$, where $a > 0$, so that the adjusted marks have mean 60 and standard deviation 12.
- (a) Find the values of a and b . [4]
- (b) Find the mark that is unchanged by the adjustment and state which students gain and which lose marks. [3]
- (c) The maximum possible mark is 100. Find the highest original mark that stays within 100 after adjustment, and comment. [3]

Large Data Sets

- 1 In the style of the weather Large Data Set, classify each variable as qualitative, discrete quantitative or continuous quantitative.
(a) Daily mean temperature ($^{\circ}\text{C}$). (b) Daily mean total cloud cover (oktas, recorded as whole numbers from 0 to 8). (c) The name of the weather station. [3]

- 2 The table shows the daily total rainfall at a weather station for eight days. The entry "tr" means that a trace of rain was recorded, that is, less than 0.05 mm.

Day	1	2	3	4	5	6	7	8
Rainfall (mm)	0.0	tr	2.4	0.2	tr	11.6	0.0	5.0

- (a) Explain why the data cannot be used to calculate the exact total rainfall for the eight days. [1]
(b) Find the least and the greatest possible values of the total rainfall. [2]

- 3 Cloud cover is recorded in oktas, the number of eighths of the sky covered by cloud, as a whole number from 0 to 8.
(a) What fraction of the sky is covered when the reading is 6 oktas? [1]
(b) A reading of 9 oktas appears in the data. What does this show and what should be done? [2]

- 4 Part of a daily weather record is shown. Identify the day with an impossible entry and explain how you know.
(Values in this question are invented, in the style of the weather Large Data Set.)

Day	Max temp ($^{\circ}\text{C}$)	Min temp ($^{\circ}\text{C}$)
Mon	18.2	9.4
Tue	17.6	10.1
Wed	9.8	15.3
Thu	19.0	11.2

[2]

- 5 The daily total hours of sunshine at a weather station on eight days are 0.0, 3.4, 8.1, 11.6, 0.5, 6.2, 9.9 and 2.3. (Values in this question are invented, in the style of the weather Large Data Set.)
Find the mean and the median, and the percentage of these days with at least 6 hours of sunshine. [4]

- 6 A wet day is one with at least 1 mm of rain. The rainfall (mm) on 20 days is 0, 0.4, 3.2, 0, 7.5, 1.0, 0, 0.2, 12.6, 0, 1.8, 0, 0, 4.4, 0.8, 0, 2.2, 0, 0, 5.6.
Find the percentage of these days that were wet. [2]

- 7 The table shows the daily maximum temperature at a weather station on 12 consecutive days in August.

Day	1	2	3	4	5	6	7	8	9	10	11	12
Max temp ($^{\circ}\text{C}$)	17.8	19.2	21.5	20.1	18.6	22.3	23.0	19.9	18.4	20.7	24.1	21.9

- (a) Calculate the mean daily maximum temperature. [2]
(b) Find the median. [1]
(c) A student says these 12 days give a good estimate of the mean daily maximum temperature for May to October. Give a reason why this is not true. [1]

- 8** An extract of daily mean temperatures from a data set is shown. "n/a" means that no reading is available.

Day	1	2	3	4	5	6	7
Mean temp (°C)	14.3	n/a	15.1	61.2	13.9	16.0	n/a

- (a) Explain how the "n/a" entries should be treated when finding the mean temperature. [1]
 (b) Identify an entry that is almost certainly an error, and say how you would deal with it. [2]
 (c) Calculate the mean of the valid temperatures. [1]

- 9** The daily mean temperatures (°C) on ten randomly chosen days in a month are 16.4, 17.9, 15.2, 18.8, 19.5, 14.7, 16.0, 17.3, 18.1, 15.9. The mean for the whole month is known to be 16.8 °C. (Values in this question are invented, in the style of the weather Large Data Set.)

- (a) Calculate the mean of the sample. [2]
 (b) Comment on the difference between the sample mean and the mean for the month. [1]

- 10** A student wants a simple random sample of 4 days from the 31 days of July, numbered 1 to 31. She reads the two-digit random numbers 17, 42, 08, 17, 31, 29 in order. Write down the days chosen and explain why choosing the first four days of July might be unsuitable. [4]

- 11** In the style of the Large Data Set, the mean daily maximum temperature in summer is 17.8 °C (standard deviation 3.1 °C) at a coastal station and 19.4 °C (standard deviation 4.6 °C) at an inland station. Compare the stations and give one reason for the difference. [3]

- 12** The cloud cover (oktas) on 40 days is summarised in the table. Find the mode, the median and the mean. (Values in this question are invented, in the style of the weather Large Data Set.)

Oktas	0	1	2	3	4	5	6	7	8
Days	1	0	2	3	2	3	6	9	14

[4]

- 13** The maximum temperatures on 36 days are classified as: below 15 °C: 6 days; 15 °C to under 20 °C: 15 days; 20 °C to under 25 °C: 10 days; 25 °C or above: 5 days. A pie chart is drawn. Find the angle for each class and the percentage of days with a maximum of at least 20 °C. (Values in this question are invented, in the style of the weather Large Data Set.) [4]

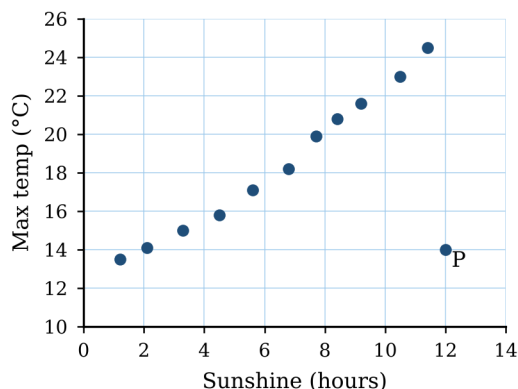
- 14** A student takes a systematic sample from the 184 days from 1 May to 31 October (numbered 1 to 184). She chooses day 3 and then every 7th day after it.

- (a) Find the size of the sample. [2]
 (b) Find the date of the last day in her sample. [2]
 (c) Give two criticisms of using this sample to describe the weather of the whole region over many years. [2]

- 15** The daily mean temperatures at two weather stations over the same six months have these summaries: Station A: mean 15.3 °C, standard deviation 3.1 °C. Station B: mean 13.8 °C, standard deviation 2.4 °C. On one day station A recorded 21.0 °C and on another day station B recorded 18.5 °C.

- (a) Find how many standard deviations each reading is above the mean of its own station. [3]
 (b) State which reading is more unusual for its station. [1]
 (c) A student claims that station A is always warmer than station B. Use the summaries to explain why this claim is not justified. [2]

- 16** The scatter diagram shows the daily hours of sunshine and the maximum temperature ($^{\circ}\text{C}$) on 12 days. (Values in this question are invented, in the style of the weather Large Data Set.)



- (a) Describe the correlation between sunshine and maximum temperature, ignoring point P. [2]
- (b) Point P is the day with 12 hours of sunshine. Suggest why P does not follow the trend, and what should be done with it. [2]
- (c) Explain why the correlation does not prove that sunshine causes the higher temperatures. [1]
-
- 17** The daily maximum temperatures t ($^{\circ}\text{C}$) for the 31 days of July are summarised by $\Sigma(t - 20) = 62$ and $\Sigma(t - 20)^2 = 412$. (Values in this question are invented, in the style of the weather Large Data Set.)
- (a) Find the mean and standard deviation of the daily maximum temperature in July. [4]
- (b) In September the mean was 18.2°C with standard deviation 2.1°C . Compare September with July. [2]

- 18** The 184 days from May to October are classified by wind direction and by whether rain was recorded. (Values in this question are invented, in the style of the weather Large Data Set.)

Wind direction	Rain	No rain
Westerly	38	46
Easterly	6	40
Other	14	40

- (a) Find the probability that a randomly chosen day had rain. [1]
- (b) Find the probability that a randomly chosen westerly day had rain. [2]
- (c) Hence state whether rain and westerly wind are independent. [2]
-
- 19** The visibility (km) over 90 days has lower quartile 9.0, median 14.0 and upper quartile 21.0. The two greatest values are 48.5 and 52.0. (Values in this question are invented, in the style of the weather Large Data Set.)
- (a) Show that both of these values are outliers. [3]
- (b) Give a possible reason for such high values, and say whether they should be removed. [2]
-
- 20** A student uses the mean daily maximum temperature for May to October of one year at one station (16.2°C) and compares it with a published long-term average of 15.5°C , concluding that "summers are getting warmer".
- (a) Find the percentage by which the student's mean exceeds the long-term average. [2]
- (b) Give two reasons why the conclusion is not justified. [2]

- 21** The daily total rainfall (mm) at a weather station on 20 days is:
0, 0, 0, 0, 0, 0, 0, 0, 0, 0.2, 0.4, 0.8, 1.2, 2.0, 3.4, 5.6, 7.8, 9.2, 14.6, 21.0.
(The nine zeros include the entries recorded as "tr", which are taken as 0.) Take quartiles to be the medians of the lower and upper halves of the data.
- Find the median and the quartiles. [3]
 - Use the $1.5 \times \text{IQR}$ rule to identify the outliers. [2]
 - Explain what $Q_1 = 0$ tells you in this context. [1]
 - Explain why a box plot is not a very informative summary of these data, and suggest one extra statistic that would help. [2]

- 22** The table classifies 184 days (May to October) at a weather station by whether rain was recorded and by the daily mean temperature.

	Rain	No rain
Mean temp $\geq 15^\circ\text{C}$	41	79
Mean temp $< 15^\circ\text{C}$	33	31

- Find the probability that a day chosen at random had rain. [1]
- Find the probability that a day had rain given that its mean temperature was below 15°C . [2]
- Hence state, with a reason, whether rain and a cool day appear to be independent. [2]
- Two different days are chosen at random from the 184. Find the probability that both had rain. [2]
- Give one reason why conclusions from these data about the whole year would be unreliable. [1]

- 23** A stratified sample of 36 days is taken from the 184 days from May to October (31, 30, 31, 31, 30, 31 days), using the months as strata. (Values in this question are invented, in the style of the weather Large Data Set.)

- Show that the sample should contain 6 days from each month, to the nearest whole number. [3]
- Describe how to choose the 6 days for May. [2]
- Explain why a systematic sample of 36 days cannot use a whole-number interval. [1]

- 24** The visibility v (km) on 184 days is summarised below; on 16 further days the entry is n/a (missing). (Values in this question are invented, in the style of the weather Large Data Set.)

Visibility v (km)	Days
$0 \leq v < 10$	28
$10 \leq v < 20$	71
$20 \leq v < 40$	56
$40 \leq v < 60$	13

- Show that the total number of days with recorded visibility is 168. [1]
- Estimate the mean visibility. [3]
- Explain why the missing days could bias the estimate, and say how the missing entries should be treated. [2]

- 25** A student summarises the rainfall at a station in the style of the Large Data Set (invented values): May had 31 days, 9 wet days (at least 1 mm) and total rainfall 62.4 mm; October had 31 days, 14 wet days and total rainfall 118.9 mm.

- Find, for each month, the mean daily rainfall and the mean rainfall per wet day (to 2 decimal places where needed). [4]
- The student claims that October is wetter than May. Comment, giving two reasons that the claim may not generalise. [3]

- 26** The mean daily wind speeds (knots) over the 92 days of summer have $\Sigma x = 1012$ and $\Sigma x^2 = 13100$. One knot is 1.15 miles per hour. (Values in this question are invented, in the style of the weather Large Data Set.)

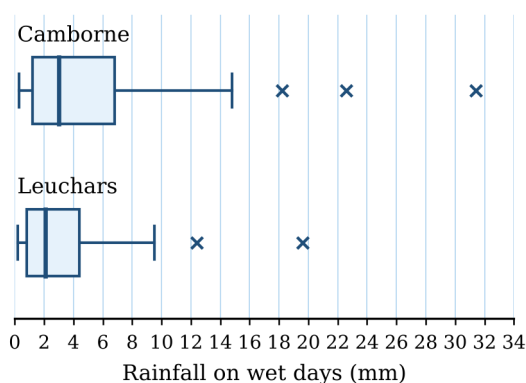
- Find the mean and standard deviation in knots. [3]
- Find the mean and standard deviation in miles per hour. [2]
- A day with a mean wind speed of 24 knots occurs. How many standard deviations above the mean is this? [2]

- 27** The daily mean wind speeds, in knots, at station A over the 92 days of June to August are summarised in the table. For the same 92 days at station B the mean was 7.5 knots and the standard deviation was 3.6 knots.

Mean wind speed (knots)	0-4	4-8	8-12	12-20
Days (station A)	10	32	30	20

- Estimate the mean and standard deviation of the wind speed at station A, in knots. [4]
- Given that 1 knot is approximately 1.15 miles per hour, convert both answers to mph. [2]
- Find the mean and standard deviation of the wind speed for all 184 station-days of the two stations, in knots. [4]

- 28** A student compares the rainfall at Camborne and Leuchars from May to October (184 days). Camborne had 78 wet days and Leuchars had 61 wet days. The box plots show the rainfall, in mm, on the wet days only. Outliers are marked with crosses.



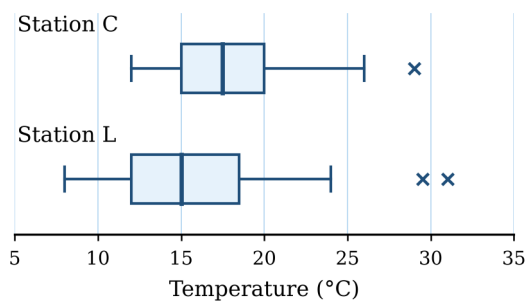
- Find the percentage of days that were wet at each place. [2]
- Compare the rainfall on wet days at the two places. [2]
- For Leuchars, show that the upper whisker should end at the largest value below the upper fence, and that 12.4 is an outlier. [2]
- Explain why the mean rainfall on wet days would be a less suitable way to compare the places than the median. [1]
- Give two reasons why these data cannot be used to say which place is wetter over a whole year. [2]

- 29** For 183 consecutive pairs of days (yesterday, today), the table shows whether rain was recorded. (Values in this question are invented, in the style of the weather Large Data Set.)

Yesterday	Today: rain	Today: no rain
Rain	45	33
No rain	32	73

- Find the probability of rain today, and the probability of rain today given rain yesterday. [4]
- Find the probability of rain today given no rain yesterday and hence comment on whether the weather on successive days is independent. [3]
- Using $P(\text{rain}) = 0.42$ on any day, assume independence and find the probability that it rains on each of 7 consecutive days. Then, using the conditional probability from (a) for each day after the first, find a second answer, and comment. [5]

- 30** The box plots show the daily maximum temperatures ($^{\circ}\text{C}$) at two stations over the same six months, in the style of the Large Data Set. A student says: "Station C is warmer than Station L because its mean is 2.6°C higher, so Station C is always warmer."



- (a) Verify that the plotted outliers for Station L are correct, given that its whiskers end at 8 and 24. [3]
- (b) Compare the distributions using medians and interquartile ranges. [3]
- (c) Evaluate the student's conclusion and suggest what extra analysis would make it more reliable. [3]