Data, sampling and graphs

SL AA · Statistics and probability · 4.1 · 4.2

Before any calculation, statistics asks two questions: where do the data come from, and what do they look like? A sample stands for a population only if it is chosen well; a graph shows at a glance the centre, the spread and the values that do not fit.

I POPULATIONS, SAMPLES AND DATA

SL 4.1 · population, sample, discrete and continuous data, sampling techniques

I.1 Population, sample, type of data

Definition 1 *population and sample*

The **population** is the whole set studied; a **sample** is the part actually measured. A sample is **biased** when it systematically over- or under-represents part of the population. **Discrete** data are counted; **continuous** data are measured.

Example 1

A factory has 500 workers. 30 of them are asked their journey time to work and their number of children. Name the population and the sample, and the type of each variable.

WHAT YOU DO	WHAT YOU WRITE
Step 1 Population: everyone studied	the 500 workers
Step 2 Sample: those asked	the 30 workers questioned
Step 3 Counted or measured?	journey time: continuous; number of children: discrete

APPLY

1 The mass of an apple: discrete or continuous?

2 The number of goals in a match?

3 A shoe size (whole and half sizes)?

I.2 Sampling techniques

Property 1 *sampling techniques*

Simple random: every sample of that size equally likely. **Systematic**: every kth member of a list, from a random start. **Stratified**: each group sampled in proportion to its size. **Quota**: a fixed number from each group, not at random. **Convenience**: whoever is easiest to reach.

Example 2

Name the sampling technique: (i) a teacher takes every 5th name on the register, starting at a random name; (ii) a reporter asks 10 men and 10 women, whoever passes first.

WHAT YOU DO	WHAT YOU WRITE
Step 1 (i) every kth, random start	systematic
Step 2 (ii) fixed numbers per group, not random	quota

APPLY

4 Names drawn from a hat.

5 The first 20 people to arrive at the gym.

I.3 The sizes in a stratified sample

Watch out 1

From each stratum take $\dfrac{\text{size of the stratum}}{\text{size of the population}} \times$ sample size, rounded sensibly. A convenience sample is the most likely to be biased: say who is left out.

Example 3

A club has 450 members: 270 seniors and 180 juniors. A stratified sample of 30 is taken. How many seniors and juniors?

WHAT YOU DO	WHAT YOU WRITE
Step 1 Fraction of the population, times the sample size	seniors: $\dfrac{270}{450} \times 30 = 18$
Step 2 The other stratum	juniors: $\dfrac{180}{450} \times 30 = 12$
Step 3 Check the total	$18 + 12 = 30$

APPLY

6 1000 people: 350 under 30, 650 aged 30 or over; a stratified sample of 40.

II GRAPHS OF DATA

SL 4.2 · histograms, cumulative frequency, box-and-whisker diagrams, outliers

II.1 A cumulative frequency table

Definition 2 *cumulative frequency*

The **cumulative frequency** of a value is the number of data $\leq$ it; its graph goes through (upper class boundary, cumulative frequency). For n data: the **median** at $\frac{n}{2}$, Q_1 at $\frac{n}{4}$, Q_3 at $\frac{3n}{4}$.

Example 4

The masses m g of 40 eggs: $m \leq 50$: 4; $m \leq 55$: 12; $m \leq 60$: 27; $m \leq 65$: 36; $m \leq 70$: 40. How many eggs have $55 < m \leq 60$? Which class contains the median? What percentage weigh more than 65 g?

WHAT YOU DO	WHAT YOU WRITE
Step 1 Subtract two cumulative frequencies	$27 - 12 = 15$ eggs
Step 2 Median: the 20th egg	$12 < 20 \leq 27$: the class $55 < m \leq 60$
Step 3 More than 65: from the total	$40 - 36 = 4$, $\dfrac{4}{40} = 10\,\%$

APPLY

1 Same eggs: how many weigh at most 60 g?

2 Same eggs: the percentage that weigh more than 55 g.

II.2 Median and quartiles from the graph

Example 5

From the cumulative frequency graph of the 40 eggs of Example 4 (points joined by straight segments), estimate the median and the interquartile range.

WHAT YOU DO	WHAT YOU WRITE
Step 1 Positions on the vertical axis	median 20, Q_1 10, Q_3 30
Step 2 Across to the graph, down: median	$55 + 5 \times \dfrac{20 - 12}{15} = 57.7$ g
Step 3 The quartiles the same way	$Q_1 = 53.75$, $Q_3 = 61.7$
Step 4 IQR $= Q_3 - Q_1$	7.92 g

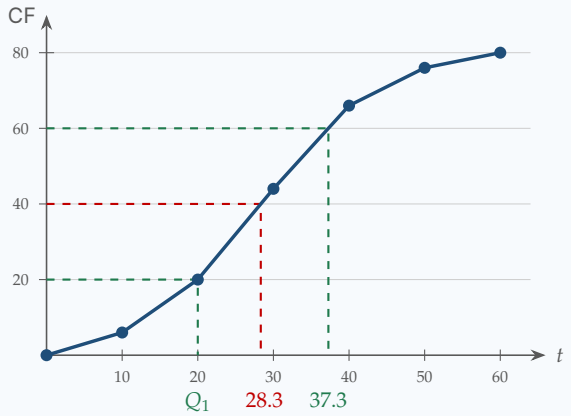


Figure 1 — Exam-style question 2's 80 times: across from 40, down to the median, about 28.3 minutes; across from 20 and 60, the quartiles.

APPLY

3 120 data: where is Q_3 read on the vertical axis?

4 What percentage of the data lies below the median?

II.3 Outliers

Property 2 *box-and-whisker diagram and outliers*

A **box-and-whisker diagram** shows the minimum, Q_1, the median, Q_3, the maximum; IQR $= Q_3 - Q_1$. An **outlier** is more than $1.5 \times$ IQR below Q_1 or above Q_3; the whisker stops at the last value that is not an outlier, the outlier is a cross.

Watch out 2

The fences use the IQR, not the range, and a whisker ends at a data value, never at a fence.

Example 6

A data set has minimum 2, $Q_1 = 15$, median 20, $Q_3 = 23$ and maximum 40. Are the minimum and the maximum outliers?

WHAT YOU DO	WHAT YOU WRITE
Step 1 IQR	$23 - 15 = 8,\ 1.5 \times 8 = 12$
Step 2 The two fences	$15 - 12 = 3$ and $23 + 12 = 35$
Step 3 Compare, in words	$2 < 3$ and $40 > 35$: both are outliers

APPLY

5 $Q_1 = 30$, $Q_3 = 42$, maximum 60: is 60 an outlier?

6 $Q_1 = 11$, $Q_3 = 19$, minimum 0: is 0 an outlier?

II.4 A box plot from a list

Example 7

Draw the box plot of $4,\ 6,\ 9,\ 9,\ 11,\ 12, 13,\ 15,\ 16,\ 16,\ 18,\ 31$.

WHAT YOU DO	WHAT YOU WRITE
Step 1 Median of the 12 values	$\dfrac{12+13}{2} = 12.5$
Step 2 Quartiles: medians of each half	$Q_1 = 9,\ Q_3 = 16$
Step 3 Fences with IQR $= 7$	$9 - 10.5 = -1.5,\ 16 + 10.5 = 26.5$: 31 is an outlier
Step 4 Draw	box 9 to 16, line at 12.5, whiskers 4 to 18, a cross at 31

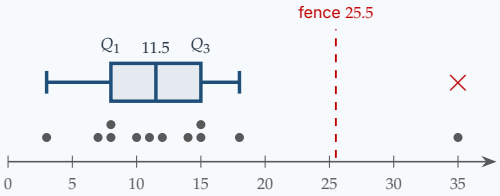


Figure 2 — Exam-style question 4's twelve data: $Q_1 = 8$, median 11.5, $Q_3 = 15$, upper fence 25.5: 35 is an outlier.

APPLY

7 $2,\ 5,\ 5,\ 7,\ 8, 9,\ 9,\ 10,\ 10,\ 22$: Q_1, Q_3 and any outlier.

III EXAM-STYLE QUESTIONS

from the shortest to the longest · every one in steps, with its marks

A Sampling

EXAM-STYLE QUESTION 1 Paper 1 · no calculator · [Maximum mark: 5]

A school has 600 students: 240 in Year 1 and 360 in Year 2. A sample of 50 students is asked how long they spend on homework each evening.

(a) A stratified sample by year group is used. Find the number of Year 1 students and the number of Year 2 students in the sample. [2]

(b) State whether the time spent on homework is discrete or continuous data. [1]

(c) Another method chooses every 12th student from an alphabetical list, starting at a random position. Name this sampling technique. [1]

(d) Give one reason why asking the first 50 students who arrive at school could give a biased sample. [1]

WORKING (a)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Fraction of the population, times the sample size	$\frac{240}{600} \times 50 = 20$ Year 1	A1
Step 2 The other stratum	30 Year 2	A1

WORKING (b)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Counted or measured?	continuous *A time is measured, not counted.*	A1

WORKING (c)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Every kth, random start	systematic sampling	A1

WORKING (d)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Say who is left out	early arrivals may differ from the others (for example, they live close, or work more) *Any reason that names who is over- or under-represented.*	R1

EXAM-STYLE QUESTION 2

Paper 1 · no calculator · [Maximum mark: 5]

The time, t minutes, that 80 students took to finish a puzzle is summarised below.

$t \leq$	10	20	30	40	50	60
cumulative frequency	6	20	44	66	76	80

(a) Find the number of students who took more than 20 minutes and at most 30 minutes. [1]

(b) Write down the modal class. [1]

(c) Write down the class that contains the median. [1]

(d) Find the percentage of the students who took more than 40 minutes. [2]

WORKING (a)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Subtract two cumulative frequencies	$44 - 20 = 24$	A1

WORKING (b)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 The largest class frequency	$20 < t \leq 30$ *Frequencies 6, 14, 24, 22, 10, 4.*	A1

WORKING (c)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 The 40th student	$20 < t \leq 30$ *The 40th and 41st students: cumulative frequency* $20 < 40 \leq 44$.	A1

WORKING (d)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 More than 40: from the total	$80 - 66 = 14$	(M1)
Step 2 Percentage	$\frac{14}{80} = 17.5\,\%$	A1

C Outliers

EXAM-STYLE QUESTION 3 Paper 1 · no calculator · [Maximum mark: 6]

For a set of data, the minimum is 5, the lower quartile is 12, the median is 18, the upper quartile is 24 and the maximum is 45.

(a) Find the interquartile range. [1]

(b) Show that 45 is an outlier. [2]

(c) Determine whether 5 is an outlier. [2]

(d) Write down the range. [1]

WORKING (a)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 $Q_3 - Q_1$	$24 - 12 = 12$	A1

WORKING (b)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Upper fence $Q_3 + 1.5\,\text{IQR}$	$24 + 1.5 \times 12 = 42$	M1
Step 2 Compare, in words	$45 > 42$, so 45 is an outlier	A1

The fence and the comparison are both needed.

WORKING (c)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Lower fence $Q_1 - 1.5\,\text{IQR}$	$12 - 1.5 \times 12 = -6$	M1
Step 2 Compare, in words	$5 > -6$: not an outlier	A1

WORKING (d)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Maximum minus minimum	$45 - 5 = 40$	A1

D A list on the GDC (Paper 2)

EXAM-STYLE QUESTION 4 Paper 2 · calculator · [Maximum mark: 6]

The numbers of books read in one year by 12 students are

$$3,\ 7,\ 8,\ 8,\ 10,\ 11,\ 12,\ 14,\ 15,\ 15,\ 18,\ 35.$$

(a) Find the mean and the standard deviation of these data. [2]
(b) Write down the lower quartile and the upper quartile. [2]
(c) Show that 35 is an outlier. [2]

WORKING (a)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 One-variable statistics on the GDC	mean 13	A1
	standard deviation 7.74 $(7.73520\ldots)$	A1
	The population standard deviation σ.	

WORKING (b)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Read the quartiles	$Q_1 = 8$	A1
	$Q_3 = 15$	A1

WORKING (c)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Upper fence	$15 + 1.5 \times 7 = 25.5$	M1
Step 2 Compare	$35 > 25.5$	A1
	So 35 is an outlier.	

IV WHICH METHOD? READ THE QUESTION

4.2 · the whole topic on one page

THE QUESTION SAYS	YOU KNOW	YOU WRITE
"stratified sample"	proportional to each group	$\frac{\text{group}}{\text{population}} \times$ sample size P1
"name the sampling technique"	how the members were chosen	random, systematic, stratified, quota or convenience P1
"discrete or continuous"	counted or measured	one word, with the reason if asked D1
"estimate the median" from a cumulative graph	$\frac{n}{2}$ on the vertical axis	across to the curve, down to the value D2
"the interquartile range"	$Q_3 - Q_1$	quartiles at $\frac{n}{4}$ and $\frac{3n}{4}$ P2
"is it an outlier?"	$Q_1 - 1.5\,\text{IQR}$, $Q_3 + 1.5\,\text{IQR}$	the fence, then the comparison in words P2
"draw a box-and-whisker diagram"	five values, outliers apart	a scale; the whisker to the last value inside the fences P2

In the exam 1 *command terms*

State: one word or phrase, no working. **Show that**: the fence computed and compared. **Estimate**: read from a graph, so not exact. **Determine**: decide, with the evidence written.

V ON THE GDC: PAPER 2

SL 4.1 · 4.2 · NumWorks · every calculator move of this topic

1 Statistics of a list

Statistics → Data: V1 → Stats

Enter the list in V1 (and the frequencies in N1 if there are any), then Stats. **On paper:** the values you use, $Q_1 = 8$, $Q_3 = 15$, and the fence.

STATISTICS · STATS

```
V1: 3,7,8,8,10,11,
12,14,15,15,18,35
mean 13 sigma 7.735201
min 3 Q1 8 Med 11.5 Q3 15 max 35
```

2 A box plot on the GDC

Statistics → Graph → Box

The Box view draws the diagram from the list. **On paper:** draw it on a scale, with the outlier as a cross.

STATISTICS · BOX

```
min 3 Q1 8 Med 11.5
Q3 15 max 35
outlier shown: 35
```

ANSWERS TO THE APPLY QUESTIONS

I.1 continuous

I.2 discrete

I.3 discrete

I.4 simple random

I.5 convenience

I.6 14 and 26

II.1 27

II.2 70 %

II.3 at 90

II.4 50 %

II.5 no: the fence is 60, and 60 is not more than it

II.6 no: the fence is -1

II.7 5; 10; 22 (fence 17.5)

Written and verified by a practising IB mathematics teacher, produced with AI assistance; every result recomputed by software, every figure computed from its function. Not affiliated with or endorsed by the International Baccalaureate Organization. © 2026 Hence