Pearson Edexcel Level 1/Level 2 GCSE (9–1)

Statistics — Paper 1: Higher Tier (1ST0/1H)

Examination Date: Tuesday 2 June 2026

Total Marks: 80 Marks

Question 1 (Total: 4 Marks)

(a) Advantage of using the internet to source data

Worked Solution: Sourcing data from the internet is quick, free, and gives immediate access to a vast, pre-existing nationwide database (secondary data) without needing to conduct expensive primary fieldwork.

> **Mark Scheme:**
> **B1:** Any valid advantage such as saving time/cost, large data availability, or access to official database sources.
>
> 1 Mark

(b) Statistical name of the diagram

Worked Solution: This diagram is a **Choropleth map**.

> **Mark Scheme:**
> **B1:** Choropleth (map).
>
> 1 Mark

(c) Suggest two improvements to the diagram

Worked Solution:

1. Add a clear title to explain what the map shows (e.g. "Percentage of UK Population Employed in Agriculture, Forestry, and Fishery by Region").
2. Clearly define the geographic boundaries or label key regions to assist the reader in identifying specific UK territories.

> **Mark Scheme:**
> **B2:** 1 mark per valid improvement (e.g. add title, label regions/map compass, state units in key, use more distinct shading/colors).
>
> 2 Marks

Question 2 (Total: 5 Marks)

(a) Percentage of petrol cars sold in Germany

Worked Solution: Looking at the "Germany" bar on the percentage composite bar chart:

- The bottom white segment (Petrol) starts at 0% and reaches 18%.

- Thus, the percentage is **18%** (allow responses in the range 17% to 19%).

> **Mark Scheme:** **2 Marks**
> **M1:** For reading the value of the bottom segment boundary correctly.
> **A1:** 18% (accept 17% - 19%).

(b) Compare fuel type percentages between countries

Worked Solution:

- **Petrol:** Visually much higher in Germany (approx. 18%) than in Great Britain (approx. 5%).
- **Diesel:** Great Britain sold a higher proportion of diesel cars (48% - 5% = 43%) compared to Germany (50% - 18% = 32%).
- **Electric:** Both countries sold a very similar proportion of electric vehicles, with Great Britain slightly higher at 52% (100% - 48%) compared to Germany's 50% (100% - 50%).

> **Mark Scheme:** **2 Marks**
> **B1:** Points out that Germany has a higher percentage of petrol cars than GB.
> **B1:** Points out that Great Britain has a higher percentage of diesel (or electric) cars than Germany.

(c) Why a time-series graph is inappropriate

Worked Solution: A time-series graph displays values over successive periods of time (longitudinal data) to observe trends, whereas this dataset only covers a single year (2022) across two distinct locations (cross-sectional comparative data).

> **Mark Scheme:** **1 Mark**
> **B1:** Explains that there is only one time period (single year 2022), making a trend analysis impossible.

Question 3 (Total: 6 Marks)

Critique of Grace's plans for data collection, processing, and presentation:

1. Collecting (Sampling and Interviewing):

- **Flaw:** Grace states she will select every 14th student to get a 10% sample of 1400 (which is 140 students). However, she does not mention selecting the first student using a random starting point between 1 and 14, which is required for systematic sampling to be unbiased.
- **Flaw:** Asking students "face-to-face" how long it takes them to get to school can introduce response bias (students may feel pressured or guess wildly) or interviewer bias. An anonymous digital survey would be more reliable.
- **Flaw:** Grace's plan aims to investigate if "average distance" and "average time" have increased, but she only collects "time taken" (omitting distance measurements).

2. Processing (Grouping Data):

- **Flaw:** The class intervals are overlapping and ambiguous. The boundaries $0 \le t \le 5$ and $5 \le t < 10$ both include the value of exactly 5 minutes. Boundaries should be mutually exclusive (e.g. $0 \le t < 5$).
- **Flaw:** The final class $t \ge 15$ is open-ended. Open-ended groups prevent accurate calculation of the mean (since midpoint cannot be determined) and make cumulative frequency plotting impossible without guessing an upper limit.

3. Presenting:

- **Flaw:** To draw a cumulative frequency curve, she needs coordinates plotted at the upper boundaries of each interval. Since the final interval $t \ge 15$ has no upper boundary, the curve cannot be closed correctly. She must define a closed limit (e.g. $15 \le t < 30$).

> **Mark Scheme:** 6 Marks
> **B2:** Cites collection flaws (lack of random starting point, face-to-face bias, or missing distance variable).
> **B2:** Cites processing flaws (overlapping boundaries at 5, open-ended interval $t \ge 15$).
> **B2:** Cites presentation flaws (cannot plot upper limit of cumulative frequency for open class) and suggests suitable fixes (e.g. rewriting intervals to be mutually exclusive).

Question 4 (Total: 9 Marks)

(a) Complete the Bergen table

Worked Solution: Bergen has a total of 30 days in September 2023.

- Lowest temperature: The cumulative frequency curve starts on the x-axis at **6 °C**.
- Lower Quartile (Q_1) (at CF = $30 / 4 = 7.5$): Reading from the curve at CF = 7.5 yields **10 °C**.
- Median (Q_2) (at CF = $30 / 2 = 15$): **12 °C** (already provided).
- Upper Quartile (Q_3) (at CF = *30 imes 0.75 = 22.5*): Reading from the curve at CF = 22.5 yields **14 °C**.
- Highest temperature: The curve terminates on the right at CF = 30 at **18 °C**.

City	Lowest	Lower Quartile	Median	Upper Quartile	Highest
Bergen	6	10	12	14	18

> **Mark Scheme:** 2 Marks
> **M1:** Attempts to read quartile values from CF = 7.5 and CF = 22.5.
> **A1:** Completes the table correctly: Lowest = 6 (or 7), LQ = 10, UQ = 14, Highest = 18.

(b) Draw box plot on the grid

Worked Solution: Sketch a standard horizontal box plot using the five-number summary from (a):

- Left whisker starts at 6.

- Left edge of the box is at 10.
- Vertical line in the box (Median) is at 12.
- Right edge of the box is at 14.
- Right whisker terminates at 18.

> **Mark Scheme:** **2 Marks**
> **M1:** Plots at least 3 values from their table on the scale correctly.
> **A1:** Draws a fully accurate, neat box plot.

(c) Compare Bergen and Hammerfest distributions

Worked Solution:

1. **Average (Median):** Bergen has a higher median temperature (12 °C) compared to Hammerfest (6 °C), showing that Bergen was warmer on average in September 2023.
2. **Spread (IQR):** Bergen's IQR is smaller (*14 - 10 = 4 ext{ °C}*) than Hammerfest's IQR (*11 - 4 = 7 ext{ °C}*). This indicates that the daily minimum temperatures in Bergen were more consistent (less variable).
3. **Range:** Bergen's range (*18 - 6 = 12 ext{ °C}*) is smaller than Hammerfest's range (*17 - 3 = 14 ext{ °C}*).

> **Mark Scheme:** **5 Marks**
> **B3:** Provides three distinct numerical comparisons (Median, IQR, Range).
> **B2:** Provides correct contextual interpretations of at least two comparisons (e.g. "Bergen is warmer on average", "Bergen has more consistent temperatures").

Question 5 (Total: 9 Marks)

(a) Name of the sampling method

Worked Solution: This method is **Cluster sampling**.

> **Mark Scheme:** **1 Mark**
> **B1:** Cluster (sampling).

(b) Two advantages of this sampling method

Worked Solution:

1. **Cost and Time efficiency:** It is far cheaper and faster to travel to and survey all students in just 2 randomly selected districts than to travel to schools spread out across all 5 districts.
2. **Admin Simplicity:** Gathering data from entire schools is simpler because student registers are already collated at the school level.

> **Mark Scheme:** **2 Marks**
> **B2:** 1 mark per valid advantage (practical convenience, cost-effective, time-saving, admin ease).

(c) Why this is not a random sample

Worked Solution: In a truly random sample, every student in the city population must have an equal chance of being selected. Since three of the five districts were excluded completely, students in those districts had a 0% chance of selection.

> **Mark Scheme:** **1 Mark**
> **B1:** Explains that students in unselected districts have zero probability of selection.

(d) Assess the validity of Amy's conclusions

Worked Solution:

1. **Conclusion 1 (Invalid):** Standard pie charts only display *proportions* (percentages) of categories, not absolute frequencies. Without knowing the total student count of School A and School B, we cannot assume that a larger Rugby slice in School A represents more actual students than in School B.
2. **Conclusion 2 (Valid):** Football occupies the largest single sector area (highest angle/proportion) in both School A and School B's pie charts. Because each pie chart represents a closed population, the largest sector always represents the most attended club.

> **Mark Scheme:** **2 Marks**
> **B1:** Concludes Conclusion 1 is invalid with reference to proportions vs counts.
> **B1:** Concludes Conclusion 2 is valid with reference to largest slice/angle.

(e) Calculate total students in School B

Worked Solution: In comparative pie charts, the total area of the circle is directly proportional to the total population size:

rac{ ext{Area}_B}{ ext{Area}_A} = rac{ ext{Total}_B}{ ext{Total}_A} \implies rac{\pi (r_B)^2}{\pi (r_A)^2} = rac{ ext{Total}_B}{ ext{Total}_A}

rac{3.6^2}{3.0^2} = rac{ ext{Total}_B}{250} \implies rac{12.96}{9.0} = rac{ ext{Total}_B}{250}

ext{Total}_B = 1.44 imes 250 = 360 ext{ students}

> **Mark Scheme:** **2 Marks**
> **M1:** For correct use of ratio of squared radii: r_B^2 / r_A^2 = 12.96 / 9.
> **A1:** 360 students.

(f) Why comparative pie charts are more appropriate

Worked Solution: Standard pie charts only display proportions. Comparative pie charts adjust circle size proportionally to total frequencies, allowing simultaneous comparison of both proportions and absolute counts between different populations.

Mark Scheme:
B1: States that comparative charts represent size/frequency differences via circle area.

1 Mark

Question 6 (Total: 16 Marks)

(a) Moving Average and Trend Line

- **(i) Calculate first 4-point moving average:**
  Sum of first 4 values (Q3 2016 to Q2 2017):

$$ext\{Sum\} = 1672 + 1034 + 895 + 1491 = 5092 \ ext\{ thousand\}$$

$$ext\{Moving\ Average\} = rac\{5092\}\{4\} = 1273.0 \ ext\{ thousand\}$$

- **(ii) Plot on graph:** Plot at 1273.0 thousand. Midpoint of the 4 quarters is between Q4 2016 and Q1 2017.
- **(iii) Trend line:** Draw a straight line of best fit through the plotted moving averages.

Mark Scheme:
M1: Computes sum = 5092.
A1: Arrives at average = 1273.0.
B1: Correctly plots point at midpoint on the x-axis.
B1: Draws a suitable line of best fit through the points.

4 Marks

(b) Describe and interpret the trend

Worked Solution: The moving averages are increasing steadily over time (from 1273.0 to 1333.25). This shows an overall upward trend, indicating that passenger numbers using Newcastle Airport generally grew over the period from 2016 to 2018.

Mark Scheme:
B1: Identifies upward trend over the period.
B1: Interprets the trend contextually (passenger growth).

2 Marks

(c) Seasonal Variation

- **(i) Two examples of seasonal variation:**
  1. Consistently high passenger volumes peaking in Quarter 3 (summer holidays) of each year.
  2. Consistently low passenger volumes dropping to their lowest point in Quarter 1 (winter period) of each year.
- **(ii) Contextual interpretation:** Quarter 3 peaks are due to school summer holidays when families travel abroad. Quarter 1 drops represent post-holiday seasonal winter lulls.

Mark Scheme:
B2: Identifies peak in Q3 and trough in Q1.
B1: Contextual explanation of holiday seasonal travel patterns.

3 Marks

(d) Why Ronan's -150k seasonal effect is wrong

Worked Solution: Quarter 3 is the peak quarter every year (e.g., 1672 in Q3 2016 vs trend ~1273; 1811 in Q3 2017 vs trend ~1329). Since actual passenger values in Q3 are consistently much higher than the average trend line, the seasonal effect for Q3 must be a **positive value**, not negative.

> **Mark Scheme:** **2 Marks**
> **M1:** Points out that Q3 actuals are always above the trend lines.
> **A1:** Correctly states that the seasonal effect must be positive.

(e) Estimate passengers for Quarter 1, 2019

Worked Solution:

1. **Estimate Trend for Q1 2019:** Extrapolating the trend line (which increases by ~5 thousand per quarter on average):

   ext{Trend Value for Q1 2019} pprox 1335 ext{ thousand}

2. **Calculate Average Seasonal Effect for Q1:**
   - Q1 2017*ext{Actual}=895ext{Trend(centered)} pprox 1290. ext{Effect} = 895 - 1290 = -395.*
   - Q1 2018*ext{Actual}=912ext{Trend(centered)} pprox 1334. ext{Effect} = 912 - 1334 = -422.*
   - Average Q1 Effect: *rac{-395+-422}{2}pprox -409 ext{ thousand}.*

3. **Estimate Value:**

   ext{Estimate} = ext{Trend} + ext{Seasonal Effect} = 1335 + (-409) = 926 ext{ thousand}

   (Accept values between 910 and 940 thousand).

> **Mark Scheme:** **3 Marks**
> **M1:** Estimates trend value for Q1 2019 (~1335).
> **M1:** Calculates a reasonable negative seasonal effect for Q1 (~ -409).
> **A1:** Arrives at final estimate in the range 910 - 940.

(f) Discuss reliability of estimate in (e)

Worked Solution: The estimate is **unreliable** because Q1 2019 is outside the range of historical data (extrapolation). Trends can change unexpectedly due to economic shocks, airport capacity adjustments, or airline route cancellations.

> **Mark Scheme:** **2 Marks**
> **B1:** Identifies the process as extrapolation.
> **B1:** Concludes it is unreliable with a valid reason.

Question 7 (Total: 2 Marks)

Comment on suitability of the company's simulation plan:

Worked Solution: The plan is **unsuitable**. The digits assigned are 0-8 (9 digits total):

- Adult tickets: 6 numbers (2, 3, 4, 5, 6, 7) ⇒ *6 / 9 pprox 66.7\%* (should be 60%).
- Child tickets: 2 numbers (0, 1) ⇒ *2 / 9 pprox 22.2\%* (should be 30%).
- Senior tickets: 1 number (8) ⇒ *1 / 9 pprox 11.1\%* (should be 10%).

Since the simulated probabilities do not match the real-world proportions (especially for child tickets: 22.2% simulated vs 30% actual), the simulation is invalid. They should assign digits 0-9 (10 digits total).

> **Mark Scheme:** **2 Marks**
> **M1:** Calculates simulated percentages (e.g. Child = 2/9 or 22.2%).
> **A1:** Compares to actual targets (30%) and correctly concludes that the plan is unsuitable.

Question 8 (Total: 9 Marks)

(a) Find the mean and standard deviation of the sample mean

Worked Solution:

- The Target/Mean of the control chart is **10.0 mm**.
- Warning limits are set at *\mu \pm 2 \sigma_{ar{x}}*.

ext{UpperWarningLimit}=10.1ext{mm}\implies 10.0 + 2\sigma_{ar{x}} = 10.1 \implies \sigma_{ar{x}} = 0.05 ext{ mm}

> **Mark Scheme:** **2 Marks**
> **B1:** Mean = 10.0 mm.
> **B1:** Standard deviation of sample mean = 0.05 mm.

(b) Plot sample 6 on the control chart

Worked Solution:

- Diameters: 10.2 mm, 10.15 mm, 10.16 mm.
- Sample mean: *rac{10.2+10.15+10.16}{3}=10.17 ext{ mm}* . Plot on chart at sample number 6 at 10.17 mm.

> **Mark Scheme:** **2 Marks**
> **M1:** Calculates sample mean = 10.17 mm.
> **A1:** Correctly plots point at 10.17 on graph scale.

(c) Draw Upper and Lower Action Lines

Worked Solution: Action limits are placed at *\mu\pm3\sigma_{ar{x}} = 10.0 \pm 3(0.05)*:

- **Upper Action Line:** 10.15 mm (Draw a solid horizontal line).
- **Lower Action Line:** 9.85 mm (Draw a solid horizontal line).

> **Mark Scheme:** **2 Marks**
> **B1:** Calculates action values (10.15 mm and 9.85 mm).
> **B1:** Draws straight horizontal action lines correctly on the grid.

(d) Comment on whether taking Sample 5 immediately was appropriate

Worked Solution: Yes, it was appropriate. Sample 4 plotted at 10.12 mm, which is above the Upper Warning Limit (10.1 mm). The standard quality control protocol dictates that if a sample mean crosses a warning line, production must not be stopped immediately, but another sample must be taken to verify if the machine is drifting out of control.

> **Mark Scheme:** **2 Marks**
> **M1:** Identifies that Sample 4 was above the warning line.
> **A1:** Explains that immediate resampling is required to confirm process stability before shutting down.

(e) Action after sample 6

Worked Solution: Since Sample 6 (10.17 mm) lies above the Upper Action Line (10.15 mm), the machine is officially out of control. **Production must be stopped immediately**, and the machine must be inspected and recalibrated.

> **Mark Scheme:** **1 Mark**
> **B1:** States stop production/recalibrate because it crossed the action line.

Question 9 (Total: 5 Marks)

Petersen Capture-Recapture Method:

1. Execution:

Capture a random sample of red squirrels (n_1), tag/mark them harmlessly, and release them back into Kielder Forest. Allow sufficient time for the marked squirrels to mix randomly with the unmarked population. Capture a second independent sample of squirrels (n_2) and record the count of marked recaptures (m).

2. Calculation:

Estimate the total forest population (N) using:

$$N = \frac{n_1 \times n_2}{m}$$

3. Reliability Assumptions:

- The forest population must be closed (no births, deaths, immigration, or emigration during the study).
- Tags must not fall off or become unreadable.
- Marked squirrels must mix randomly and be equally likely to be recaptured as unmarked ones (no trap-shyness/trap-happiness).

Mark Scheme: **5 Marks**

B2: Explains the capture-tag-release-recapture process clearly.

B1: States the formula $N = \frac{n_1 n_2}{m}$ with defined terms.

B2: Lists at least two valid reliability assumptions in context.

Question 10 (Total: 11 Marks)

(a) Three characteristics of a normal distribution

Worked Solution:

1. It is symmetric and bell-shaped.
2. Unimodal (single central peak where Mean = Median = Mode).
3. The tails are asymptotic to the x-axis (approaching but never touching).

Mark Scheme: **3 Marks**

B3: 1 mark per valid characteristic (symmetry, unimodal, bell shape, asymptotic tails, total area = 1).

(b) Assess Helen's 5% conclusion

Worked Solution: Heights $X \sim N(180, 8^2)$.

$$Z = \frac{196 - 180}{8} = \frac{16}{8} = 2$$

The proportion of males taller than 196 cm corresponds to $P(Z > 2)$. From standard normal tables $P(Z > 2) \approx 2.28\%$ (or 2.3%). Since 2.3% is much less than 5%, Helen's conclusion is incorrect.

Mark Scheme: **4 Marks**

M1: Subtracts mean and divides by standard deviation.

A1: Obtains Z-score = 2.

M1: Identifies tail area $P(Z > 2) \approx 2.3\%$ (or 2.5% using the empirical rule).

A1: Concludes Helen's statement is incorrect.

(c) Assess Helen's two conclusions using the sketch

Worked Solution: From the symmetric female sketch: Mean/Median height of females = 170 cm (the line of symmetry). Range of females is approx. 155 to 185 cm. Since approx. 99.7% of normal data falls within 6σ, we can estimate:

$$6\sigma_{\text{female}} \approx 185 - 155 = 30 \text{ cm} \implies \sigma_{\text{female}} \approx 5 \text{ cm}$$

- **Conclusion 1 (Valid):** Mean male height (180 cm) is greater than mean female height (170 cm). Thus, on average, males are taller.
- **Conclusion 2 (Invalid):** Standard deviation of male heights is 8 cm, whereas standard deviation of female heights is approx. 5 cm. Since $5 < 8$, female heights are actually **more consistent** (less variable) than male heights.

Mark Scheme: **4 Marks**

B1: Identifies female mean = 170 cm and supports Conclusion 1.

M1: Estimates female standard deviation (approx 5 cm) using range limits.

A1: Compares standard deviations ($5 < 8$).

A1: States Conclusion 2 is invalid because females are more consistent.

Question 11 (Total: 4 Marks)

(a) Complete the tree diagram

Worked Solution:

- Tulip branch = 0.6 ⇒ Daffodil branch = 0.4.
- Tulip → White = 0.3 ⇒ Tulip → Yellow = 0.7.
- Given $P(\text{Daffodil} \cap \text{Yellow}) = 0.04$. Since $P(\text{Daffodil}) = 0.4$:

$$P(\text{Yellow} \mid \text{Daffodil}) = \frac{0.04}{0.4} = 0.1$$

$$P(\text{White} \mid \text{Daffodil}) = 1 - 0.1 = 0.9$$

Flower	Color	Probability Path
Tulip (0.6)	White (0.3)	0.18
Tulip (0.6)	Yellow (0.7)	0.42
Daffodil (0.4)	White (0.9)	0.36
Daffodil (0.4)	Yellow (0.1)	0.04

Mark Scheme: **2 Marks**

M1: Calculates $P(\text{Yellow} \mid \text{Daffodil}) = 0.1$.

A1: Completes all branches correctly (0.6, 0.4, 0.3, 0.7, 0.9, 0.1).

(b) Conditional Probability

Worked Solution: Given selecting a white flower, find the probability it is a tulip:

$$P(\text{White}) = P(\text{Tulip} \cap \text{White}) + P(\text{Daffodil} \cap \text{White}) = 0.18 + 0.36 = 0.54$$

$$P(\text{Tulip} \mid \text{White}) = \frac{P(\text{Tulip} \cap \text{White})}{P(\text{White})} = \frac{0.18}{0.54} = \frac{1}{3} \approx 0.333$$

Mark Scheme: **2 Marks**

M1: Calculates total probability of white flowers = 0.54.

A1: Arrives at fraction $\frac{1}{3}$ or decimal 0.333 (3 s.f.).