Pearson Edexcel Level 3 GCE

StatisticsAdvanced—Paper1:DataandProbability (9ST0/01)

Examination Date: Thursday 4 June 2026 **Total Marks:** 80 Marks

Document Type: Official Mark Scheme & Worked Solutions

Question 1 (Total: 15 Marks)

(a) Complete the Venn diagram representing the numbers of cards

Worked Solution:

- Total deck size = 108 cards. Black cards = 8 (all of which are special).
- Remaining cards = $108 - 8 = 100$. Divided equally between 4 colors (Red, Blue, Green, Yellow) = 25 cards each.
- Total Red cards (R) = 25.
- Special cards (S): 8 Black + 6 of each other color ($6 \times 4 = 24$) ⇒ Total $S = 32$.
- Intersection ($R \cap S$): 6 (since 6 of the red cards are special).
- Only Red ($R \cap S'$) = $25 - 6 = 19$.
- Only Special ($S \cap R'$) = $32 - 6 = 26$.
- Outside both ($(R \cup S)'$) = $108 - (19 + 6 + 26) = 57$.

Mark Scheme: 3 Marks

- **B1:** 6 correctly placed in the intersection.
- **B1:** 19 and 26 correctly placed in their respective circles.
- **B1:** 57 correctly placed in the outer box.

(b) Find the probability the card is:

(i) A special card:

P(S) = rac{32}{108} = rac{8}{27} pprox 0.296

Mark Scheme: 1 Mark

B1: *rac{32}{108}* or equivalent fraction/decimal (3 s.f.).

(ii) Neither a red nor a special card:

P(R' \cap S') = rac{57}{108} = rac{19}{36} pprox 0.528

Mark Scheme: 1 Mark

B1: *rac{57}{108}* or equivalent fraction/decimal (3 s.f.).

(iii) A red card given that it is not special:

$$P(R \mid S') = \frac{P(R \cap S')}{P(S')} = \frac{19/108}{76/108} = \frac{19}{76} = 0.25$$

> **Mark Scheme:** **2 Marks**
> **M1:** Correct conditional probability formula with a denominator of 76 (or $108-32$).
> **A1:** $\frac{19}{76}$ or 0.25 exactly.

(c) Show that R and S are not statistically independent

Worked Solution:

If independent, then $P(R \cap S) = P(R) \times P(S)$.

$$P(R) \times P(S) = \frac{25}{108} \times \frac{32}{108} = \frac{800}{11664} \approx 0.0686$$

Since $P(R \cap S) = \frac{6}{108} \approx 0.0556$, we see that $0.0556 \neq 0.0686$. Thus, they are not independent.

> **Mark Scheme:** **2 Marks**
> **M1:** For calculation of product $P(R) \times P(S)$ or conditional probability comparisons.
> **A1:** Correct numerical values shown and clear conclusion drawn.

(d) Probability that exactly 2 of 3 cards are red (without replacement)

Worked Solution:

$$P(\text{Exactly 2 Red}) = 3 \times \frac{25}{108} \times \frac{24}{107} \times \frac{83}{106} = \frac{149400}{1224720} \approx 0.122$$

> **Mark Scheme:** **2 Marks**
> **M1:** Correct setup using hypergeometric method or combinations $\frac{\binom{25}{2}\binom{83}{1}}{\binom{108}{3}}$.
> **A1:** 0.122 (3 s.f.).

(e) Contextual meaning of $(S \cup R)'$

Worked Solution: This represents a card that is **neither a special card nor a red card** (i.e., a normal blue, green, or yellow card).

> **Mark Scheme:** **2 Marks**
> **B1:** Mentions "not red" and "not special".
> **B1:** Fully contextualizes (e.g., normal blue, green, or yellow cards).

(f) Find the probability that Sam is able to play his card on this turn

Worked Solution:

- Remaining cards = $108 - 20 = 88$.

- Sam needs a Blue or a Black card.
- Blue cards remaining: $25 - 6 = 19$. Black cards remaining: $8 - 2 = 6$.
- Winning cards remaining = $19 + 6 = 25$.
- $P(ext{Playable}) = rac{25}{88} pprox 0.284$.

Mark Scheme: **2 Marks**
M1: Correctly calculating remaining valid cards (25) and remaining total (88).
A1: $rac{25}{88}$ or 0.284.

Question 2 (Total: 9 Marks)

(a) Normal Distribution calculations

(i) Systolic blood pressure of more than 140 mmHg:

$$Z = rac{140 - 126.6}{16.76} = 0.80 \implies P(Z > 0.80) = 1 - 0.7881 = 0.2119 pprox 0.212$$

Mark Scheme: **1 Mark**
B1: Correct probability 0.212 (or 0.2119).

(ii) Mean of sample of 100 people is less than 128 mmHg:

$$\sigma_{ar{x}} = rac{16.76}{\sqrt{100}} = 1.676 \implies Z = rac{128 - 126.6}{1.676} = 0.835 \implies P(Z < 0.835) pprox 0.798$$

Mark Scheme: **3 Marks**
M1: Use of standard error formula $rac{\sigma}{\sqrt{n}} = 1.676$.
M1: Correct standardization calculation for Z-score.
A1: 0.798 (accept answers in range 0.797 - 0.799).

(b) Determine whether 179 mmHg should be considered an outlier

Worked Solution: Using standard $\mu \pm 3\sigma$ threshold limit:

$$ext{Upper Limit} = 126.6 + 3(16.76) = 176.88 ext{ mmHg}$$

Since $179 > 176.88$, the value lies beyond 3 standard deviations (Z-score = 3.13) and should be considered an outlier.

Mark Scheme: **2 Marks**
M1: For calculation of a suitable limit (e.g. $\mu + 2\sigma$ or $\mu + 3\sigma$) or Z-score (3.13).
A1: Correct comparison and valid conclusion.

(c) Explain whether the reading of 179 mmHg should be removed

Worked Solution: It should **not** be removed. Although it is an outlier, it represents a genuine medical condition (a real heart condition) and is a valid measurement in the target population. Removing it would introduce selection bias.

> **Mark Scheme:** **2 Marks**
> **B1:** Clear choice (Should not be removed).
> **E1:** Valid reason (e.g. real/natural biological variation, not a recording error).

(d) Criticism of Normal distribution assumption using Figure 1

Worked Solution: The distribution has a very sharp central peak (leptokurtic) and a distinct right-hand skew, which violates the symmetry assumption of the Normal distribution.

> **Mark Scheme:** **1 Mark**
> **B1:** Identifies the skewness or extreme peakedness.

Question 3 (Total: 18 Marks)

(a) Why expected correlation coefficient is positive

Worked Solution: A mock exam measures academic preparation; students who perform well on the mock are logically expected to also perform well on the actual exam.

> **Mark Scheme:** **1 Mark**
> **B1:** Reasonable logical connection between mock and actual performance.

(b) PMCC Calculations

(i) Calculate PMCC value:

$$r pprox 0.885$$

> **Mark Scheme:** **2 Marks**
> **M1:** Formula application or calculation setup.
> **A1:** 0.885 (allow 0.88 - 0.89).

(ii) Interpret this value in context:

There is a strong positive linear correlation between mock exam scores and actual exam scores.

> **Mark Scheme:** **1 Mark**
> **B1:** Mentions "strong positive" and links to both variables.

(c) Usefulness of scatter diagram

Worked Solution: It lets us check if a linear model is appropriate, or if non-linear rank-based correlation (Spearman's) should be used instead.

> **Mark Scheme:** **1 Mark**
> **B1:** Mention of verifying linearity or identifying outlier impact.

(d) Estimate actual exam score for a mock score of 60

$$y = -15.63 + 1.05(60) = 47.37 pprox 47.4$$

> **Mark Scheme:** **2 Marks**
> **M1:** Substitution of $x=60$ into the regression equation.
> **A1:** 47.4.

(e) Comment on reliability of estimate in (d)

Worked Solution: The estimate is reliable because 60 lies within the range of the observed data (interpolation) and the correlation coefficient is high ($r pprox 0.885$).

> **Mark Scheme:** **2 Marks**
> **B1:** Identifies as interpolation.
> **B1:** Concludes it is reliable.

(f) Interpretation of 1.05 in regression equation

Worked Solution: For every 1 mark increase in the mock exam score, the actual exam score is predicted to increase by 1.05 marks on average.

> **Mark Scheme:** **2 Marks**
> **B1:** Mentions increase of 1.05 actual marks per mock mark.
> **B1:** Uses contextual terminology.

(g) Why regression cannot be used for mock scores < 15

Worked Solution: Subbing in $x < 14.89$ results in an impossible negative predicted actual exam score.

> **Mark Scheme:** **1 Mark**
> **B1:** Identifies that it yields a negative score.

(h) Calculate the residual for Student E

Worked Solution: For Student E, $x = 43$, observed $y = 16$.

$$\backslash hat\{y\} = -15.63 + 1.05(43) = 29.52 \ \backslash implies\ ext\{Residual\} = 16 - 29.52 = -13.52$$

Mark Scheme:
M1: Calculates predicted value 29.52.
A1:-13.52.

2 Marks

(i) Probability that total score of 5 students is > 250

Worked Solution: Let $X \sim N(60, 13^2)$. For $n = 5$, Total $T \sim N(5 imes 60, 5 imes 169) \implies T \sim N(300, 845)$.

$$Z = rac{250 - 300}{\sqrt{845}} = rac{-50}{29.07} = -1.72 \implies P(Z > -1.72) = \Phi(1.72) \quad pprox 0.957$$

Mark Scheme:
M1: Mean of total = 300.
M1: Variance of total = 845.
M1: Correct standardization Z-score setup.
A1:0.957 (3 s.f.).

4 Marks

(j) Assumption required for (i)

Worked Solution: The scores of the 5 students must be independent of each other.

Mark Scheme:
B1: Explicitly states independence.

1 Mark

Question 4 (Total: 14 Marks)

(a) Standard deviation of earthquakes

$$\sigma = \sqrt{\lambda} = \sqrt{148} \ pprox 12.2$$

Mark Scheme:
B1:12.2.

1 Mark

(b) Find the probability that:

(i) At least 150 earthquakes in 1 year:

Using Normal approximation: $Y\sim N(148, 148)$. Applying continuity correction:

$$P(X \ge 150) \quad pprox P(Y > 149.5) \implies Z = rac{149.5 - 148}{\sqrt{148}} = 0.123 \implies P(Z > 0.123) \quad pprox 0.451$$

Mark Scheme:
B1:0.451 (allow 0.44 - 0.46).

1 Mark

(ii) No more than 260 earthquakes in 2 years:

For 2 years, \lambda = 296. Using Y_2 \sim N(296, 296) with continuity correction:

P(X_2 \le 260)   pprox P(Y_2 < 260.5) \implies Z = rac{260.5 - 296}{\sqrt{296}} = -2.063 \implies P(Z < -2.063)   pprox 0.020

> **Mark Scheme:** **2 Marks**
> **M1:** Correctly setting parameter to 296 and standardizing with continuity correction.
> **A1:** 0.020 (or 0.0196).

(iii) Gap of more than 6 days between consecutive earthquakes:

Rate per day \lambda_d = rac{148}{365} pprox 0.4055.

P(T > 6) = e^{-6 \lambda_d} = e^{-6(0.4055)} = e^{-2.433}   pprox 0.0878

> **Mark Scheme:** **3 Marks**
> **M1:** Finding daily rate 0.4055.
> **M1:** Use of exponential tail formula e^{-\lambda t}.
> **A1:** 0.0878.

(c) Does this gap of 6 days support the use of this model?

Worked Solution: Yes, because the probability of having a gap of more than 6 days is approximately 8.78\%. Since this is not a highly unlikely event (i.e. > 5\%), it is perfectly consistent with standard Poisson model fluctuations.

> **Mark Scheme:** **3 Marks**
> **M1:** Mentions calculated probability from (b)(iii).
> **M1:** Logical comparison with standard significance criteria.
> **A1:** Concludes that it supports the model.

(d) Comment on the seismologist's claim using Figure 4

Worked Solution: The monthly frequencies vary significantly (ranging from 8 in July to 17 in October). A true Poisson process requires a constant average rate. However, seasonal variation or clustering of earthquakes (aftershocks triggering others) suggests the independence/constant rate assumptions might be violated, supporting the claim.

> **Mark Scheme:** **4 Marks**
> **B1:** Identifies variance across months from the chart.
> **B1:** Links to "constant average rate" assumption.
> **B1:** Discusses real-world factors like aftershocks (independence).
> **B1:** Evaluates claim validity based on evidence.

Question 5 (Total: 13 Marks)

(a) Relative advantages, disadvantages, and recommendation

Method	Advantages	Disadvantages
Method A (Immediate Survey)	• Extremely cheap and easy. • High response rate.	• High pressure to complete, leading to biased answers. • Non-representative of departments.
Method B (Random Email)	• Unbiased random selection. • Comfortable environment for response.	• Typically very low response rates. • Risk of non-response bias.
Method C (Stratified Inter view)	• Proportional/representative of company. • Rich, detailed qualitative feedback.	• Very expensive and time-consuming. • Risk of interviewer bias.

Recommendation: Use **Method C** because the sample is small enough (50 people out of 390 total) for interviews to be viable, ensuring all departments are fairly represented.

Mark Scheme: **8 Marks**
B1-B6: Up to 6 marks for detailed discussion of advantages/disadvantages.
M1: Recommends a specific method.
A1: Provides solid reasoning in context.

(b) Calculate how many employees will be selected from IT

ext{IT Sample} = rac{42}{390} imes 50 pprox 5.38 \implies 5 ext{ employees}

Mark Scheme: **2 Marks**
M1:*rac{42}{390} imes 50.*
A1: 5 (must be an integer).

(c) State whether you agree with the company's conclusion

Worked Solution: Disagree. A mean of 3.8 does not prove that the *majority* agreed. It is possible that extreme ratings (many 5s and some 1s/2s) skewed the mean, or that the majority selected 3 (neutral) and 4 (agree), rather than specifically "agreeing or strongly agreeing". The median or mode is needed for ordinal data.

Mark Scheme: **3 Marks**
B1: Explicitly states disagreement.
E1: Explains that mean is heavily affected by outliers.
E1: Explains that mean does not equal majority on Likert scales.

Question 6 (Total: 11 Marks)

(a) Show that a suitable estimation for cancellation is 0.0156

P(ext{Cancelled}) = rac{6400}{409000} pprox 0.015648 pprox 0.0156

Mark Scheme: 1 Mark

B1: Show calculation setup leading to 0.0156.

(b) Use binomial distributions to find the probability that:

(i) None of the flights will be cancelled:

P(C = 0) = (1 - 0.0156)^5 = 0.9844^5 pprox 0.924

Mark Scheme: 1 Mark

B1: 0.924.

(ii) At least one flight will be cancelled:

P(C \ge 1) = 1 - P(C = 0) = 1 - 0.9242 pprox 0.0758

Mark Scheme: 1 Mark

B1: 0.0758.

(iii) Exactly two flights do not operate on time:

Probability not on time $q = 1 - 0.71 = 0.29$. Let W \sim ext{B}(5, 0.29).

P(W = 2) = inom{5}{2} (0.29)^2 (0.71)^3 = 10 imes 0.0841 imes 0.3579 pprox 0.301

Mark Scheme: 2 Marks

M1: Setting up correct binomial parameters and formula.

A1: 0.301.

(c) Assumptions for binomial validity

Worked Solution:

1. The outcome of each flight booking is independent of the others.
2. The probability of flight delays/cancellation remains constant over time.

Mark Scheme: 1 Mark

B1: Lists any two valid conditions (Independence, Constant Probability, Fixed Trials).

(d) Why Normal approximation is unsuitable for (b)

Worked Solution: For a Normal approximation, we require $np > 5$ and $n(1-p) > 5$. Here, $np = 5 imes 0.0156 = 0.078$, which is far below 5, making the distribution highly skewed.

> **Mark Scheme:** **2 Marks**
> **M1:** Calculates np or mentions small trial size.
> **A1:** Concludes that criteria are not met, rendering approximation invalid.

(e) Probability that it is Kenya Airways given it was not on time

Worked Solution: Let K = Kenya Airways, T' = Not on time.

$$P(K) = rac{180}{409000}, \quad P(T' \mid K) = 0.55, \quad P(T') = 0.29$$

$$P(K \mid T') = rac{P(T' \mid K) imes P(K)}{P(T')} = rac{0.55 imes rac{180}{409000}}{0.29} \qquad pprox 0.000835$$

> **Mark Scheme:** **3 Marks**
> **M1:** Setting up Bayes' Theorem ratio.
> **M1:** Correct calculation of numerator (99 flights) and denominator ($118,610$ flights).
> **A1:** 0.000835 (or $8.35 imes 10^{-4}$).