Data Detective — Worked Solutions

MYP Year 10 Mathematics

Side 1: Anscombe's Quartet

Q1. Summary statistics (to 2 d.p.)

	Set I		Set II		Set III		Set IV	
Statistic	x	y	x	y	x	y	x	y
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std. dev.	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
Correlation r	0.82		0.82		0.82		0.82	

Observation: All four columns are identical to 2 d.p. From the summary statistics alone, the four data sets appear to be the same.

Q2. Predict the scatter plots

Students' predictions vary. Most expect four roughly similar linear-with-scatter clouds (the "honest" Set I shape), because the statistics suggest the same underlying relationship in every case. The point of the prediction is to expose the gap between this expectation and what the plots actually show.

Q3. Plot them for real

The four scatter plots reveal completely different structures:

- **Set I:** a roughly linear point cloud with natural scatter — the only set where a linear model is genuinely appropriate.
- **Set II:** a clear **curve** (parabolic shape). The relationship is not linear, so the regression line fits poorly even though r looks high.
- **Set III:** ten points lie on a **perfect straight line**, and a single outlier pulls the line of best fit away. Without that one point the relationship is exact.
- **Set IV:** ten points sit in a **vertical column** at $x = 8$, plus one outlier at $(19, 12.5)$. The correlation is created entirely by that single outlier — with that point removed, r would be undefined.

Key idea: Summary statistics are blind to the *shape* of the data. Curves, outliers, and unusual clusters all collapse into the same mean, standard deviation, and correlation. The lesson is to **always plot the data** before reaching for a numerical model.

Side 2: Simpson's Paradox

Q4. Calculate the pass rates

Method	Multiple choice	Long answer	Overall passed / total	Overall rate
A	$\frac{81}{87} = 93.1\%$	$\frac{192}{263} = 73.0\%$	273 / 350	78.0%
B	$\frac{234}{270} = 86.7\%$	$\frac{55}{80} = 68.8\%$	289 / 350	82.6%

Q5. Which method is better?

(a) Multiple choice: $93.1\% > 86.7\%$, so **Method A is better**.

(b) Long answer: $73.0\% > 68.8\%$, so **Method A is better**.

(c) Overall: $82.6\% > 78.0\%$, so **Method B is better**.

Q6. What is going on?

Look at how the students were distributed across the two exams:

- **Method A:** 87 took MC and 263 took LA — about **75% of A's students sat the harder long-answer exam**.
- **Method B:** 270 took MC and 80 took LA — about **77% of B's students sat the easier multiple-choice exam**.

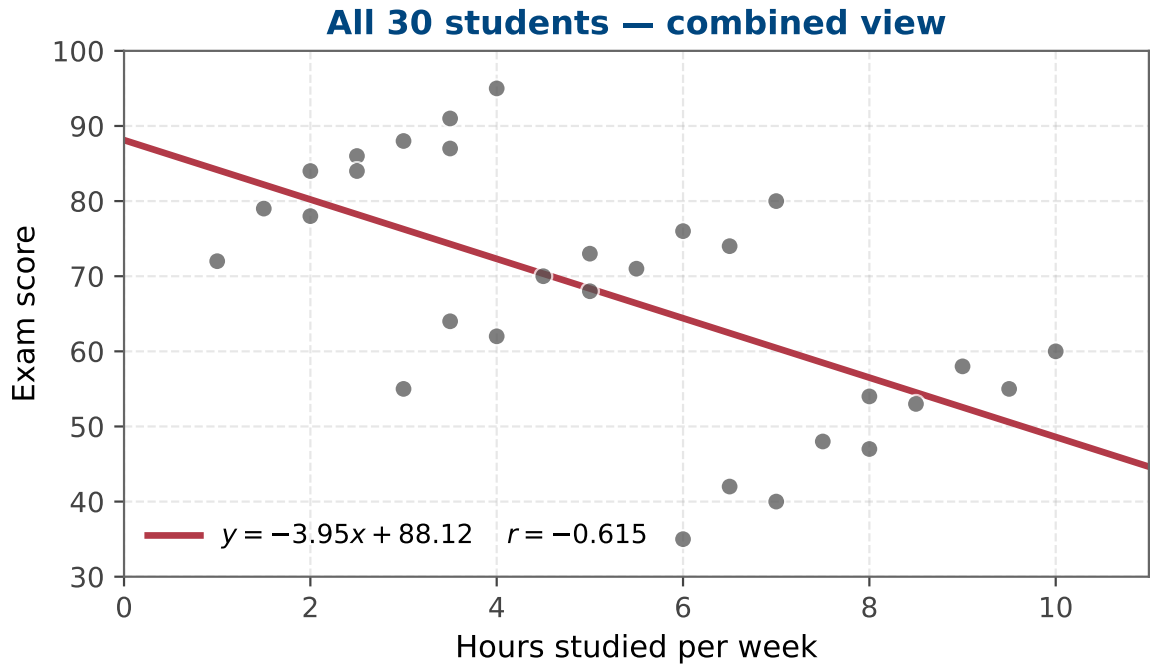
The MC exam has much higher pass rates than the LA exam for both methods. Since Method B's students were mostly given the easier paper, the overall pass rate for B is pulled *up* by that lopsided weighting, while Method A's overall rate is dragged *down* by having more LA candidates. The exam type is a **lurking variable** hiding behind the headline "overall" figures.

Recommendation: Method A is genuinely better. It wins both like-for-like comparisons, where students sat the same exam. Method B's apparent advantage in the overall figure is an artefact of how students were allocated to the two methods, not of teaching quality.

Side 3: Hidden Groups — plot all the data together

Q7. The combined scatter

Using all 30 students, the scatter shows a clear **negative** pattern: points run from top-left (low hours, high scores) to bottom-right (high hours, low scores).



The linear trendline (computed from `=SLOPE` and `=INTERCEPT` on `J6:K35`):

$$\boxed{y = -3.95\,x + 88.12}$$

Q8. The combined correlation

`=CORREL(J6:J35, K6:K35)` = **r ≈ −0.615**

This is a **moderately strong negative** correlation. Reading the plot alone, we would conclude that *more* study hours seem to lead to *lower* exam scores. **That conclusion should feel wrong** — common sense says studying helps, not hurts. That uneasy feeling is the signal that something is hidden in the data.

Side 4: Hidden Groups — group the same data

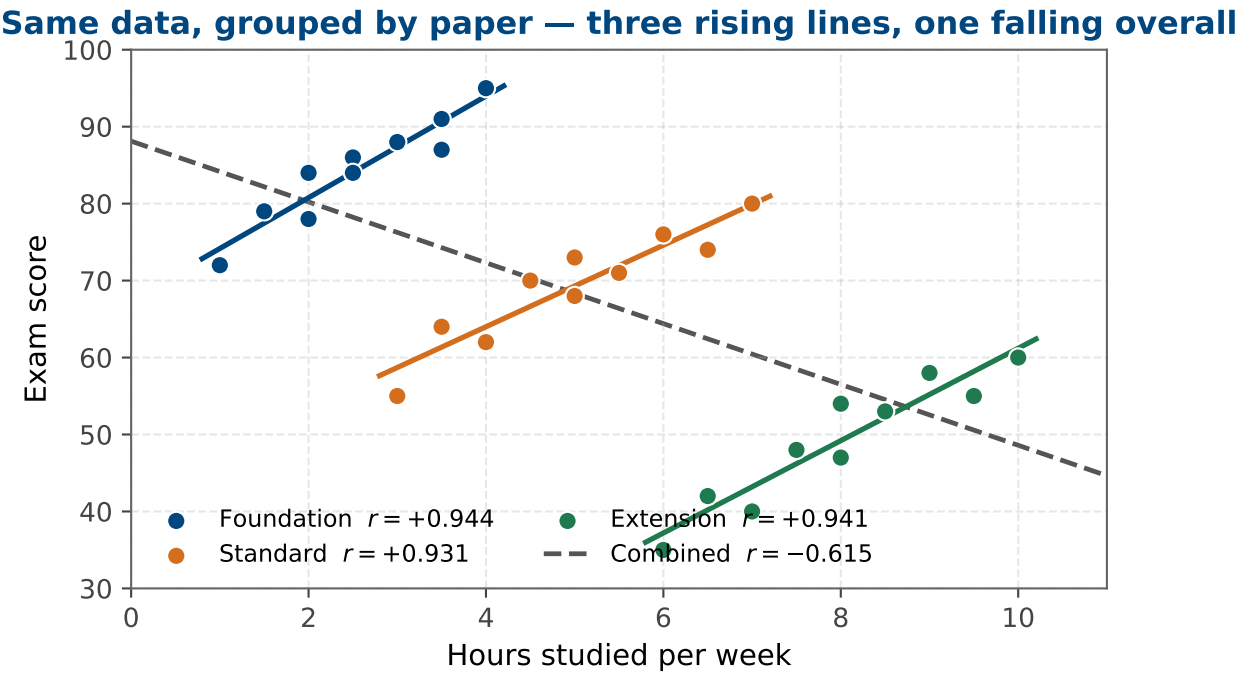
Q9. Per-group correlations

	Foundation	Standard	Extension
r	+0.944	+0.931	+0.941

All three group correlations are **strong and positive** — the *opposite sign* to the combined $r \approx -0.615$. Within each paper, more study time *is* associated with higher scores, as expected.

Q10. The grouped scatter

The three group trendlines all slope **upwards** (positive). The clusters sit at very different heights: Foundation top-left (low hours, high scores), Standard middle, Extension bottom-right.



The group means are:

- Foundation: mean hours ≈ 2.55, mean score ≈ 84.4
- Standard: mean hours ≈ 5.00, mean score ≈ 69.3
- Extension: mean hours ≈ 8.00, mean score ≈ 49.2

The line connecting those three group means slopes *downwards* sharply — that is what drives the misleading combined trendline.

Q11. Explain the paradox

The hidden variable is the **paper the student sat**. The Foundation paper is easier so students study less but score highly. The Extension paper is harder so students study more but score lower in absolute terms. When all 30 students are pooled, the *difference in difficulty between papers* dominates the picture and the within-group "more study → higher score" pattern is hidden.

The numbers tell the story:

- Within each paper, r is strongly positive (+0.944, +0.931, +0.941).
- Across all 30 students, $r = -0.615$: the *group means* sit on a downward line, swamping the within-group trends.

Connect: a general lesson for the newspaper reader.
Across the three tasks the same principle keeps appearing: a headline statistic can hide structure. Anscombe shows that identical summary numbers can come from completely different shapes — always ask to see the plot. Simpson (proportions and regression) shows that pooling data across genuinely different groups can reverse the apparent conclusion — always ask how the groups were formed, and whether subgroups would tell a different story.

A good response should pick up on at least the following ideas:

- Ask **what the underlying data looks like** before trusting a single summary number.
- Ask **how the groups were formed**, and whether subgroups would tell a different story.
- Where possible, ask to **see the plot**, not just the headline figure.