

Unit 1A Review — Solutions

AP Statistics · Mr. Levy

Check your reasoning against this, not just your answer. For several of these the number is the easy part and the justification is the whole question — questions **7c**, **11c** and **12b** have more than one correct response, and what makes an answer right there is the argument you give for it.

1. **Mean 50 and range 98 for all three classes. Standard deviation, largest first: B, then C, then A.**

All three classes have the same mean and the same range, so neither of those numbers tells them apart. Standard deviation asks a different question: how far is a typical score from 50?

In **B** almost everybody sits 49 points away from the mean, so the standard deviation is enormous — about 49. In **C** the scores are spread evenly from 1 to 99, so a typical distance is about 29. In **A**, 97 of the 99 students scored exactly 50, so almost every deviation is 0 and the standard deviation is only about 7.

The table is the argument: identical mean, identical range, three completely different spreads. Range looks only at the two most extreme students and is blind to the other 97.

2. **Yes, it can be true.**

It takes a right-skewed distribution. A long tail of high scores drags the mean up past the middle of the class, so most students end up sitting below it. Any description with the bulk of the scores clustered low and a few very high ones will do it.

The trap is reading the mean as “the middle.” The middle is the median. The mean is the balance point, and a few extreme values can pull it a long way off centre.

3. **I = T · II = F · III = F**

I. The median is the 50th percentile. That is what median means.

II. An IQR of 20 tells you only that $Q3 - Q1 = 20$. It does not tell you the quartiles are 260 and 280 — they could just as easily be 265 and 285.

III. Seventy-five percent of the data sit at or below $Q3$, and you were never told that $Q3$ is 280.

Both wrong answers come from assuming the quartiles sit symmetrically around the median. They do that only when the distribution is symmetric, and nothing here says it is.

4. **I = T · II = F · III = T**

Reading the plot: minimum 245, $Q1$ 260, median 270, $Q3$ 273, maximum 281.

I. 273 is $Q3$, and $Q3$ is the 75th percentile.

II. A boxplot cannot show a mean. The line inside the box is the median, 270. This distribution is skewed left, so the mean is probably a little below 270 — but nothing on this graph tells you what it actually is.

III. The left whisker ends at 245.

II is the one worth sitting with. Every number a boxplot shows is a position in the ordered list. The mean is not one of them.

5. **Only the third box: the mean and the median can both be unchanged.**

Take two outliers placed symmetrically about the centre — say 10 and 90 in a set centred at 50. Removing both leaves the balance point exactly where it was, and taking one value off each end leaves the middle where it was too.

The other three all need the standard deviation to hold still or grow. It cannot: you just removed the two values furthest from the mean, so the typical distance from the mean has to come down.

“Resistant” means a statistic **can** survive an outlier — not that it must be unaffected by removing one.

6. Ten 2s.

A standard deviation of 0 means every deviation from the mean is 0. So every value equals the mean, and the mean is 2.

The tempting wrong answer is five 1s and five 3s. That does have a mean of 2, which is why it looks right — but its standard deviation is 1, not 0.

7. a) Age, height and sleep are quantitative; travel method is categorical. student_id is not a variable at all.

b) The height of 17 cm, for student 40670.

a) An id is unique to each row and measures nothing about the student. It is a label that happens to be written with digits. Ask yourself what the mean student id would tell you — nothing, and that is the test.

b) 17 cm is about 7 inches, which is impossible for a person. Every other height sits between 162 and 183 cm, so this is almost certainly 170 cm with a digit dropped.

c) There is no single right answer here, and that is the point. The strongest responses say investigate first: go back to the source rather than silently deleting a student or inventing a replacement value. Deleting data changes your sample.

8. a) Mean 40 minutes, median 25 minutes. b) Mean \$263,000, median \$224,000.

Both distributions are skewed right, so in both the mean is the larger number. A few very long commutes — and a few very expensive houses — pull the mean up while leaving the median down where the bulk of the data actually is.

There is no calculation in this question. The whole answer is deciding which direction the tail goes, and in both cases the tail runs toward the expensive, slow end.

9. a) Median $\approx$ 6 letters; mean $\approx$ 6.3 letters. b) Quote the median.

a) There are 500 sites, so the median sits between the 250th and 251st value. Counting bars from the left, you reach 250 inside the 6-letter bar. The mean lands a little above it, because the tail runs right to 15 letters with nothing balancing it on the left.

b) The median is the smaller number here, and the honest one: the long right tail inflates the mean, so the median better describes a typical popular site.

Any median of 6, and any mean between 6 and 7 with sound reasoning, is a correct answer. The method matters more than the number.

10. 69.86 inches.

The 500 students total $500 \times 67.8 = 33,900$ inches. The 150 girls account for $150 \times 63 = 9,450$ of those inches, which leaves 24,450 inches spread across 350 boys. So the boys average $24,450 \div 350 = 69.86$ inches.

Two shortcuts that fail: averaging the two means, and subtracting 63 from 67.8 and doubling it. The group sizes have to carry their weight — the boys outnumber the girls more than two to one.

11. a) No outliers. b) 80 is an outlier. c) The two rules disagree here.

a) $Q1 = 20$ and $Q3 = 50$, so $IQR = 30$. The fences are $20 - 45 = -25$ and $50 + 45 = 95$. Nothing is beyond 95, so this rule flags nothing.

b) The mean is 37 and $s \approx 16.45$, so the bounds are 37 ± 32.9 , that is 4.1 and 69.9. 80 is past the upper bound, so this rule does flag it.

c) The middle half of this data is wide (IQR 30), which pushes the $1.5 \times IQR$ fence all the way out to 95, so that rule sees nothing unusual. The 2-standard-deviation rule measures against a much smaller yardstick and catches the 80.

Neither rule is “right.” They are two conventions, and calling a value an outlier is a judgement you have to defend — not a fact the data hands you. Either report is correct as long as you name the rule you used.

12. Route B is typically faster; Route A is far more reliable.

Centre. B’s median is 14 minutes against A’s 20, so a typical B trip is about 6 minutes shorter.

Spread. B’s IQR is 17 minutes against A’s 6, and B’s range is 39 against A’s 14. B is hugely more variable.

Shape. A is roughly symmetric. B is clearly skewed right — its box runs 4 minutes below the median and 13 above, and its upper whisker is 17 minutes long.

b) Route A wins whenever being late is expensive: a test first period, an exam, a shift you cannot miss. Three quarters of A’s trips are done by 23 minutes and its worst is 27, so you can plan around it. Route B wins when only the typical day matters and arriving late now and then costs nothing — it is about 6 minutes quicker at the median, so across a term of ordinary mornings it gives back real time.

The distinction to draw: **A is the better route on its worst day, B on its typical day.** Any pair of scenarios that turns on that difference is correct. A pair that turns only on which median is smaller is not.

13. a) Maya $z = -1.50$, Jordan $z = -1.75$. Jordan had the better run. b) 64.0 seconds, and it was a poor run.

a) Maya: $(53.5 - 58.0) \div 3.0 = -1.50$. Jordan: $(128.0 - 142.0) \div 8.0 = -1.75$. Because a lower time is a better run, the more negative z-score is the better performance — Jordan is further ahead of his field than Maya is of hers.

b) $58.0 + 2.00(3.0) = 64.0$ seconds. That is two standard deviations *slower* than the mean.

Nothing on the sheet tells you that a lower time is better — noticing that is the question. A z-score has a sign, and what the sign *means* depends on the variable. Ask yourself what a z-score of +3 would mean for a marathon time.

14. a) Mean 14.3 lb, standard deviation 3.96 lb, median 13.64 lb, IQR 5.5 lb. b) Still the 80th percentile.

a) Every one of them gets multiplied by 2.2, spread included: $6.5 \times 2.2 = 14.3$, $1.8 \times 2.2 = 3.96$, $6.2 \times 2.2 = 13.64$, $2.5 \times 2.2 = 5.5$.

b) Multiplying every weight by the same positive number does not reorder anybody. The heaviest backpack is still the heaviest, and 80% of them are still lighter than this one. Only the numbers on the axis changed — not the picture.

Two errors to avoid. Leaving the standard deviation alone confuses this with *adding* a constant. Multiplying the percentile by 2.2 to get 176 confuses a position with a measurement — a percentile has no units, so there is nothing to convert.