Standard deviation of the sum of uncorrelated data

Let's say five students took English test and math test. You can see their scores, the sum of the scores and their average scores in Table 1. You see that there is a clear tendency in the scores of students. Usually, the higher someone's English score, the higher his/her math score. The only exception is student 2 and 3. Student 3's English score is slightly higher than student 2's, but student 3's math score is slightly lower than student 2's. The English + math score in the last column is obtained by adding the English score and the math score. In the bottom row, you see that the average of English score is 78, the average of Math score is 69, and the average of English + math score is 147.

Notice that the average of English + math score is the sum of the average of English score and the average of math score. This is expected as it doesn't matter in which order we add numbers when adding a group of numbers. In other words, if we denote the English score by E and the math score by M, we have

$$\begin{aligned}&(E_1 + E_2 + E_3 + E_4 + E_5) + (M_1 + M_2 + M_3 + M_4 + M_5)\\ &\quad = (E_1 + M_1) + (E_2 + M_2) + (E_3 + M_3) + (E_4 + M_4) + (E_5 + M_5)\end{aligned} \tag{1}$$

From Table 1, you will see that the above formula corresponds to

$$390 + 345 = 110 + 135 + 135 + 165 + 190 = 735 \tag{2}$$

If you divide the left-hand side of the above formula by 5, you get 78+69=147.

So far, I said something that was very obvious. The average of the sum is the sum of average. Then, what about the variance? Can we say something similar to the case of

Student	English	math	English + math
1	60	50	110
2	70	65	135
3	75	60	135
4	85	80	165
5	100	90	190
Sum	390	345	735
Average	78	69	147

Table 1: English and math scores of five students

	E	M	ΔE	ΔM	$\Delta E + \Delta M$	$(\Delta E)^2$	$(\Delta M)^2$	$(\Delta E + \Delta M)^2$
1	60	50	-18	-19	-37	324	361	1369
2	70	65	-8	-4	-12	64	16	144
3	75	60	-3	-9	-12	9	81	144
4	85	80	7	11	18	49	121	324
5	100	90	22	21	43	400	400	1600
Sum	390	345	0	0	0	930	1020	3830
Sum/5	78	69	0	0	0	186	204	766

Table 2: English and math scores of five students with the deviations from the average

	E	M	ΔE	ΔM	$2\Delta E\Delta M$	$(\Delta E)^2$	$(\Delta M)^2$	$(\Delta E + \Delta M)^2$
1	60	50	-18	-19	684	324	361	1369
2	70	65	-8	-4	64	64	16	144
3	75	60	-3	-9	54	9	81	144
4	85	80	7	11	154	49	121	324
5	100	90	22	21	924	400	400	1600
Sum	390	345	0	0	1880	930	1020	3830
Sum/5	78	69	0	0	376	186	204	766

Table 3: Analysis of the variance of English score, math score and "English + math" score

average? Let's find out. See Table 2. We denoted the deviation of E from the average, i.e., $E - \langle E \rangle$ by ΔE, and similarly for M.

From the last row, you see that the variance of English score is 186, and the variance of math score is 204, and the variance of "English + Math" score is 766. Thus, we see $186 + 204 \neq 766$. Then, maybe, does the standard deviation add up? By using a calculator, you can easily check $\sqrt{186} + \sqrt{204} \neq \sqrt{777}$. So, the standard deviation doesn't add up either.

What is happening here? The column $\Delta E + \Delta M$ is the sum of the column ΔE and the column ΔM. However, the column $(\Delta E + \Delta M)^2$ is *not* the sum of the column $(\Delta E)^2$ and the column $(\Delta M)^2$, because

$$(\Delta E + \Delta M)^2 = \Delta E^2 + \Delta M^2 + 2\Delta E \Delta M \neq (\Delta E)^2 + (\Delta M)^2 \tag{3}$$

In other words, there is the difference of $2\Delta E\Delta M$.

To see this more closely, let's make a column "$2\Delta E\Delta M$." See Table 3. Now, you see that the columns $2\Delta E\Delta M$, $(\Delta E)^2$, $(\Delta M)^2$ add up to $(\Delta E + \Delta M)^2$. Now, notice that, in our case of Table 3, $(\Delta E + \Delta M)^2$ is bigger than the sum of $(\Delta E)^2$ and $(\Delta M)^2$. Why is this so? It's because each entry for $2\Delta E\Delta M (= (\Delta E + \Delta M)^2 - (\Delta E)^2 - (\Delta M)^2)$ is bigger than 0. So why is each entry for $2\Delta E\Delta M$ positive?

When is $2\Delta E\Delta M$ be positive? It's when both ΔE and ΔM are positive, or when both ΔE and ΔM are negative. When are both ΔE and ΔM positive? When the English score

and the math score of a student are both above average. When are ΔE and ΔM? When the English score and the math score of a student are both below average. So, we see that, if students with higher scores in English tend to get higher scores in math, while students with lower scores in English tend to get lower scores in math, most or all of the entries for $2\Delta E \Delta M$ will be positive.

This is our case. That is why $2\Delta E\Delta M = (\Delta E + \Delta M)^2 - (\Delta E)^2 - (\Delta M)^2$ is bigger than zero. In other words, this is the reason why

$$\mathrm{Var}(E+M) > \mathrm{Var}(E) + \mathrm{Var}(M) \tag{4}$$

Problem 1. Explain similarly, why $2\Delta E\Delta M$ will be negative if students with higher score in English tend to get lower scores in math. In other words, show why the following must be satisfied in such a case:

$$\mathrm{Var}(E+M) < \mathrm{Var}(E) + \mathrm{Var}(M) \tag{5}$$

If neither is the case, i.e., if there is no strong correlation between English scores and math scores, we will get

$$\mathrm{Var}(E+M) \approx \mathrm{Var}(E) + \mathrm{Var}(M) \tag{6}$$

Let's derive this more rigorously. This time, E will be the SAT English score and M the SAT math score. They both can be from 200 to 800, and their average is both 500. However, I will assume that there is no correlation between the SAT English score and the SAT math score, which is *not* the actual case. In other words, I will assume, both the average SAT math score of students who got 800 in the SAT English test and the average SAT math score of students who got 200 in the SAT English test are 500. It's just an assumption for the sake of explanation. Now, let's calcualte the variance of $E + M$.

$$\begin{aligned}
\langle (E+M)^2 \rangle - \langle E+M \rangle^2 &= \langle E^2 + M^2 + 2EM \rangle - (\langle E \rangle + \langle M \rangle)^2 && (7)\\
&= \langle E^2 \rangle + \langle M^2 \rangle + \langle 2EM \rangle - (\langle E \rangle^2 + \langle M \rangle^2 + 2\langle E \rangle \langle M \rangle) && (8)\\
&= (\langle E^2 \rangle - \langle E \rangle^2) + (\langle M^2 \rangle - \langle M \rangle^2) + 2\langle EM \rangle - 2\langle E \rangle \langle M \rangle && (9)\\
&= \mathrm{Var}(E) + \mathrm{Var}(M) + 2(\langle EM \rangle - \langle E \rangle \langle M \rangle) && (10)
\end{aligned}$$

So, let's calculate $\langle EM \rangle$. Let's say the total number of students is N. If $N(E)$ students got a score of E in English test, $N(E, M)$ students got a score of E in English test and M in math test, we certainly have

$$N = \sum_{E=200}^{800} N(E), \qquad N(E) = \sum_{M=200}^{800} N(E, M) \tag{11}$$

and if we denote the average math score of students who got E for English score by $\langle M \rangle_E$,

$$\langle E \rangle = \frac{1}{N} \sum_{E=200}^{800} EN(E), \qquad \langle M \rangle_E = \frac{1}{N(E)} \sum_{M=200}^{800} MN(E, M) \tag{12}$$

Then,

$$\langle EM\rangle = \frac{1}{N}\sum_{E=200}^{800}\sum_{M=200}^{800} N(E,M)EM \tag{13}$$

$$= \frac{1}{N}\sum_{E=200}^{800} EN(E)\frac{\sum_{M=200}^{800} N(E,M)M}{N(E)} \tag{14}$$

$$= \frac{1}{N}\sum_{E=200}^{800} EN(E)\langle M\rangle_E \tag{15}$$

Now, recall our assumption. We assumed that the average math score doesn't depend on what you got in your English test. Thus, $\langle M\rangle_E = \langle M\rangle$. Then, (15) becomes

$$\langle EM\rangle = \frac{1}{N}\sum_{E=200}^{800} EN(E)\langle M\rangle \tag{16}$$

$$= \left(\frac{1}{N}\sum_{E=200}^{800} EN(E)\right)\langle M\rangle \tag{17}$$

$$= \langle E\rangle\langle M\rangle \tag{18}$$

Therefore, (10) becomes

$$\mathrm{Var}(E+M) = \mathrm{Var}(E)+\mathrm{Var}(M)+2(\langle E\rangle\langle M\rangle - \langle E\rangle\langle M\rangle) \tag{19}$$

$$= \mathrm{Var}(E)+\mathrm{Var}(M) \tag{20}$$

This is when there is absolutely no correlation between E and M. If we just say there is no *strong* correlation between E and M, this relation essentially becomes (6).

Summarizing, if there is no correlation between data A and data B, we will have

$$\mathrm{Var}(A+B) = \mathrm{Var}(A)+\mathrm{Var}(B) \tag{21}$$

Similarly, if there is no correlation among data A, B, and C, there will be no correlation between the data $A+B$ and C. Thus,

$$\mathrm{Var}(A+B+C) = \mathrm{Var}(A+B)+\mathrm{Var}(C) = \mathrm{Var}(A)+\mathrm{Var}(B)+\mathrm{Var}(C) \tag{22}$$

where we used (21) in the last step.

A similar relation can be written for arbitrary number of independent data.

Summary

- If there is no correlation between the data A, B. We have

  $$\mathrm{Var}(A+B) = \mathrm{Var}(A)+\mathrm{Var}(B)$$

  A similar relation is true for arbitrary number of independent data; their variances add up.