How does the change of a single data point affect the variance, and why?

Hans Humenberger

University of Vienna, Vienna, Austria

Correspondence
Hans Humenberger, University of Vienna, Vienna, Austria.
Email: hans.humenberger@univie.ac.at

Abstract
In this paper, we investigate an interesting question that came up when reading a problem in a school textbook: What happens to the variance of a dataset in the case of changing one single data point, and why? Some of the answers are not surprising but here we find the full answer and demonstrate the understanding of it suitable for school students.

KEYWORDS
teaching, statistics, variance

1 INTRODUCTION

Let us start with the analogous question concerning the mean $\bar{x}$ of a dataset, where everything is straightforward. It is immediately clear in which direction $\bar{x}$ changes, if exactly one value of $x_i$ is increased and the others are regarded as constant, it will increase, too. A common analogy for this is, if one has the concept of $\bar{x}$ as the center of gravity of a bar with weights. The positions, $x_i$, are interpreted as points of equal masses, and $\bar{x}$ corresponds to that point on the number line in which this line must be supported in order that it—thought as a bar with weights—is in balance. If the bar is in perfect balance and one data point (weight, no matter whether it is on the right hand or left hand side of $\bar{x}$) is moved to the right (left), then the whole bar tips to the right (left). Moreover, it can be easily determined how much the mean changes in case of $x_n \rightarrow x_n + h$, this is $\frac{h}{n}$. It is arbitrary which point to move. For convenience, in this article, the last observed data point, $x_n$, will be moved.

In the following paragraphs, we deal with the analogous situation with the variance instead of the mean. We need to consider only the sum of the $n$ squared deviations to the mean value: $v = \sum_{i=1}^{n} (x_i - \bar{x})^2$.

The actual trigger for all these considerations was a problem in a school textbook (grade 8), in the section Descriptive Statistics (here presented only with regard to contents, not literally cited): Thomas and Carina played the same computer game 20 times each and recorded their results in a table. This table says of both how many times they got the possible numbers of points (100, 200, 300, 400, 500) in the 20 trials.

1. Determine for Thomas and Carina the arithmetic mean, $\bar{x}$, and the variance of their point numbers!
2. Carina has made a mistake, in one game she noted 200 instead of 300 points. What will be the effect of repairing this mistake: will the new (correct) value of $\bar{x}$ be smaller or bigger than the old (false) one? Will the new (correct) value of the variance be smaller or bigger than the old (false) one? Give a conjecture before calculating!
3. Give arguments for your conjecture!

It is not difficult to give reasons for the mean, but in the case of the variance, the argument seems to be less simple. Let us assume that Carina’s mean—with the wrong data point (200 points)—was 320 points. Then, it
is obvious that the new (correct) data point, 300, is nearer to her old mean; therefore, the conjecture that the variance gets smaller due to the change 200 \rightarrow 300 will come up quickly, because one crucial squared distance gets smaller. Perhaps, this was meant by the authors of the school textbook. But is it really correct and that easy? Is the following always the case (independent of the position of the other data points): whenever one data point comes closer to the mean, the variance always decreases? Or formulated the other way round: whenever one data point is moved away from the mean, the variance always increases? After all, the mean itself also changes if one data point moves, and therefore all (squared) distances change. Considering all that it is not so easy at all to reasonably estimate the effects to the sum of all the squared deviations, and hence the new variance.

This is a classical example that the extent of doubt can increase by thinking more deeply. Sometimes such doubts are highly appropriate and help to find fallacies and misconceptions in arguing intuitively. In this case, we will see that the above intuition is usually correct, but not always. In the above example, with the original mean at 320 and the data point, 200, moved to 300, the variance does decrease. However, we will also see that if the original mean was, say 248, so that the data point is moved further away from the original mean (but on the other side), the variance does not increase, but decreases. Seeing when and why this happens helps understand both the intuitive and non-intuitive aspects.

2 | QUALITATIVE AND QUANTITATIVE ASPECTS

First of all, it is clear that one can move the set of all values along the number line without affecting the variance. That means one can put—without loss of generality—the mean, \( \bar{x} \), to the zero position, \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 0 \). The \( n \)-th value, \( x_n \), is regarded as variable, the other \( x_i \) stay fixed. We are interested in the corresponding change, \( \Delta v \), when moving \( x_n \).

In the original variance, we separate the contribution of \( x_n \): 
\[
v = \left( \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) + x_n^2.
\]
Now, the \( n \)-th value is changed \( (x_n \rightarrow x_n + h) \), the resulting new statistics (mean, sum of squares of deviations from the mean) we denote by \( \bar{x}_{new} \) and \( v_{new} \). Then, we have \( \bar{x}_{new} = \frac{\bar{x}}{n} + \frac{h}{n} \) and 
\[
v_{new} = \frac{1}{n} \sum_{i=1}^{n-1} (x_i - \frac{h}{n})^2 + (x_n + h)^2 - \frac{h^2}{n},
\]
which simplifies via \( \sum_{i=1}^{n-1} x_i = 0 \) to 
\[
v_{new} = \sum_{i=1}^{n-1} x_i^2 + (x_n + h)^2 - \frac{h^2}{n}.
\]
We are interested in \( \Delta v = v_{new} - v \), if this is positive the variance increases, in case of \( \Delta v < 0 \) it decreases.

We get 
\[
\Delta v = (x_n + h)^2 - \frac{h^2}{n} - x_n^2 \]
or simplified
\[
\Delta v = h \left( 2x_n + \frac{n-1}{n} h \right)
\]
(1)

Here, one can see immediately: both cases, \( h, x_n > 0 \) and \( h, x_n < 0 \) yield \( \Delta v > 0 \).

Now, consider \( h \) and \( x_n \) of opposite signs, for example, \( x_n < 0 \) and \( h > 0 \). For values of \( h \) not too large, this is the situation of moving closer to the original mean. If the new data point, \( x_n + h \), is still on the same side of the original mean, that is, is \( < 0 \), we have \( h < -x_n \), and 
\[
2x_n + (n-1)h/n < 2x_n - (n-1)x_n/n < 0.
\]
This means that if the \( n \)-th data point moves closer to the original mean but stays on the same side, \( \Delta v < 0 \).

Hence intuition is confirmed in this case: whenever the moveable data point stays on the same side of the original mean and is moved away from it (brought closer to it), then the variance increases (decreases)!

Without our assumption, \( \bar{x} = 0 \), (which makes algebraic calculations easier) we would get instead of (1):
\[
\Delta v = h \left( 2\bar{x} - x_n + \frac{n-1}{n} h \right)
\]
(2)

Again, the intuition is confirmed in this general formula: if \( x_n \) is above the original mean (ie, \( x_n - \bar{x} > 0 \)) and you move it away from the mean (that means \( h > 0 \)), then the variance increases. Likewise, the same holds, of course, if \( x_n \) is below the original mean and you move it away from the original mean. The relations, (1) and (2), give not only qualitative answers to the question of the variance change but also quantitative ones.

In the following paragraphs, we do not use the assumption \( \bar{x} = 0 \) anymore, in order that readers need not answer the question by themselves: how are the conclusions made in the general case \( \bar{x} \neq 0 \)?

Note that the above considerations include the special case, \( x_n = \bar{x} \): if one data point which lies exactly at \( \bar{x} \) is changed, the variance increases: \( \Delta v = \frac{n-1}{n} h^2 > 0 \). This special case, \( x_n = \bar{x} \), could be dealt with even before one considers the above already more general case.

The algebraic workload that was necessary for the findings above is not high but still there is one. It would be nice if there were an easy and correct argument for the phenomenon without algebraic calculations (similar to the mean above), but unfortunately we did not find such a possibility. We are not sure whether there exists one.

From now on we will also consider motions of \( x_n \) which can go beyond the mean, and look what happens. Let us assume that \( x_n \) is below the original mean, \( \bar{x} \), and \( h > 0 \), which means the motion of \( x_n + h \) goes toward \( \bar{x} \) and even beyond it (we skip the presentation of the
analogous situation where $x_n$ is above the original mean and we move $x_n$ below it, as the details are nearly identical). Initially, $v$ will indeed decrease (we know already: at least until the new value, $x_n + h$, is at $\bar{x}$), but what happens afterwards?

From (2) we find that $\Delta v \geq 0$ if and only if

$$h \geq 2 \cdot \frac{n}{n-1} (\bar{x} - x_n) = 2(\bar{x} - x_n) + \frac{2}{n-1}(\bar{x} - x_n)$$

That is, we must move the $n$-th data point not to its symmetric counterpart with respect to the original mean (ie, to $x_n + 2(\bar{x} - x_n)$) but at least an additional $\frac{2}{n-1}(\bar{x} - x_n)$ beyond (see Figure 1).

For determining this additional amount explicitly, we had to use Formula (2). If one is not particularly interested in the exact amount of this additional shift, but looks for an intuitive explanation for the existence of such a shift, one could argue like this: *By increasing the data point $x_n$ we increase the mean, too.* Therefore, for producing the same $v$ the symmetric point, $x_n + 2(\bar{x} - x_n)$, can hardly be a candidate, one has to go even a bit beyond it.

Returning to the textbook question, which motivates this article, suppose that we know $x_n = 200$, $n = 20$, and $h = 100$, but we do not know the value of the original mean, $\bar{x}$. Below we consider two possible values of $\bar{x}$ to illustrate the result shown in Figure 1.

i. $\bar{x} = 320$. Hence (2) yields $\Delta v = -14500$. The data point has moved from 120 to 20 points distance from the original mean (on the same side!), and the sum of squares of deviations has decreased by 14500.

ii. $\bar{x} = 248$. Hence (2) yields $\Delta v = -100$. The data point has moved from 48 to 52 points distant from the original mean but on the other side, and the sum of squares of deviations has decreased by 100.

Hence the intuitive conjecture that moving a single point closer to the original mean decreases the variance is correct, no matter what the direction of movement is. Moving a single point further away from the original mean always increases the variance if the point stays on the same side. If the point is moved to the other side of the original mean, there is an initial interval of values further away from the original mean in which the variance decreases, but once the point is beyond this interval, the variance increases.

3 | COMMENTS OF INTEREST

$\Delta v$ is a quadratic (or parabola) in $h$ with zeroes at $h_1 = 0$, and $h_2 = \frac{2n}{n-1}(\bar{x} - x_n)$, and hence has a minimum at $h^* = \frac{n}{n-1}(\bar{x} - x_n)$. Therefore, $v_{new}$ is also a minimum there, where the data point has been moved to

$$x_n + h = x_n + \frac{n}{n-1}(\bar{x} - x_n) = \frac{n\bar{x} - nx_n}{n-1}$$

This is the mean of the other $n-1$ data points, $x_1, ..., x_{n-1}$.

This can also be expressed as $\bar{x} + \frac{\bar{x} - x_n}{n-1} = \bar{x} + \frac{h^*}{\pi}$, which is the new mean, $\bar{x}_{new}$.

That is, when the $n$-th data point is moved to the new mean, the variance is a minimum. The only way this can happen is as in Figure 1 (or the reverse of Figure 1, with the $n$-th data point moving from the right of the original mean to the left of it). The $n$-th data point moves to the right toward the original mean and then passes it, and the new mean also moves to the right, but the $n$-th data point “catches up” with the new mean and then “passes it”.

This minimum value of $v_{new}$ can be readily obtained in different ways.

First, if a data point is at the data mean (of any dataset), it contributes nothing to the sum of squares of deviations from the mean. In our case, (3) tells us this is the mean of the other data points. Hence the minimum value of $v_{new}$ is the sum of squares of deviations from the mean of the other $n-1$ data points, namely, $\frac{n-1}{n-1} \times \text{variance of the other } n-1 \text{ data points. (Aside: “variance” is the sample variance = the sum of the squared deviations divided by } n-1 \text{ in the case of } n \text{ data points).}$

Another way connects nicely with [1], which gave recursive formulae for mean and variance of a dataset of size $n$ when a new data point is added. If the added data point is the mean of the dataset of size $n$, then from [1] we have $n s^2_{n+1} = (n-1)s^2_n$. In our case, if we consider removing the $n$-th data point so that we have a sample of size $n-1$, and then adding a data point at the mean of this dataset of size $n-1$, we have, as above, $(n-1)s^2_n = (n-2)s^2_{n-1}$. 

4 | CONCLUSION

A problem of a school textbook led us to think of the effect of moving a data point more deeply for the first
time. The effects of moving, removing, or adding data points form part of a very large area of statistics called robust statistics, with extensive literature, but this school textbook problem is a very specific question.

It would be beautiful to be able to understand and argue for the mentioned phenomena without calculations (algebra), but we are not sure whether this is possible because sums of squared distances are not so easy for illustrations. Once more it is confirmed: even, intuitively, clear phenomena that seem rather likely need a certain extent of mathematics for valid argumentation.

The mathematics (in Section 2) involved in obtaining the answers to the question raised is middle school level, so is ideal to investigate an intuitive conjecture: whenever one moves a single data point closer to the mean (or moves the data point away from it), the variance will decrease (increase).

The result that the conjecture holds when the data point is moved closer to the original mean, and mostly holds when the data point is moved away from the original mean, demonstrates the importance of checking intuition. Examples illustrating the results, including when intuition does not hold, are readily accessible to middle school students. In addition, perhaps for older school students, the investigation leads to some interesting further considerations in Section 3.

ACKNOWLEDGEMENTS
I would like to thank the editor and the reviewers for valuable comments and suggestions which considerably improved the paper.

ORCID
Hans Humenberger https://orcid.org/0000-0001-6798-6933

REFERENCE
1. Yiu-Man Chan, Combining means and variances of samples, Teaching Statistics 16(3) (1994), 80.

How to cite this article: Humenberger H. How does the change of a single data point affect the variance, and why? Teaching Statistics. 2020;42: 87–90. https://doi.org/10.1111/test.12230