CURRICULUM INSPIRATIONS: www.maa.org/ci

MATH FOR AMERICA_DC: www.mathforamerica.org/dc

INNOVATIVE CURRICULUM ONLINE EXPERIENCES: www.gdaymath.com
TANTON TIDBITS: www.jamestanton.com

TANTON'S TAKE ON …

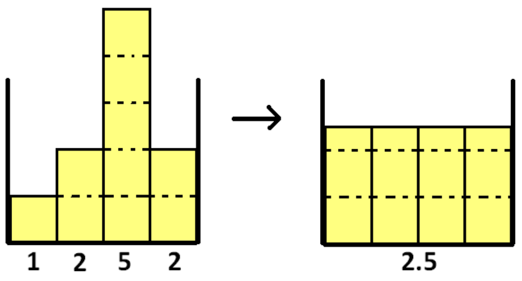
MEAN and VARIATION

JULY 2014

In our early grades we learn that the *average* of a collection of data measurements represents, in some way, a "typical" or "middle" value for the data. For example, the average of the numbers 1, 2, 5, 2 is:

$$\frac{1+2+5+2}{4}=2.5 .$$

Geometrically, the average is the level of a sand-box after we smooth out columns of sand of heights given by the data:



In a statistics class the average value of a collection of data values is called the *mean* of the data. (The word still means "average.") One denotes the mean of the data by putting a bar over whichever letter is being uses to denote the data. For example, the mean of a_1, a_2, and a_3 is:

$$\bar{a}=\frac{a_1+a_2+a_3}{3}$$

and the mean of $x_1, x_2, \ldots, x_n$ is:

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n} .$$

If the data set is extraordinarily large and one doesn't have any hope of determining the mean of the full data set, then that true, but unknown, mean is usually denoted with the Greek letter μ. For example, we have no hope of knowing the average height of all humans on this planet at this very moment. But we can measure the height of 1200 humans and collect 1200 data values $h_1, h_2, \ldots, h_{1200}$. We then hope that the sample mean $\overline{h}$ approximates the true mean μ to some reasonable degree.

Exercise: Data values x_1, x_2, $\cdots$, x_n have mean $\overline{x}$. Prove that the sum of the difference of each data value from the mean is sure to be zero:

$$\left(x_1 - \overline{x}\right) + \left(x_2 - \overline{x}\right) + \cdots + \left(x_n - \overline{x}\right) = 0 .$$

Exercise: Some texts might give the following formula for mean:

$$\overline{x} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_n x_n}{f_1 + f_2 + \cdots + f_n}$$

Can you interpret what the symbols in this formula mean and why the formula is correct?

SIMPSON'S PARADOX

Two students Albert and Bilbert each took a sample of math questions over a series of two days. There were 100 questions in total and Albert scored 65% and Bilbert 64% overall. Thus Albert proved himself a better test taker.

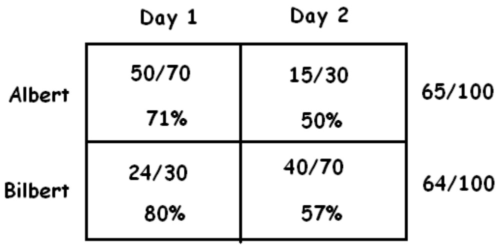
But here are the scores day-by-day:

FIRST DAY:
Albert = 71%
Bilbert = 80%

SECOND DAY:
Albert = 50%
Bilbert = 57%

So each day Bilbert did a better job than Albert, but did not prove to be the better test-taker after the two days combined! How is this possible?

The following table shows raw data of their test results.



	Day 1	Day 2	
Albert	50/70 71%	15/30 50%	65/100
Bilbert	24/30 80%	40/70 57%	64/100

This paradox arises because Albert and Bilbert did not complete the same number of questions each day and the averages computed are not equally weighted. This curious phenomenon is known as Simpson's paradox and was discovered by the Statistician Simpson in the 1960s after examining graduate school admission rates for men and women into UC Berkeley.

ASIDE:

There are several other measures of a "typical" or "central" value of a data set.

The **mode** of a set of data values is the value in the set that occurs most often (if there is one).

- For the ten data values 3, 6, 5, 3, 1, 6, 5, 3, 8, 3 the mode is 3.
- The data set 5, 5, 6, 6, 9, 9, 3, 3, 2, 2 has no mode.
- The data set 1, 1, 1, 1, 5, 5, 7, 7, 7, 7, 8, 8, 9, 9, 9 is <u>bimodal</u>.

(Is the second example quinti-modal?)

For non-numerical data, such as colours, or letters of the alphabet, the mode is the only measure of central tendency available.

If we arrange the data set in increasing order of values, then the **median** of the data is the middle value of the ordered sequence or the average of the two middle values if there are an even number of terms.

- The median of the data set 3, 3, 5, 6, **7**, 16, 16, 19, 37 is 7.
- The median of 3, 4, 4, **5**, **8**, 8, 10, 12 is $\frac{5+8}{2}=6.5$.

The median is a value that divides the data set into two equally sized groups.

The **midrange** of a data set is the average of the smallest and largest values.

- The midrange of the data set 5, 6, 9, 9 is $\frac{5+9}{2}=7$.

The midrange provides a quick estimate to a central value. It is easy to compute, but is highly affected by extremely low or high values in the data set.

Exercise: a) Find FIVE data values with:

Median = 10
Mode = 10
Mean = 1000

b) Now find five data values with median = 10, mode = 1000 and mean = 10.

c) Can you find five data values with median = 1000, mode = 10, mean = 10?

COOL Exercise: Repeat the previous exercise but this time for SIX data values.

DEVIATION FROM THE MEAN

The data set $1,2,5,2$ has mean 2.5. So too does the data set: $-101,\ 0,\ 1,\ 110$. These are two very different data sets, with the second being much more "spread out" than the first. We can measure the degree of spread by calculating the average deviation from the mean for each.

DATA SET $1,2,5,2$:

Deviations:

$$|1-2.5|=1.5$$
$$|2-2.5|=0.5$$
$$|5-2.5|=2.5$$
$$|2-2.5|=0.5$$

Average deviation:

$$\frac{1.5+0.5+2.5+0.5}{4}=1.25\ .$$

DATA SET $-101,\ 0,\ 1,\ 110$:

Deviations:

$$|-101-2.5|=103.0$$
$$|0-2.5|=2.5$$
$$|1-2.5|=1.5$$
$$|110-2.5|=107.5$$

Average deviation:

$$\frac{103.0+2.5+1.5+107.5}{4}=28.625$$

The numbers 1.25 and 28.625, the average deviations from the mean, do give a quantitative measure of the amount of "spread" of each data set.

THE POINT OF THIS ESSAY

Using the absolute value, the distance of a particular data value from the mean value of the data, is the natural and appropriate way to measure data variation. **But statisticians DON'T use absolute values in their work**! This is very strange and confusing for students. (There is also a second piece of confusion, which we shall leave to later in this essay.)

Here are two rationales for the switch away from absolute values:

RATIONALE ONE:
Working with absolute values is hard. Can we avoid them?

Indeed, working with absolute values in mathematical equations is really tough!

Optional Exercises:

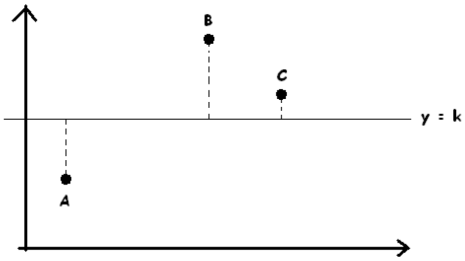
a) Sketch the "curve" $|x-1|+|y-2|=3$.

b) Find all values of w which satisfy:

$$\big|\,|w-2|-3w\,\big|-|5-w|=7\,.$$

c) (From last month's essay)

Three data points $A=(2,3)$, $B=(5,8)$ and $C=(7,5)$ are plotted on a graph.



A horizontal line $y=k$ will be drawn but a value k needs to be chosen so that the sum of the three vertical deviations from the horizontal line is at a minimum.
(NOTE: We've drawn the horizontal line so that A lies below it and B and C above it. This need not be the case.)

On a calculator, type in a function that represents the sum of these three deviations and graph it.

Which value of k seems to give a minimum value for this sum of three deviations?

But we still need a measure, a positive number that represents the deviation of each data value from the mean. If we want to avoid absolute value, how else can we obtain positive values? Answer: Square the values!

Let's square all the deviations and take the average of those squared deviations:

DATA SET $1,2,5,2$:

Deviations squared:

$$(1-2.5)^2=2.25$$
$$(2-2.5)^2=0.25$$
$$(5-2.5)^2=6.25$$
$$(2-2.5)^2=0.25$$

Average squared deviation:

$$\frac{2.25+0.25+6.25+0.25}{4}=2.25\,.$$

DATA SET $-101,\ 0,\ 1,\ 110$:

Deviations squared:

$$(-101-2.5)^2=10609$$
$$(0-2.5)^2=6.25$$
$$(1-2.5)^2=2.25$$
$$(110-2.5)^2=11556.25$$

Average squared deviation:

$$\frac{10609+6.25+2.25+11556.25}{4}$$
$$=5543.4375$$

These average squared deviations still give a good sense of the different spreads the two data sets possess.

One subtle point: Data often comes from physical measurements – the height of a person, the speed of a car on a highway, and so on – and so has units associated with them.

If x_1, x_2, … , x_n are in units of inches, say, then the mean $\overline{x}=\dfrac{x_1+x_2+\cdots+x_n}{n}$ also has units of inches, but the average squared deviation:

$$\frac{\left(x_1-\overline{x}\right)^2+\left(x_2-\overline{x}\right)^2+\ldots+\left(x_n-\overline{x}\right)^2}{n}$$

has units of inches squared. To bring all quantities and comparisons between quantities back to the same units, statisticians will take the square root of the average squared deviation:

$$\sqrt{\frac{\left(x_1-\overline{x}\right)^2+\left(x_2-\overline{x}\right)^2+\ldots+\left(x_n-\overline{x}\right)^2}{n}}$$

This quantity now has units of inches and is called the **standard deviation** of the data.

WARNING: Statisticians might raise an eyebrow or two over at what I just said. They might prefer to call the quantity:

$$\sqrt{\frac{\left(x_1-\overline{x}\right)^2+\left(x_2-\overline{x}\right)^2+\ldots+\left(x_n-\overline{x}\right)^2}{n-1}}$$

the standard deviation of the data set. (Note "$n-1$" in the denominator, rather than "n.") This change - the second confusion for students studying statistics - is discussed at the end of this essay.

RATIONALE TWO:
Abstract mathematics tells us it is natural to work with quantities squared.

Suppose we run an experiment or poll some people and gain from the exercise n data values:

$$x_1, x_2, \ldots, x_n .$$

We, not being omniscient, know nothing about the data values we shall obtain: we don't know what to expect for the mean of the values (what is the true average height of all humans on this planet?), what variation from the mean to expect, what the frequencies of particular values should be, and so on.

But if the experiment was ideal or the population we were polling from is truly uniform, then the experiment or polling would be absolutely and utterly repeatable and we'd expect no variation in data values at all. That is, in the perfect ideal, all measurements would adopt exactly the same value q, say, over and over again.

Let's ask: *How close is our data* $\left(x_1, x_2, \ldots, x_n\right)$ *from some ideal set of repeatable data* $\left(q, q, \ldots, q\right)$*?*

Now we learned last month that, in two-dimensional geometry, the distance between two points $A=\left(a_1, a_2\right)$ and $B=\left(b_1, b_2\right)$ is given by:

$$d\left(A,B\right)=\sqrt{\left(a_1-b_1\right)^2+\left(a_2-b_2\right)^2} .$$

And the distance between two points $A=\left(a_1, a_2, a_3\right)$ and $B=\left(b_1, b_2, b_3\right)$ in three-dimensional space is:

$$d\left(A,B\right)=\sqrt{\left(a_1-b_1\right)^2+\left(a_2-b_2\right)^2+\left(a_3-b_3\right)^2}$$

And so on, for any dimension of space.

So to answer this question we seek a value q so that the point $M=\left(q, q, \ldots, q\right)$ is as close as possible to our point $P=\left(x_1, x_2, \ldots, x_n\right)$ in n-dimensional geometry.

We want to choose a value q that minimizes the distance:

$$d\left(P,M\right)=\sqrt{\left(x_1-q\right)^2+\left(x_2-q\right)^2+\cdots\left(x_n-q\right)^2}$$

It is easier to just to minimize the quantity under the square root sign. Notice that we are now led to study a sum of quantities squared.

Expand the sum under the root and collect terms:

$$\begin{aligned}&(x_1-q)^2+(x_2-q)^2+\cdots(x_n-q)^2\\&\quad=nq^2-2(x_1+\cdots+x_n)q+\left(x_1{}^2+\cdots+x_n{}^2\right)\end{aligned}$$

We see that the sum we wish to minimize is just a quadratic in q. It has minimum value for:

$$q=\frac{2(x_1+\cdots+x_n)}{2n}=\frac{x_1+\cdots+x_n}{n}=\overline{x}$$

- the data’s mean!

We have:

The mean of a data set is the value of closest ideal, repeatable, experiment to the given data.

From this perspective we see that it is natural to think about sums of deviations squared. Dividing through by n, we call:

$$\frac{\left(x_1-\overline{x}\right)^2+\cdots+\left(x_n-\overline{x}\right)^2}{n}$$

the **variance** of the data. And to match units, we take the square root and call this the **standard deviation** of the data:

$$\sqrt{\frac{\left(x_1-\overline{x}\right)^2+\cdots+\left(x_n-\overline{x}\right)^2}{n}}\ .$$

Comment: We have now seen that the mean $\overline{x}$ of a set of data values $x_1,x_2,\ldots,x_n$ has two properties:

i) The sum

$$\left(x_1-\overline{x}\right)+\left(x_2-\overline{x}\right)+\cdots+\left(x_n-\overline{x}\right)$$

is zero.

ii) Of all the sums of the form:

$$(x_1-q)^2+(x_2-q)^2+\cdots(x_n-q)^2.$$

the sum

$$\left(x_1-\overline{x}\right)^2+\left(x_2-\overline{x}\right)^2+\cdots\left(x_n-\overline{x}\right)^2$$

has the smallest value.

ON n VERSUS $n-1$

Some text authors will argue that it is better to divide by $n-1$ in the formulas for variance and standard deviation rather than by n for the following philosophical reason:

We have that

$$\left(x_1-\overline{x}\right)+\left(x_2-\overline{x}\right)+\cdots+\left(x_n-\overline{x}\right)$$

is sure to equal zero. This means that if one knows the first $n-1$ values $x_1-\overline{x},\ x_2-\overline{x},\ \ldots,\ x_{n-1}-\overline{x}$, then the value of the n th one, $x_n-\overline{x}$, is forced.

So among the values

$$\left(x_1-\overline{x}\right),\left(x_2-\overline{x}\right),\ldots,\left(x_n-\overline{x}\right)$$

there are only $n-1$ real pieces of information. To reflect this, let’s divide by $n-1$ rather than n and set the variance as:

$$\frac{\left(x_1-\overline{x}\right)^2+\cdots+\left(x_n-\overline{x}\right)^2}{n-1}$$

and the standard deviation as:

$$\sqrt{\frac{\left(x_1-\overline{x}\right)^2+\cdots+\left(x_n-\overline{x}\right)^2}{n-1}}\ .$$

But this seems unsatisfactory an explanation.

Text authors will often add:

If the data sets are large, that is, if n is a large number, then there will be little difference in dividing through by $n-1$ over dividing through by n.

The correct student response to this add on is: “So, really, why bother making this change?”

To understand why statisticians prefer to divide by $n-1$,not n, let’s go back to a previous example.

Because the data set is so large, we have no hope of knowing the true average height μ of all humans on this planet right at this moment. All we can do is measure the heights of a sample of humans, compute the data mean $\overline{h}$ of that sample, and hope that $\overline{h}$ offers a good approximation for μ.

We would expect there to be some uniformity among all the possible samples we could work with. Certiainly, if we select a sample of 1200 humans and measure their heights we would obtain a sample mean $\overline{h}$. If we chose a different collection of 1200 people we would probably obtain a slightly different mean $\overline{h}$. In fact, if we looked at every possible collection of 1200, we'd have a whole spread of values for $\overline{h}$, all approximating the true mean value μ. Since the set of all samples of 1200 humans well and truly covers the entire human population, it would be a shock if, on average, the set of all possible values of $\overline{h}$ turned out to be different from μ.

The same should be true for variance. We can't possibly know the true variance of the entire set of human population heights, but we can take a sample of 1200 heights and find the value of the variance for that sample. And it would be a shock if again, on average, the variances over all possible samples of 1200 people turned out to be a value different from the true variance of the entire population.

Let's see what can happen with some actual numbers.

EXAMPLE: Consider the data set $1,2,2,3$.

This is a set of $n=4$ data values with "true" mean $\mu=2$ and true variance, when dividing by $n=4$:

$$V_n=\frac{(1-0)^2+(2-2)^2+(2-2)^2+(3-2)^2}{4}$$

$$=\frac{1}{2}$$

and true variance, when dividing by $n-1=3$:

$$V_{n-1}=\frac{(1-0)^2+(2-2)^2+(2-2)^2+(3-2)^2}{3}$$

$$=\frac{2}{3}$$

But suppose we don't know these values – a data set of four values is too large for us to manage – so we decide to look at samples of size three instead and work out their sample means and sample variances.

Here is a table of all possible subsets of size three (handling the repeated 2 s) and the sample means and variances we would see:

	$\overline{x}$	V_n	V_{n-1}
$\{1,2,2\}$	$5/3$	$\frac{4/9+1/9+1/9}{3}$ $=2/9$	$\frac{4/9+1/9+1/9}{2}$ $=1/3$
$\{1,2,3\}$	2	$4/9$	1
$\{1,2,3\}$	2	$4/9$	1
$\{2,2,3\}$	$7/3$	$2/9$	$1/3$
Average	**2**	**1/3**	**2/3**

We see that the means and the variances do depend on which sample of three you happen to choose.

We also see, in this example, that our first dream is true: the average of all the sample means matches $\mu=2$ on the nose.

And our second dream is true too if we divide by $n-1$ instead of n when computing variances: the average of the values of V_{n-1} over all samples matches the value of V_{n-1} for the overall data.

These two claims are not a coincidence for our particular example: they are true in general. It is for this reason that statisticians prefer to work with the formula:

$$\frac{\left(x_1-\overline{x}\right)^2+\cdots+\left(x_n-\overline{x}\right)^2}{n-1}$$

for variance and the square root of this for standard deviation.

Exercise: There are six two-element subsets of the data set $1,2,2,3$ (if you handle the repeated 2 s appropriately). List all six subsets, compute the mean and variance V_{n-1} of each, and take the average value of these six means and variances. Show these average values match $\mu=2$ and $V_{n-1}=2/3$ of the original data set.

MATHEMATICAL PROOFS:

The mathematics here is tedious algebra and is hard to read. One can phrase the algebra in terms of expected values and variances of random variables ($E(X)$ and $Var(X)$) and make matters less complicated visually, but one does this at the price of obscuring the conceptual straightforwardness.

If you are game, here's how these proofs proceed.

Suppose a population possesses a total of N data points and has mean:

$$\mu=\frac{y_1+y_2+\cdots+y_N}{N}.$$

Our job is to look at a subset of n data points, $x_1,x_2,\ldots,x_n$, compute their data mean $\overline{x}$, and take the average of all possible values for $\overline{x}$ over all possible subsets and show this average equals μ. We must also compute the variances

$$\frac{\left(x_1-\overline{x}\right)^2+\cdots+\left(x_n-\overline{x}\right)^2}{n-1}$$

over all subsets and show that their average equals:

$$\frac{\left(y_1-\mu\right)^2+\cdots+\left(y_N-\mu\right)^2}{N-1}.$$

Now there are ${}_N C_n=\dfrac{N!}{n!\left(N-n\right)!}$ subsets of size n among N data points, so in each case, our average is a sum divided by this number.

For the sample means we need to show:

$$\frac{\dfrac{x_1+x_2+\cdots+x_n}{n}+\dfrac{x'_1+x'_2+\cdots+x'_n}{n}+\cdots}{\dfrac{N!}{n!\left(N-n\right)!}}$$

equals μ where the numerator is the sum of sample means over all possible subsets. (There is a similar, but more complicated formula, for the average of the variances.)

This expression is equivalent to:

$$\frac{(n-1)!(N-n)!}{N!}\left(\left(x_1+x_2+\cdots+x_n\right)+\left(x'_1+x'_2+\cdots+x'_n\right)+\cdots\right)$$

Now a particular data point x appears in ${}_{N-1}C_{n-1}=\dfrac{(N-1)!}{\left(n-1\right)!(N-n)!}$ subsets of size n. So in the sum we have each data point mentioned this many times. Our expression is thus equivalent to:

$$\frac{(n-1)!(N-n)!}{N!}\left(\frac{\left(N-1\right)!}{\left(n-1\right)!\left(N-n\right)!}x+\frac{\left(N-1\right)!}{\left(n-1\right)!\left(N-n\right)!}y+\cdots\right)$$

where the sum is over each and every data point in the set.

This simplifies to:

$$\frac{1}{N}\left(x+y+\cdots\right)$$

which is indeed μ !

For the average value of the variances, we need to work with:

$$\frac{N!}{n!(N-n)!}\left(\begin{array}{l}\dfrac{\left(x_1-\overline{x}\right)^2+\cdots+\left(x_n-\overline{x}\right)^2}{n-1}\\ +\dfrac{\left(x'_1-\overline{x'}\right)^2+\cdots+\left(x'_n-\overline{x'}\right)^2}{n-1}\\ +\cdots\\ \end{array}\right)$$

This is equivalent to:

$$\frac{N!}{(n-1)n!(N-n)!}\left(\begin{array}{l}\left(x_1-\overline{x}\right)^2+\cdots+\left(x_n-\overline{x}\right)^2\\ \quad+\left(x'_1-\overline{x'}\right)^2+\cdots+\left(x'_n-\overline{x'}\right)^2\\ \quad+\cdots\end{array}\right)$$

$$=\frac{N!}{(n-1)n!(N-n)!}\times$$

$$\left(\begin{array}{l}\left(x_1-\frac{1}{n}\left(x_1+\cdots+x_n\right)\right)^2+\cdots+\left(x_n-\frac{1}{n}\left(x_1+\cdots+x_n\right)\right)^2\\ \quad+\left(x'_1-\frac{1}{n}\left(x'_1+\cdots+x'_n\right)\right)^2+\cdots+\left(x'_n-\frac{1}{n}\left(x'_1+\cdots+x'_n\right)\right)^2\\ \quad+\cdots\\ \end{array}\right)$$

$$=\frac{N!}{n(n-1)n!(N-n)!}\times$$

$$\left(\begin{array}{l}\left((n-1)x_1-x_2-\cdots-x_n\right)^2+\left(-x_1+(n-1)x_2-\cdots-x_n\right)^2+\cdots\\ \quad\left((n-1)x'_1-x'_2-\cdots-x'_n\right)^2+\left(-x'_1+(n-1)x'_2-\cdots-x'_n\right)^2+\cdots\\ \quad+\cdots\end{array}\right)$$

By expanding terms and counting how many times a particular data point squared $x_1^{\ 2}$ appears and how many times the pair x_1x_2 appears (and these counts are the same for all data points), one can show that this expression does indeed equal:

$$\frac{\left(y_1-\mu\right)^2+\cdots+\left(y_N-\mu\right)^2}{N-1},$$

the variance over all the data points.

We'll leave the details to the truly gung-ho reader!

Exercise: To get a (manageable) feel for the algebra, do work through the details for the case of $N=4$ data points: x_1,x_2,x_3,x_4. Write down and simplify the formulas for the variances of each of the subsets $\{x_1,x_2,x_3\}$, $\{x_1,x_2,x_4\}$, $\{x_1,x_3,x_4\}$, $\{x_2,x_3,x_4\}$ and an expression for the average of these four values. Show this average equals:

$$\frac{1}{3}\left(\begin{array}{l}\left(x_1-\dfrac{x_1+x_2+x_3+x_4}{4}\right)^2\\ +\left(x_2-\dfrac{x_1+x_2+x_3+x_4}{4}\right)^2\\ +\left(x_3-\dfrac{x_1+x_2+x_3+x_4}{4}\right)^2\\ +\left(x_4-\dfrac{x_1+x_2+x_3+x_4}{4}\right)^2\end{array}\right).$$