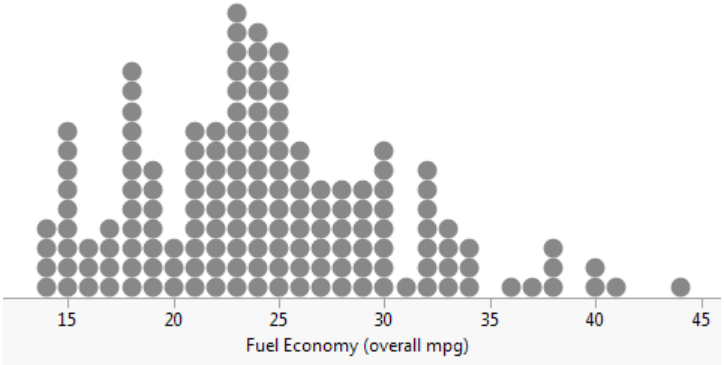
More from the Stats Fan…
What, Exactly, IS a Standard Deviation, Anyhow?

Ruth Miller, Kent Denver School in Englewood, CO
rmiller@Kentdenver.org

Data is, in its raw form, an ugly, often large, and unwieldy set of facts, usually represented numerically. But in that form, data is impossible to work with, so we try to organize it.
Data falls into a distribution, which we can look at as a histogram or dot plot (for example); there are lots of ways to display data (maybe this is the subject of another article).



But even this is difficult to use, we can see a shape, and make some statements, like, "The best fuel economy in this data set is just a little less than 45mpg." But we want some consistent ways to describe distributions of data, and the ones that statisticians have settled on are *Center*, *Spread*, *Shape*, and *Unusual Features*. We quantify Center by using the mean (average) or the median (middle data point). Shape has to do with symmetry and the clusters that exist in the distribution, and Unusual Features are things like gaps or outliers. But today's topic is a measure of spread, the *Standard Deviation.*

Some things to know first: You really can't talk about measures of center without also talking about measures of spread, and there are two ways to do this, the median is paired with a measure of spread called the inter-quartile range, and the mean (average) is paired with the standard deviation. Every teacher understands how to find the average, but the standard deviation is tricky to calculate, and so we don't always attach it to our discussion of the mean. But we should! Because if I tell you that the average score on a test was 72%, and you earned 76%, that might mean you scored above average, but it might not, *depending on the standard deviation*.

"Deviation" means difference, and "Standard" means usual. The standard deviation is essentially the average space between the mean of a data set and the individual date points. If

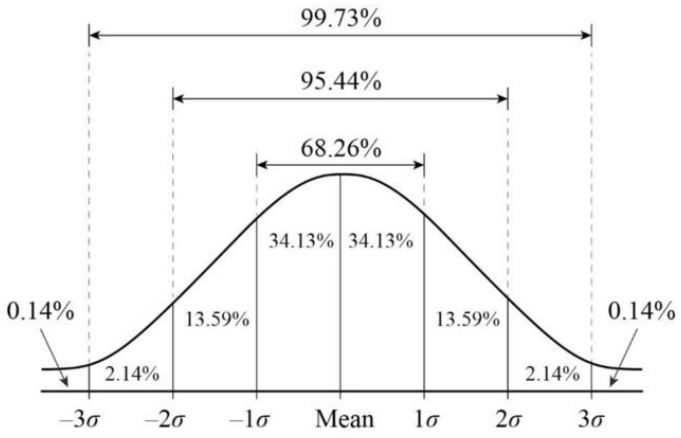
we revisit the dot plot above, the mean of the data points is $\bar{x} = 24.4$. ($\bar{x}$ is pronounced "*x*-bar" and is the symbol that we use to denote the mean of our distribution.) Let's consider a few of our individual data points (called the "*x*, sub *i*'s").

x_i	$\bar{x}$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$\sqrt{(x_i - \bar{x})^2}$
15	26.7	$(15 - 24.4)$	88.125	9.3
21	26.7	$(21 - 24.4)$	11.475	3.4
24	26.7	$(24 - 24.4)$	.15016	1.4
27	26.7	$(27 - 24.4)$	6.8252	2.6
31	26.7	$(31 - 24.4)$	43.725	6.6
41	26.7	$(41 - 24.4)$	275.98	16.6
44	26.7	$(44 - 24.4)$	384.65	19.6

The difference between the mean $(\bar{x})$ and the individual data points $(x_i's)$ is in the third column. We'd like to find the average amount that the $x_i's$ deviate from the mean, so maybe we just add them all up and divide by however-many-there-are? But there is a problem- Because of what the mean is, a measure of center, the sum of all these differences will be zero, some are lower than the mean and others are higher, the sum ends up being zero. So, we essentially take the absolute value of these differences by squaring and then rooting them. The last column in the chart represents the distance between each x_i and the mean. These are the values that we'll take the average of, to establish the usual difference (or *standard deviation*) of the $x_i's$ from the mean.

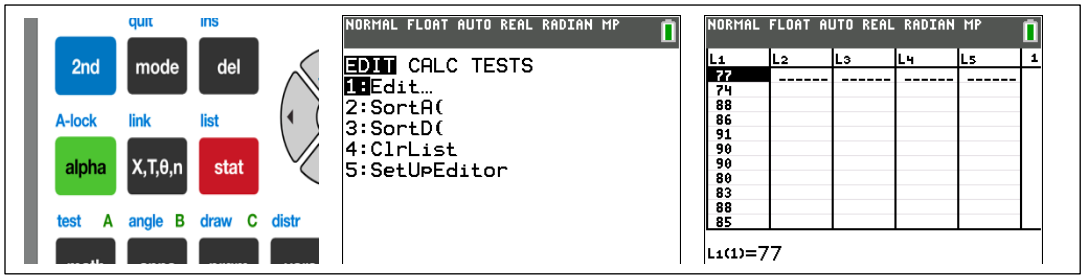
The equation for the standard deviation of a sample is $s_x = \sqrt{\frac{\sum_1^N (x_i - \bar{x})^2}{N}}$. Look at the symbols, see how they tell us to add up all the individual deviations and divide by however-many-there-are? And see how we have avoided the problem of the sum adding to zero by squaring and rooting? Cool, no?

So, what does this tell us? In a Normal Distribution (the "Bell Curve") about $\frac{2}{3}$ of the data will fall between the first standard deviation from the mean. An additional 28% falls (14% on either side of the mean) within 2 standard deviations, and virtually all the rest of the data falls within the next standard deviation. About .3% will fall outside of the third Standard Deviation, and that is considered to be very unusual data.

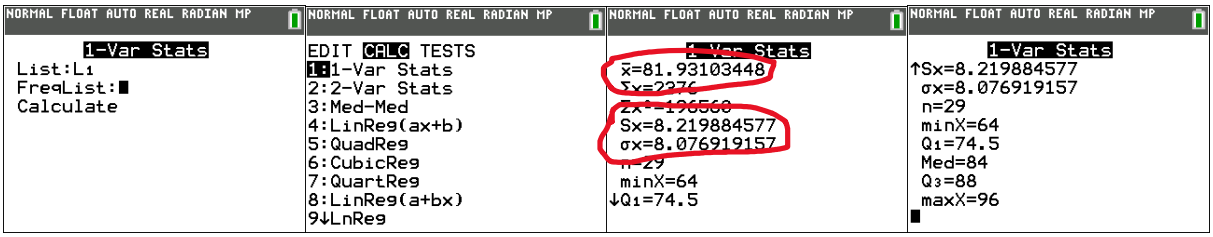


So back to the question of whether my 76% on a test with a mean score of 72% was "above average" or not. This entirely depends on the standard deviation. If the mean is 72% and the standard deviation is 2, then about $\frac{2}{3}$ of the class scored between 70% and 74%, and my score is truly above average. But if the standard deviation on the test was 5 points, then about $\frac{2}{3}$ of the class scores between 67% and 77%, and I did OK, but not really "above average." A working statistician will never talk about a measure of center without adding a measure of spread, because without the spread, the center has no context.

The calculator will give you all of this information, and more!
STAT EDIT will allow you to enter your data (here are some recent test scores).



Then, back to the STAT button, to CALC, to 1-Var(iable) Stats, and the calculator will report the mean, $\bar{x}$, and two versions of the standard deviation, s_x and σx. The Greek symbol σx represents an estimate of the standard deviation for a population, and the English symbol s_x represents the standard deviation of the particular data you have stored in the calculator.



I can see that on this test, the average score was about 82%, with a standard deviation of about 8 points. About $\frac{2}{3}$ of the kids scored between 74% and 90%, which is $\bar{x} \pm s_x$, and almost all of students will have scored between $\bar{x} \pm 2s_x$, or between 66% and 98%. A student in the 90's, on this test, will have done "better than average, and one in the low 70's may cause me some concern. This information helps me to decide whether I want to "curve" the test, but that, too, may be a topic for another time ☺

By the way, the equation for the Normal Curve is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, and you can see that both μ and σ (the mean and standard deviation of the population) are constants that affect the parent curve via transformations. This equation was developed by Carl Gauss, a German born around the time of the American Revolution. In Germany it's not called the Normal Curve,

rather it's called the Gauss Curve! Gauss (and his curve) were featured on a 10 Mark note before the advent of the Euro.

Feel free to contact me with any questions or suggestions for further snippets-of-Stats!
rmiller@kentdenver.org
@rm11235813 on twitter ☺

AND: #shamelessplug; I will be leading some Calculus APSI's this summer; I'd love to "see" some Michigan peeps!

AP Calculus AB/ BC online 7/6 – 7/9
https://www.hpu.edu/cps/outreach/apsi/index.html

AP Calculus BC online 7/12 – 7/16
https://www.utep.edu/extendeduniversity/professional-and-public-programs/programs/advanced-placement.html

AP Calculus BC online 7/19 – 7/23
https://www.augsburg.edu/ecs/apsi/mpls/

Ruth Miller is the Head of the Math Department at Kent Denver School in Englewood, CO. From 2011 to 2019, Ruth was the Chair of the Department of Mathematics and Computer Science at Greenhills School in Ann Arbor, and Ruth also spent part of that time as the Region II Representative to the Michigan Council of Teachers of Mathematics Board of Directors. She is an AP Calculus Reader, Table Leader, and Consultant, and has also taught AP Stats off and on since the course's inception.