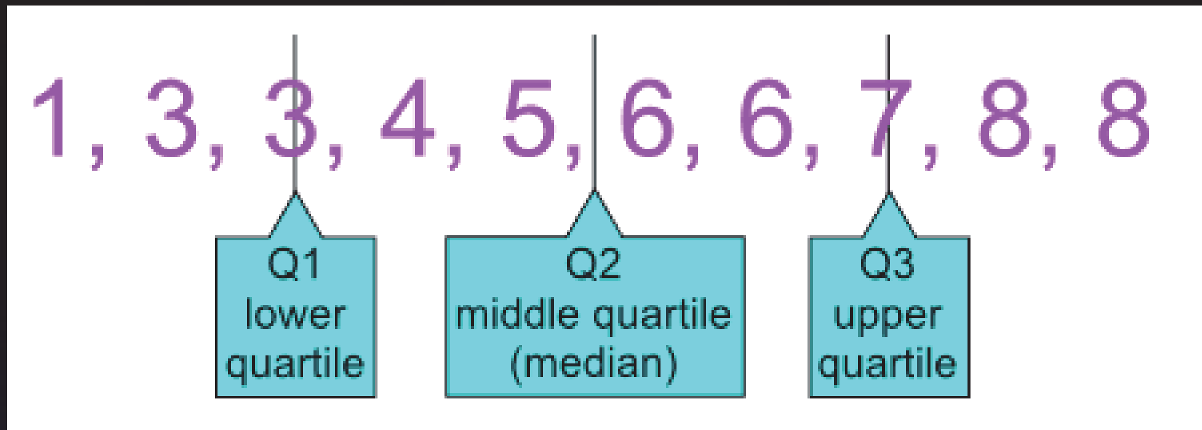
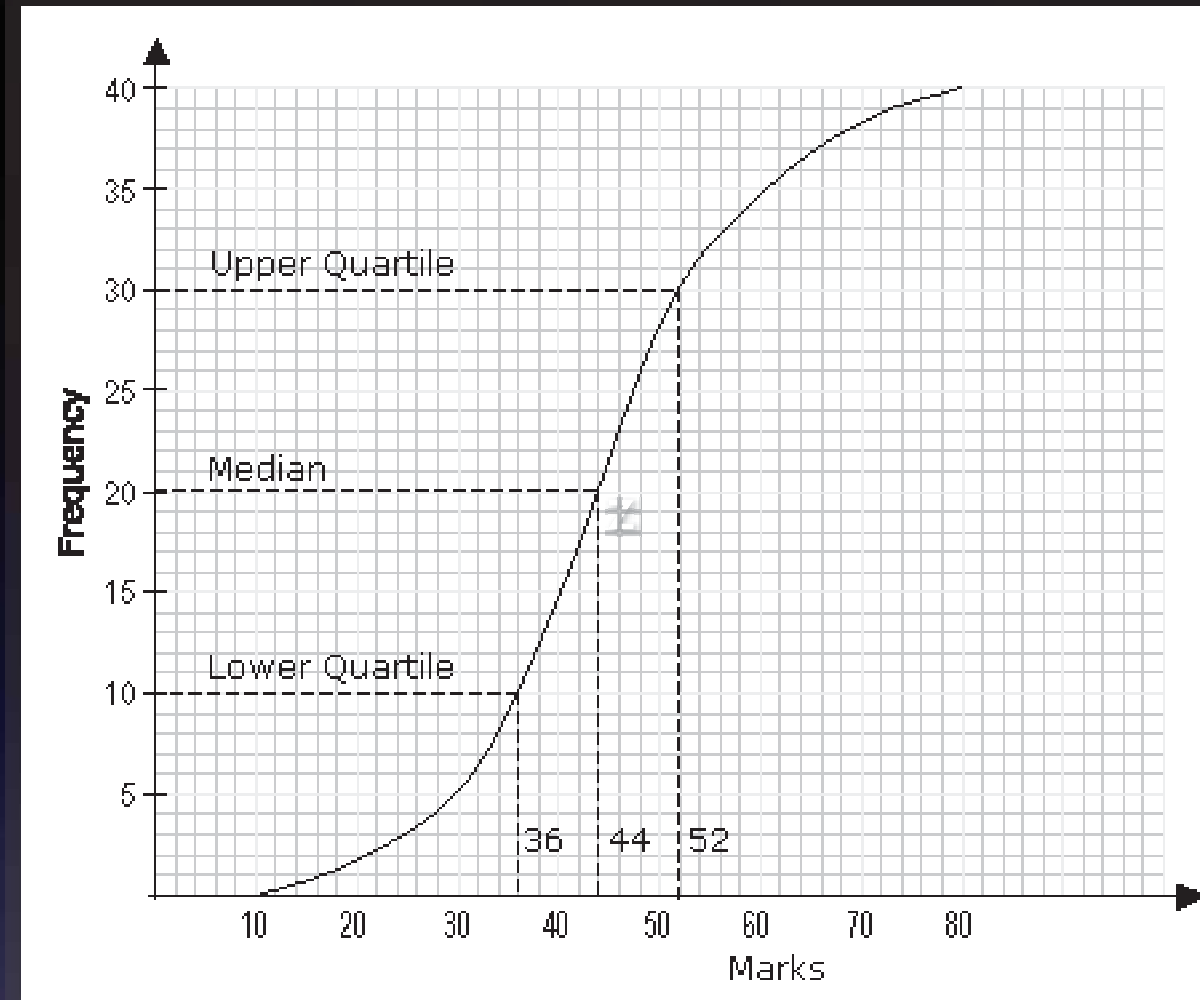
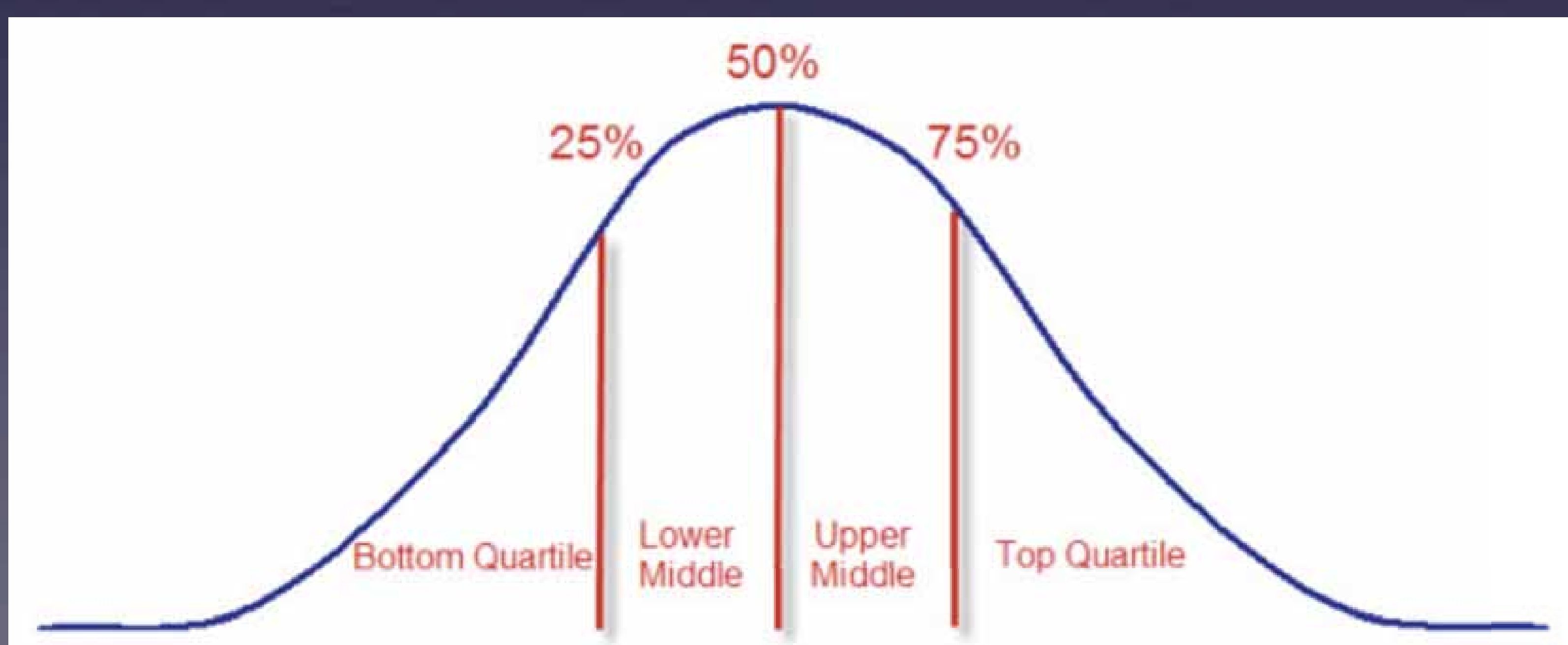


1, 3, 3, 4, 5, 6, 6, 7, 8, 8
Q1 lower quartile
Q2 middle quartile (median)
Q3 upper quartile



Frequency
40
35
30
25
20
15
10
5
Upper Quartile
Median
Lower Quartile
36
44
52
10
20
30
40
50
60
70
80
Marks



50%
25%
75%
Bottom Quartile
Lower Middle
Upper Middle
Top Quartile

Range
Spread
IQR = $Q_3 - Q_1$
Standard deviation - s, σ
Variance - s^2, σ^2

Measures of Spread/Variability

How does Standard Deviation differ from IQR?

Lower Quartile - Q_1
Upper Quartile - Q_3
2nd Quartile = Median

Measures of Position

Remember this data?
Use 1-Var-Stats to find center and spread.

Question: Why "Standard" Deviation? Why not just "Average" Deviation?

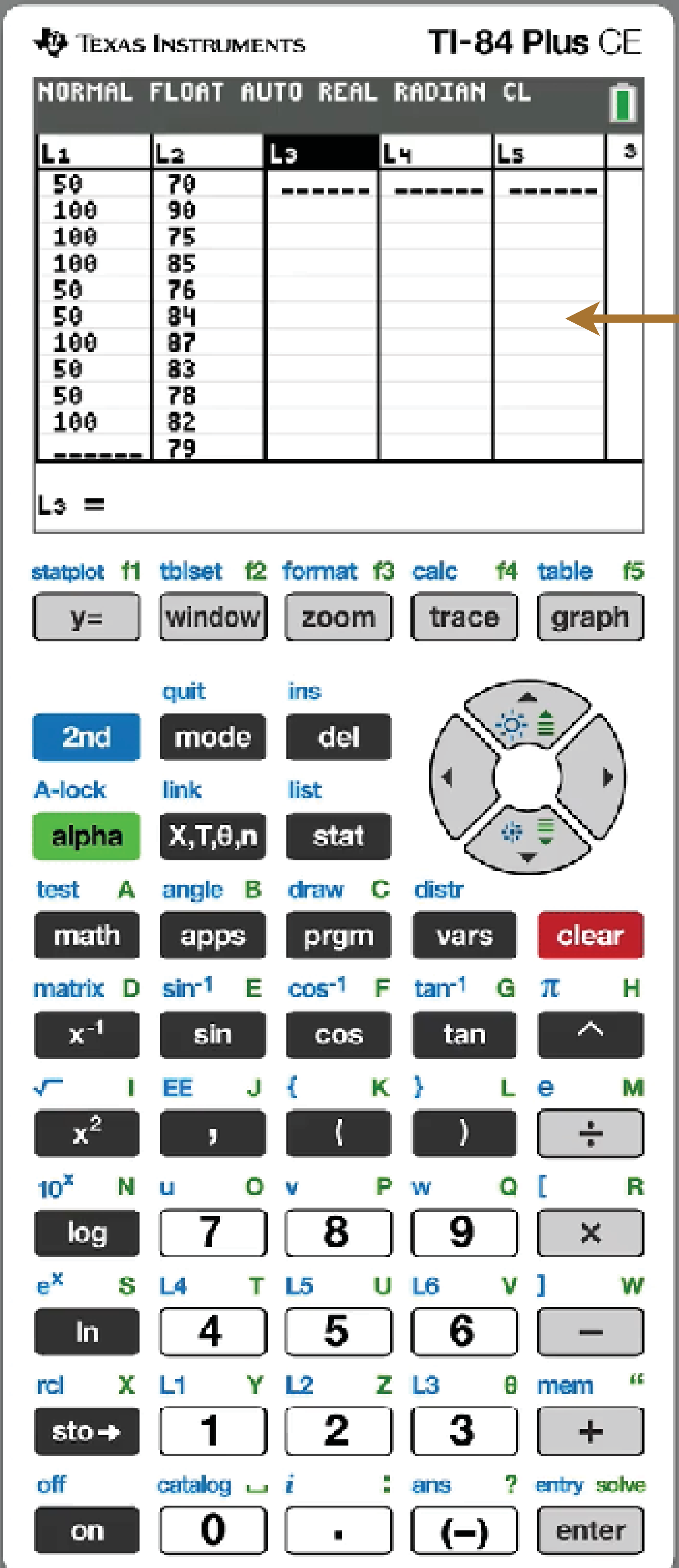
Here's the formula for the average deviation:

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n}$$

For Test 1, the mean is 75 so it would look like this:

$$\frac{(50-75)+(100-75)+(100-75)...}{n} = ?$$

Test 1	Test 2
50	70
100	90
100	75
100	85
50	76
50	84
100	87
50	83
50	78
100	82
	79
	81
	80
	80
	80
	88

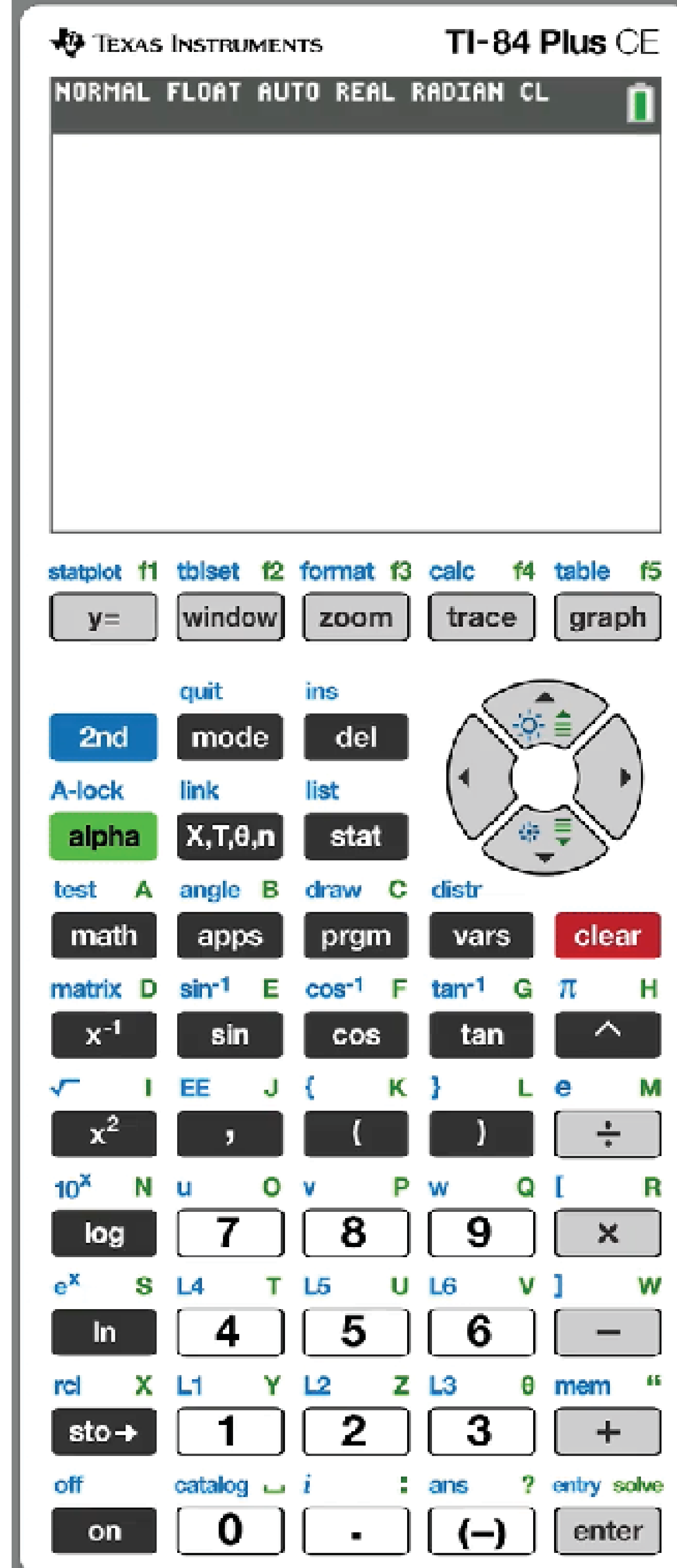


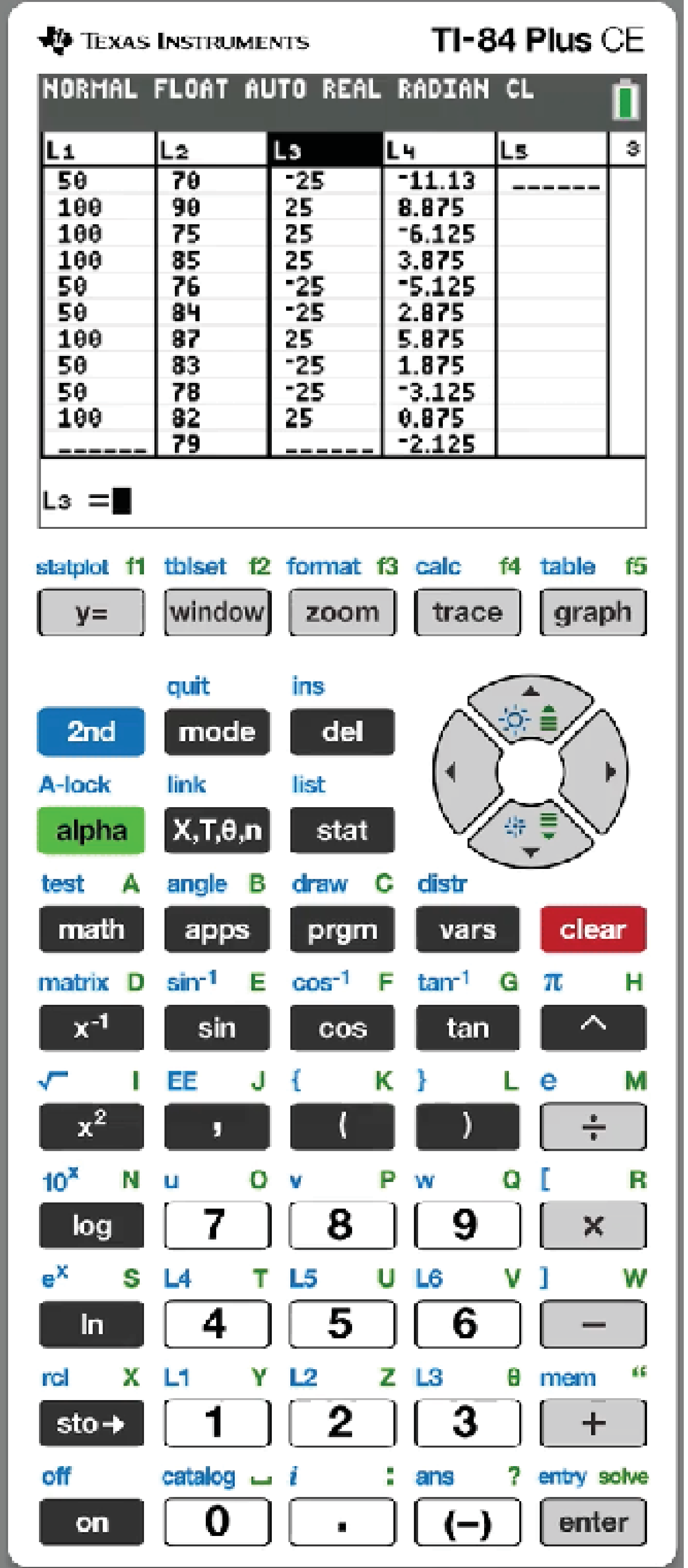
Let’s try it on our calculators

If we just calculated average deviation, we would subtract each score from the mean like this

But look what happens when we add them up to average them

So this method won’t work arithmetically.





Turns out that squaring each term to make it positive and then adding helps.

We can then add, divide by the sample size, and then take the square root giving us this:

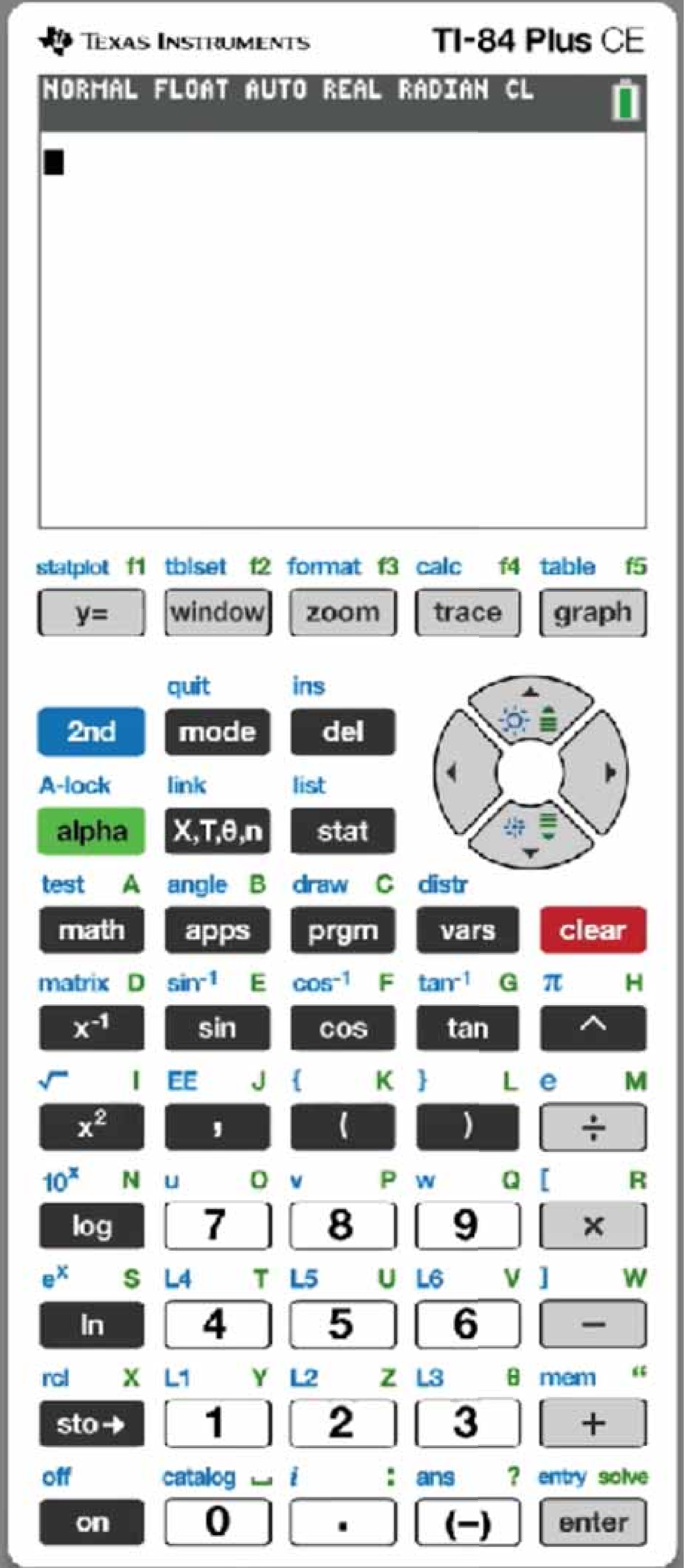
Sample Standard Deviation

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Wait, why the $n - 1$?

We will answer that shortly. There are two formulas for standard deviation which we will look at more closely a bit later.

But let's finish this on the calculator first.



Turns out that squaring each term to make it positive and then adding helps.

Now let's add, divide, and take the square root according to the formula

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Now let's compare to what 1-Var-Stats gives us

Note also that the calculator gives us two standard deviation results. Let's look further into that

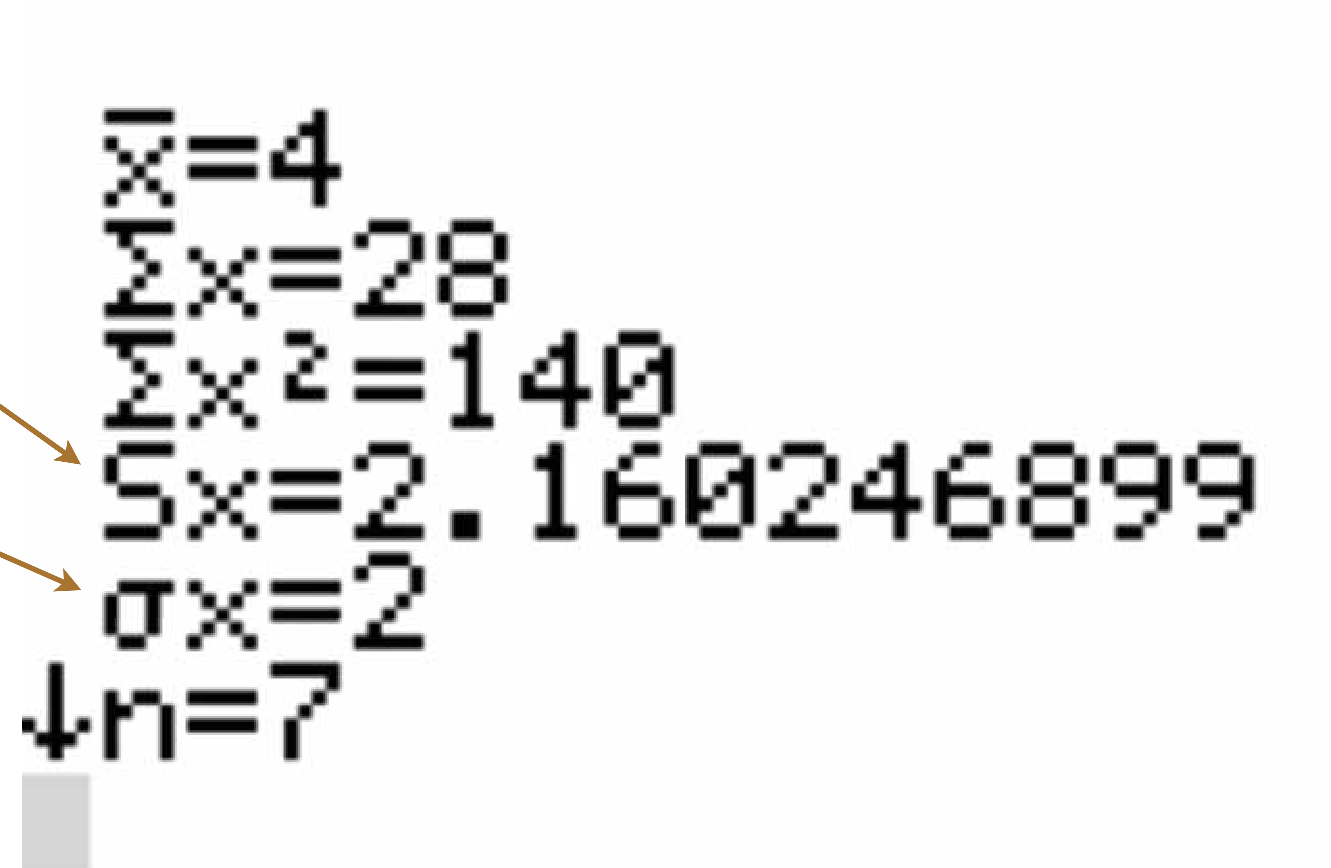
There are actually two Standard Deviation formulas

Let’s try this with just a simple data list of numbers from 1 to 7

This is called the sample standard deviation (the one we’ve been doing)

This is called the population standard deviation

Notice that the sample SD is larger than the population SD



Here’s how they are different:

Sample Standard Deviation

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Sample Standard Deviation

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

$\bar{x} =$ The sample mean

$\mu =$ The population mean

$n =$ # of elements in the sample

$N =$ # of elements in the population

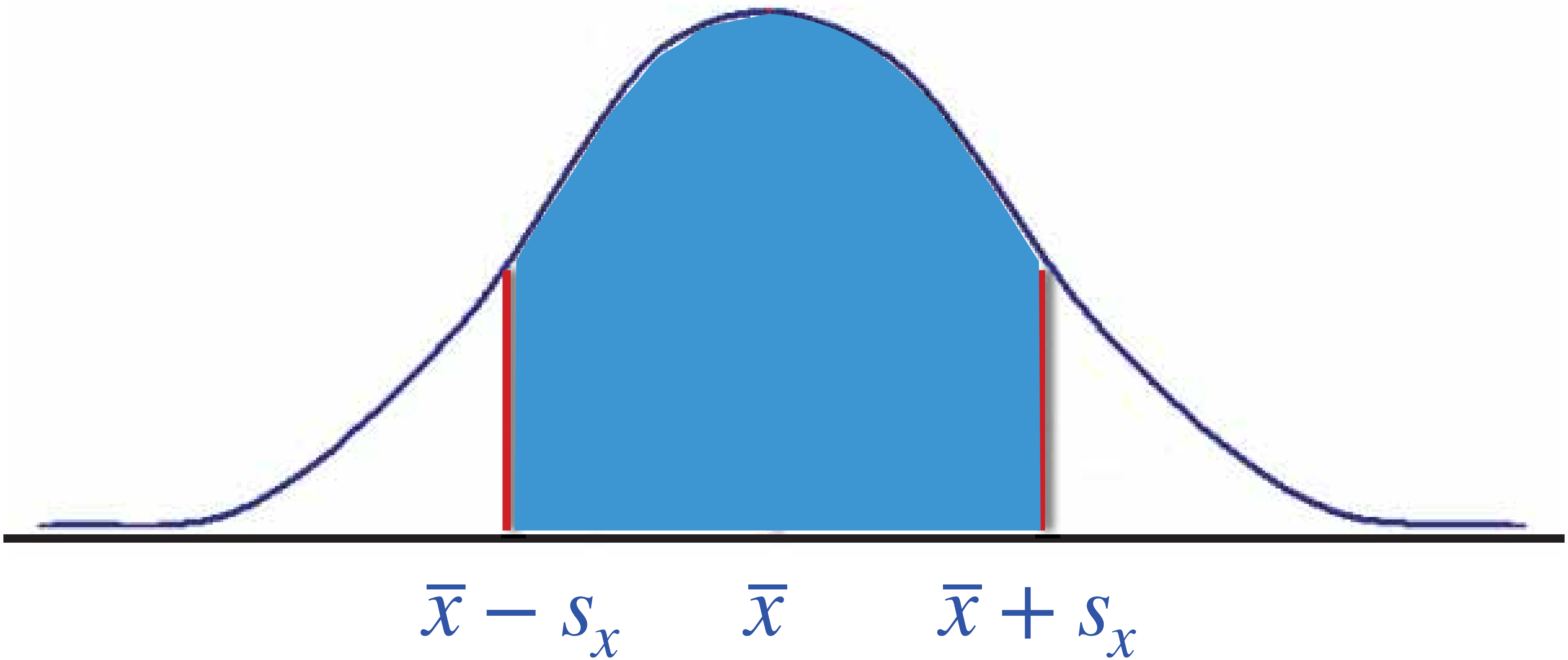
We add the squares and take the square root because to simply add up the variations by themselves always gives us zero (as we saw on the calculator)

We divide by n – 1 with the sample SD in part to account for additional unknown variation between sample and population

There are other explanations beyond the scope of this class.

But the square root of the sum isn't the same as the sum of the square roots so this still doesn't match the average deviation

This is in part because the natural curvature of a normal distribution means that the deviation will be symmetric but not uniform



Note that more data is within one standard deviation of the mean

In fact we shall soon see that just over 2/3 of the data is within one standard deviation of the mean

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

And just to show you what it looks like to calculate them by hand:

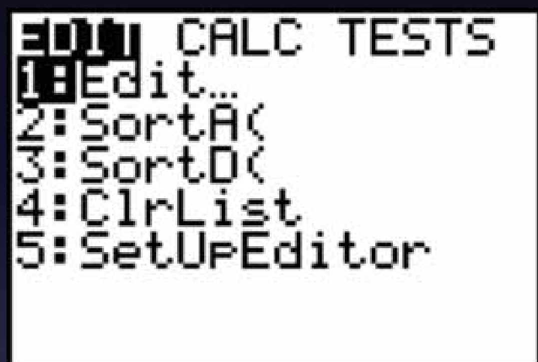
$$= \sqrt{\frac{(1-4)^2 + (2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2}{7-1}} \approx 2.16$$

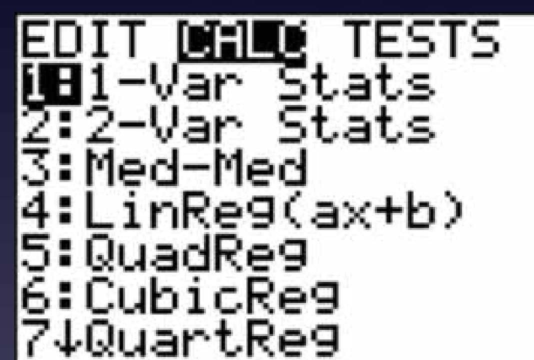
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

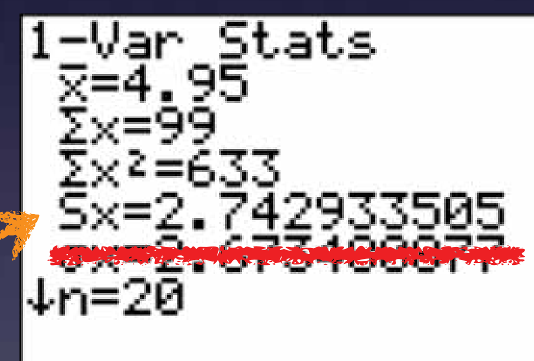
We can do this on the calculator in more than one way. More to come on this when we are doing the assignment problems.

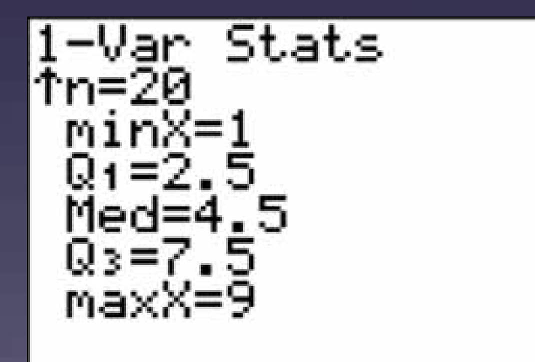
$$= \sqrt{\frac{(1-4)^2 + (2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2}{7}} = 2$$

For large amounts of data, let's just use 1-Var Stats









And since we aren't discussing population data just yet, we'll just stick with S_x

But you do need to know that there is a difference between population s.d. and sample s.d.

Why do we need *BOTH* a measure of center *and* a measure of spread?

Which class scored better on their exam?

A: The class with a mean of 70 and a standard deviation of 4

If a 70 average is ok and you want most to be within that range then this is the outcome you want

B: The class with a mean of 70 and a standard deviation of 12

If you are looking for higher scores, a higher range of scores, and are willing to live with half of your data being considerably below the mean then you can live with this outcome.

Mean - $\frac{\sum_{all\ x} x_i}{n}$

not scary - just means to add all the *x* terms.

Median - Middle #

Mode - # occurring most

Midrange

Measures of Center

Resistant ⟵ to outliers ⟶ Non-Resistant

Resistant Statistics

- Median
- IQR
- Mode

Non-Resistant Statistics

- Range
- Minimum
- Maximum
- Mean
- Standard Deviation
- Variance

As the formula showed us, any change in the data would change these scores

Recall how outliers moved the mean away from the median in the two skewed graphs?