## Other Measures of Spread

#### 5 Number Summary

In the previous sections, we have seen how to calculate the values associated with the **five number summary** (**min**, Q\_1*Q*1​, Q\_2*Q*2​, Q\_3*Q*3​, **max**), as well as the measures of spread associated with these values (**range** and **IQR**).

For datasets that are **not symmetric**, the five number summary and a corresponding box plot are a great way to get started with understanding the spread of your data. **Although I still prefer a histogram in most cases, box plots can be easier to compare two or more groups.** You will see this in the quizzes towards the end of this lesson.

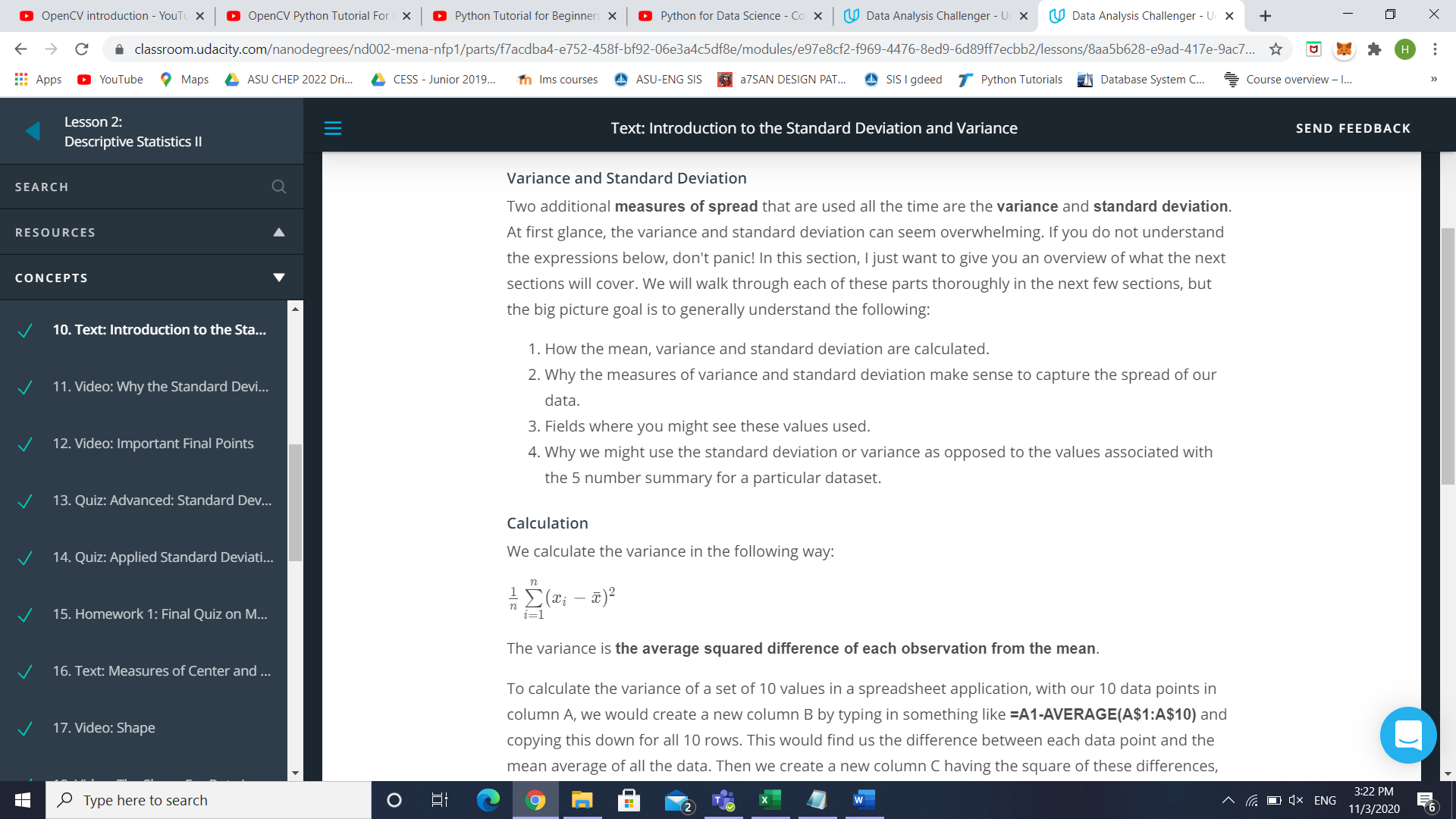
#### Variance and Standard Deviation

Two additional **measures of spread** that are used all the time are the **variance** and **standard deviation**. At first glance, the variance and standard deviation can seem overwhelming. If you do not understand the expressions below, don't panic! In this section, I just want to give you an overview of what the next sections will cover. We will walk through each of these parts thoroughly in the next few sections, but the big picture goal is to generally understand the following:

1. How the mean, variance and standard deviation are calculated.
2. Why the measures of variance and standard deviation make sense to capture the spread of our data.
3. Fields where you might see these values used.
4. Why we might use the standard deviation or variance as opposed to the values associated with the 5 number summary for a particular dataset.

#### Calculation

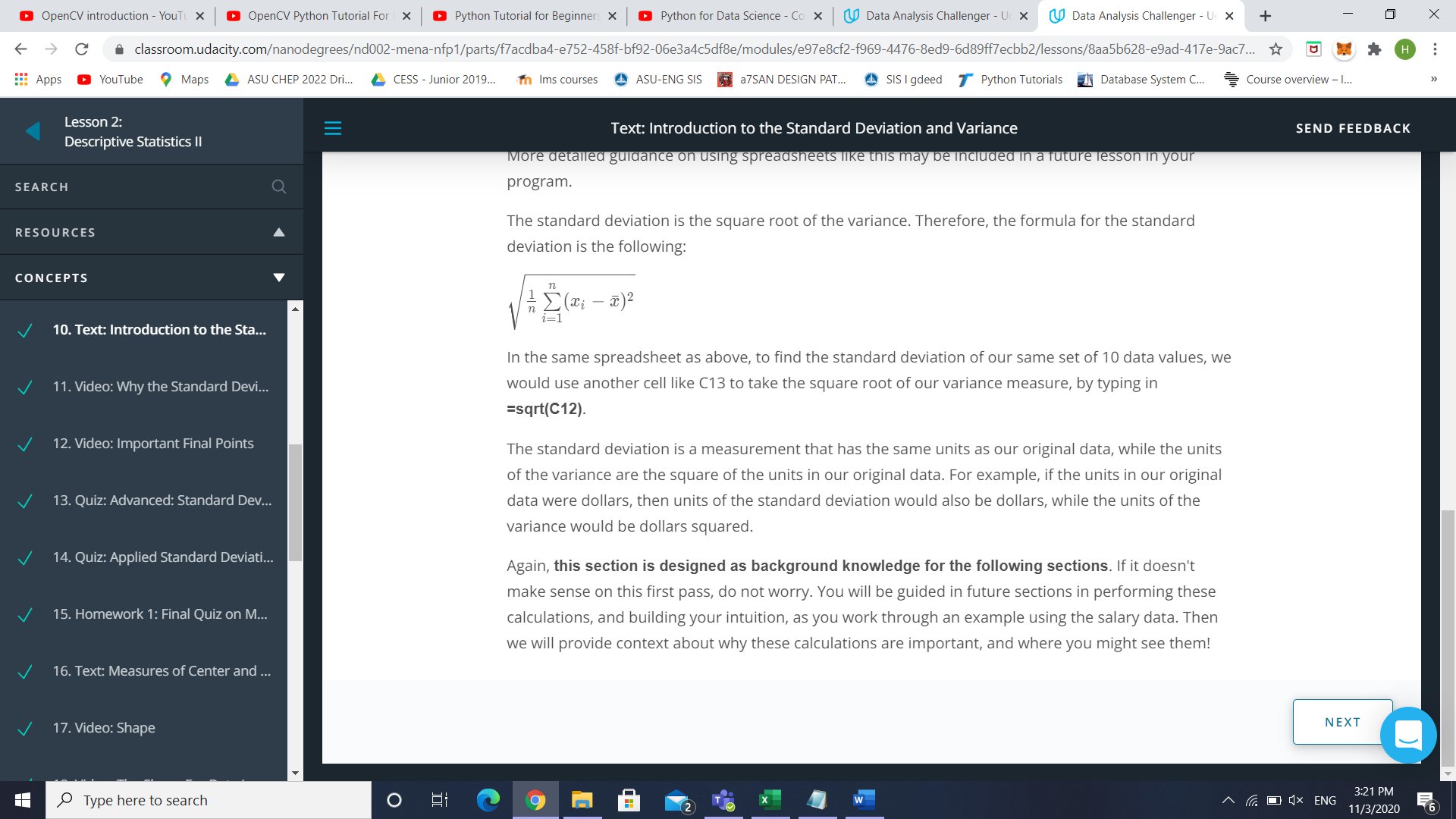
We calculate the variance in the following way:

The variance is **the average squared difference of each observation from the mean**.

To calculate the variance of a set of 10 values in a spreadsheet application, with our 10 data points in column A, we would create a new column B by typing in something like **=A1-AVERAGE(A$1:A$10)** and copying this down for all 10 rows. This would find us the difference between each data point and the mean average of all the data. Then we create a new column C having the square of these differences, using the formula **=B1^2** in cell C1, and copying that down for all rows. Then in the cell below this new column, cell C11, type in **=SUM(C1:C10)**. This adds up all these values in column C. Finally in cell C12, we divide this sum by the number of data points we have, in this case ten: **=C11/10**. This cell C12 now contains the variance for our 10 data points.

More detailed guidance on using spreadsheets like this may be included in a future lesson in your program.

The standard deviation is the square root of the variance. Therefore, the formula for the standard deviation is the following:

In the same spreadsheet as above, to find the standard deviation of our same set of 10 data values, we would use another cell like C13 to take the square root of our variance measure, by typing in **=sqrt(C12)**.

The standard deviation is a measurement that has the same units as our original data, while the units of the variance are the square of the units in our original data. For example, if the units in our original data were dollars, then units of the standard deviation would also be dollars, while the units of the variance would be dollars squared.

**Variable Types**

We have covered a lot up to this point! We started with identifying data types as either categorical or quantitative. We then learned, we could identify quantitative variables as either continuous or discrete. We also found we could identify categorical variables as either ordinal or nominal.

**Categorical Variables**

When analyzing categorical variables, we commonly just look at the count or percent of a group that falls into each **level** of a category. For example, if we had two **levels** of a dog category: lab and not lab. We might say, 32% of the dogs were lab (percent), or we might say 32 of the 100 dogs I saw were labs (count).

However, the 4 aspects associated with describing quantitative variables are not used to describe categorical variables.

**Quantitative Variables**

Then we learned there are four main aspects used to describe quantitative variables:

1. Measures of **Center**
2. Measures of **Spread**
3. **Shape** of the Distribution
4. **Outliers**

We looked at calculating measures of Center

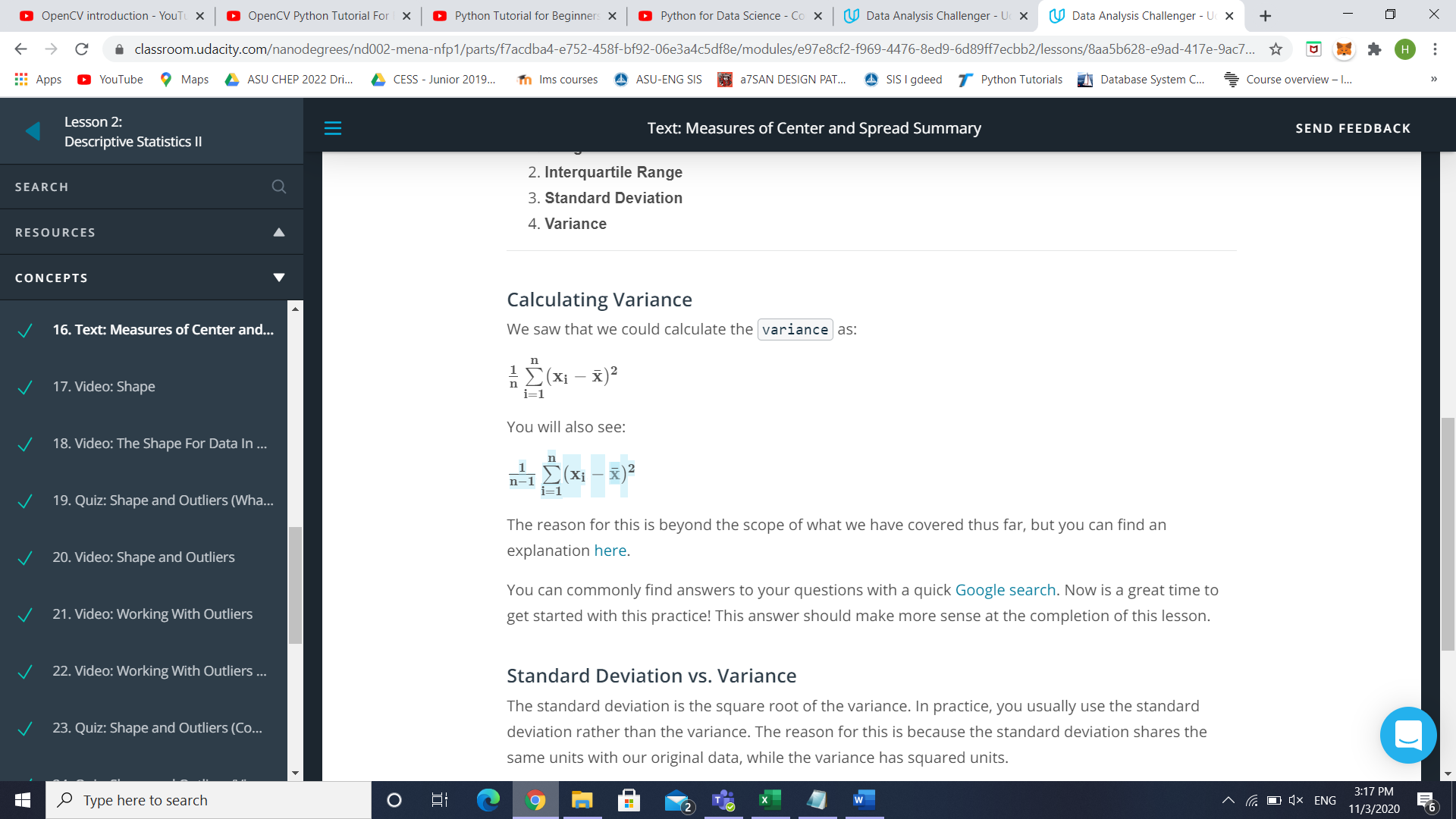
1. **Means**
2. **Medians**
3. **Modes**

We also looked at calculating measures of Spread

1. **Range**
2. **Interquartile Range**
3. **Standard Deviation**
4. **Variance**

**Calculating Variance**

We saw that we could calculate the variance as:



The reason for this is beyond the scope of what we have covered thus far, but you can find an explanation [here](https://stats.stackexchange.com/questions/3931/intuitive-explanation-for-dividing-by-n-1-when-calculating-standard-deviation).

You can commonly find answers to your questions with a quick [Google search](https://www.google.com/). Now is a great time to get started with this practice! This answer should make more sense at the completion of this lesson.

**Standard Deviation vs. Variance**

The standard deviation is the square root of the variance. In practice, you usually use the standard deviation rather than the variance. The reason for this is because the standard deviation shares the same units with our original data, while the variance has squared units.